Filosofisk tidskrift

Årgång 40 Nr 1 Februari 2019

Filosofisk tidskrift

Utges av Stiftelsen Filosofisk tidskrift och Stiftelsen Bokförlaget Thales

Redaktör och ansvarig utgivare: Lars Bergström

Filosofisk tidskrift utkommer med fyra nummer per år

Prenumeration per år inom landet: 200 kr (inkl. moms); utom landet: 300 kr
Lösnr: 55 kr (inkl. moms)
För prenumerationsärenden kontakta:
Nätverkstan ekonomitjänst, Box 31120, 400 32 Göteborg
Telefon 031-743 99 05 Fax 031-743 99 06
ekonomitjanst@natverkstan.net

Förlagsadress: Bokförlaget Thales, Box 50034, 104 05 Stockholm
info@bokforlagetthales.se www.bokforlagetthales.se

Grafisk form och produktion: Ulf Jacobsen
Satt med Arnhem Pro Blond
Tryck: Carlshamn Tryck & Media AB, Karlshamn 2019
ISSN 0348-7482

ANDERS HANSSON

PARFITS MEDGIVANDEPRINCIP

I förordet till *On What Matters*[1] skriver Derek Parfit att Kant och Sidgwick är hans läromästare. Inflytandet från Sidgwick är uppenbart, det genomsyrar hela Parfits projekt. Att Kant ska ha gjort något vidare avtryck på Parfits teori är kanske inte lika uppenbart. Han ägnar visserligen en stor del av *On What Matters*, volym 1, åt Kant. Men i slutändan frågar man sig vad som egentligen blev kvar, teorin han landar i känns inte särskilt kantiansk.

Jag tror inte heller att den är det.[2] Men det finns en avgörande idé som Parfit, i reviderad form, lånar från Kant, nämligen vad han kallar

> *Medgivandeprincipen*: Det är fel att behandla någon på sätt som denna person inte rationellt kan lämna sitt medgivande till. (OWM1, s. 181)

Parfit menar att det är den rimligaste tolkningen av Kants ”Ändamål i sig-formulering” av det kategoriska imperativet. Som jag förstår Parfit är principen dessutom viktig för att rättfärdiga de olika reglerna i hans regelkonsekventialism. Principen ska se till att regelsystemet förhåller sig korrekt till våra partiska intressen, som vårt välmående, våra projekt och hur det går för just de personer vi har särskilda relationer till.

Jag ska försöka förklara hur principen förhåller sig till de grundläggande problemen i Parfits teori, och hur han tänker sig att vi med den kan rättfärdiga ett optimalt regelsystem. Principen har att göra med de problem som Parfit analyserar i *Reasons and Persons* och som teorin i *On What Matters* är tänkt att lösa.

1. Vid citat och hänvisningar refererar jag till *On What Matters* som OWM och *Reasons and Persons* som RP.

2. Jag tror att Parfit missförstår en hel del av Kants resonemang, liksom många andra har gjort före honom. Parfit tänker sig att Kant i första hand är ute efter principer som talar om vad som gör handlingar riktiga, och han förväntar sig att imperativen ska säga något om detta. Men det är inget imperativen uttalar sig om. De handlar om hur en förnuftig person med fri vilja ska förhålla sig till riktiga handlingar. En sådan person, säger imperativen, ska välja riktiga handlingar av rätt anledning, av plikt. När Parfit läser Kant på detta sätt blir resultatet att han missar vad Kant vill ha sagt i vissa grundläggande passager. Jag diskuterar dessa sätt att missförstå Kant närmare i *Tre slags etiska teorier?,* del 3.

Filosofisk tidskrift 2019 nr 1, 3–15

1. MISSTAGEN VI SKA UNDVIKA

I *Reasons and Persons* går Parfit igenom en rad misstag i vårt moraliska tänkande. De olika misstagen framstår som ett antal kluriga tankenötter, men de ska i själva verket visa på svårigheterna vi står inför. Han skriver att fram till förra århundradet levde större delen av mänskligheten i små grupper och våra handlingar påverkade endast ett fåtal andra. Nu har förutsättningarna ändrats (RP, s. 86). Dessvärre är våra begrepp och problemformuleringar alltjämt desamma. En av våra vardagsuppfattningar är till exempel att om ingen tar uppenbar skada av våra handlingar så kan inte handlingen vara felaktig. Men det stämmer inte längre, menar Parfit. I vårt globala och högteknologiska samhälle kan vem som helst påverka hur många som helst, vilket gör att även omärkbara effekter av våra handlingar spelar roll. Vi måste därför undvika vad vi kan kalla

> *Bortse från små effekter-misstaget*: Om en handling har effekter på andra som är omärkbara, så kan inte handlingen vara moraliskt fel på grund av att den har dessa effekter. (RP, s. 75)

Med hjälp av snillrika exempel visar han hur detta misstag får märkliga konsekvenser.

Ett besläktat misstag är att bedöma enbart enskilda handlingars effekter. Parfit menar att vi även måste undvika

> *Enskilda handlingars direkta effekter-misstaget*: Om en handling är rätt eller fel på grund av sina effekter, så är de enda relevanta effekterna just denna handlings effekter. (RP, s. 70)

Om vi inte undviker dessa misstag så blir konsekvenserna att vi tillsammans kommer att handla mycket fel. Det är bland annat fallet med vår klimatpåverkan. Vi måste inse att alla enskilda handlingar som bidrar till utsläppen är felaktiga, på grund av de totala konsekvenser de bidrar till. ”Varje liten höjning av temperaturen”, skriver Parfit, ”kommer att göra att saker går sämre för väldigt många människor, genom att orsaka torka, bränder, översvämningar, och så vidare. Några av dessa effekter kommer att döda många människor. När vi bidrar till den globala uppvärmningen – genom att använda luftkonditionering, till exempel, eller bilar, eller flygplan – är vi några av de personer som är ansvariga för att orsaka dessa framtida dödsfall” (OWM3, s. 430).

Ytterligare ett misstag vi ofta gör leder till vad Parfit kallar Envar-Alla-dilemman (RP, s. 87). I dessa dilemman handlar var och en av oss bäst i en teoris termer, men tillsammans handlar vi sämre i teorins termer. Problemet är att om alla är ”disponerade eller motiverade att till exempel göra det som vore bättre för oss själva eller vår familj eller dem vi älskar, så kommer det att gå sämre för oss alla. Om inget ändras i förutsättningarna så kommer utfallet att bli sämre för oss alla” (RP, s. 87). Frågan är då vad som ska ändras?

Detta problem uppstår när var och en av oss har *olika* mål. Många av de vanligaste teorierna säger med andra ord att vi har agent-relativa skäl som är skäl för olika personer att ha och att försöka uppfylla olika mål (se till exempel RP, s. 27, 104). Det *relativa* består i att vissa fakta ger skäl som refererar till de personer som har dessa skäl. Parfits exempel är bland annat var och ens intresse i att förhindra sitt eget lidande. Att vi alla har olika mål innebär dock att min framgång kommer att vara beroende av vad andra gör. Parfit ber oss till exempel att tänka på en situation där vi har två alternativ. Antingen ser var och en till att utföra några av sina egna plikter, eller så ser vi alla till att göra det möjligt för andra att utföra fler av deras plikter. Resultatet kan bli så att om alla enbart bryr sig om sina plikter, så kommer var och en av oss att kunna utföra färre av våra egna plikter (RP, s. 98).

Bäst för alla vore att vi försökte göra det möjligt för andra att utföra fler av deras plikter. Teorier som ger oss gemensamma mål, som säger att vi har agent-neutrala skäl, undviker alltså detta misstag. Parfit skriver att agent-neutrala skäl är skäl för alla att omfatta eller att försöka uppnå ett gemensamt mål. Det *neutrala* består här i att vissa fakta eller mål ger oss skäl utan att referera till någon särskild person. Parfits vanligaste exempel är att alla har skäl att bry sig om lidande. Förekomsten av lidande är nämligen i sig något dåligt. Det faktum att det förekommer lidande någonstans är dåligt oavsett vem lidandet hör till och alla har skäl att önska att det upphör.

Att vi gör misstag av dessa slag talar för att många av våra begrepp inte längre håller måttet. Omständigheterna är inte längre sådana att de fyller sin funktion. Parfit skriver att givet den vetenskapliga, teknologiska och ekonomiska utvecklingen de senaste två århundradena står vi nu inför moraliska problem som vår vardagsmoral inte kan lösa. Så vi måste försöka förstå och acceptera några nya principer (OWM3, s. 422). Han menar bland annat att vi måste inse att handlingar ”kan vara mycket fel, på grund av dess effekter på andra människor, även

om ingen av dessa människor någonsin märker av dessa effekter. Våra handlingar kan tillsammans göra att dessa människor får det mycket sämre" (RP, s. 83).

2. KONFLIKTER MELLAN OLIKA SLAGS SKÄL

Vi måste även uppmärksamma en annan viktig komponent i Parfits resonemang, nämligen att olika slags skäl kan hamna i konflikt med varandra. Vi kan ha skäl av flera slag i samma situation.

För Sidgwick var detta ett olösligt problem. Han såg de olika slags skälen som ojämförbara och då, resonerade han, tycks vi kunna ha lika goda skäl att utföra två olika handlingar i samma situation.

Parfit håller inte med. En sådan konflikt, menar han, innebär inte att vi har *mest* skäl att utföra två olika handlingar, utan bara att det finns omständigheter som talar för två olika handlingar. Det finns dock alltid ett korrekt svar på vilken handling vi har mest skäl att välja.[3] I vissa lägen kan det emellertid i praktiken vara svårt för oss att avgöra vilken handling vi har mest skäl att utföra. Orsaken är att praktiska frågor inte erbjuder den precision som gör det möjligt att komma fram till exakt vad vi har skäl att göra. Det betyder dock inte att det inte finns ett svar, bara att vi kan ha svårt att nå fram till det.

I sådana lägen, menar Parfit, skulle vi ändå ofta kunna ha tillräckliga skäl att välja vilket som helst av alternativen. Detta innebär att det ofta skulle kunna vara rimligt att handla så att det gynnar andra mer än det gynnar oss själva. Parfit godtar nämligen något i stil med

> *Parfits opartiskhetsprincip*: det måste inte vara irrationellt att välja att befrämja andras intressen framför våra egna, och vi kan därför fullt rationellt välja handlingar eller principer som gynnar andra mer än oss själva. (RP, s. 131)

Vi skulle alltså kunna ge samma eller till och med större vikt till någon främlings välmående (OWM1, s. 139). Men han hänvisar lika ofta till en princip som tillåter att vi tar hänsyn till våra partiska skäl och befrämjar våra egna eller närståendes intressen framför flera andras, även i de fall priset skulle bli högt för andra. Vi kan kalla det

3. Konflikten mellan partiska och opartiska skäl är i själva verket Parfits huvudproblem. Utgångspunkten är Sidgwicks resonemang. Jag diskuterar Sidgwicks problem och Parfits lösning närmare i *Tre slags etiska teorier?*, kapitel 15 och 16.

> *Parfits partiskhetsprincip*: I vissa fall har vi partiska och agent-relativa skäl som väger mycket tyngre än relevanta opartiska skäl, och då är det rationellt att befrämja våra partiska intressen framför många andras intressen som vi har opartiska skäl att befrämja.

Detta leder till vad vi kan kalla

> *Principen om tillräckliga skäl*: Vi har alltid mest skäl att göra vad som helst som skulle bli opartiskt bäst, såvida inte någon annan handling skulle vara bäst för oss själva. I sådana fall skulle vi ha tillräckliga skäl att handla på båda sätten. Om vi kände till relevanta fakta, så skulle båda handlingarna vara rationella. (OWM1, s. 131)

Parfits tanke är alltså att utgångsläget alltid är att vi ska göra mest gott, vilket pekar mot någon sorts konsekventialistisk teori (OWM1, s. 246). Men det finns två viktiga undantag.

För det första menar Parfit att vi har skäl att bry oss om en mängd olika saker. Hans konsekventialism är pluralistisk och säger att det finns en mängd saker som är värdefulla.

Och för det andra säger ju principen om tillräckliga skäl att vi ibland kan ha partiska skäl att bortse från det övergripande målet att göra så mycket gott som möjligt. Detta innebär att Parfits konsekventialism är av ett annat slag än till exempel den klassiska utilitarismen.

3. PARFITS REGELKONSEKVENTIALISM

De flesta av våra moraliska föreställningar har sådana brister att de måste justeras, eller ges upp. Däremot skulle en konsekventialistisk teori klara sig relativt bra. Problemet är att en *rent* konsekventialistisk teori ofta skulle föreslå handlingar som vi inte vill utföra, sådana som de flesta tycker är moraliskt felaktiga, som mord, tortyr, bedrägeri och liknande. Dessa anses vara fel i sig, det vill säga oavsett deras konsekvenser. Många skulle vilja att en teori bättre kan redogöra för detta än vad till exempel utilitarismen gör genom att hänvisa till de dåliga sidoeffekterna av sådana handlingar, som att de sprider skräck och otrygghet i samhället.

Parfit väljer att formulera en regelkonsekventialistisk teori som först och främst syftar till att göra mest gott, opartiskt sett. Men för att undvika de problem jag tagit upp måste den ha några inbyggda restriktioner. Det är här Kants medgivandeprincip kommer in genom

att principen skulle utesluta regler som någon skulle ha konklusiva partiska skäl att underkänna. Till exempel skulle den som riskerar att bli torterad ha sådana skäl att invända mot tortyr, varför regelsystemet måste innefatta principer som förbjuder detta.

Det är dock inte helt klart vad för sorts regelkonsekventialism Parfit tänker sig. Jag tror det kan behövas några ord om det. För det första är det fråga om en kollektiv variant av konsekventialismen. En sådan uttalar sig om vad en grupp personer bör sikta mot under ideala förutsättningar, det vill säga om alla andra skulle följa samma regler. Detta till skillnad från till exempel handlingskonsekventialismen som säger att var och en av oss i varje situation ska försöka avgöra vilken av våra enskilda handlingar som har bäst konsekvenser, givet vad alla andra gör.

Parfits konsekventialism säger att var och en ska försöka ha den uppsättning önskningar och dispositioner sådan att om alla hade en sådan uppsättning skulle utfallet bli bättre än om alla hade någon annan uppsättning (RP, s. 30). En kollektiv teori skulle undvika vad Parfit kallade Envar-Alla-dilemman (RP, s. 87).

Det skulle inte handlingskonsekventialismen. Det kan nämligen vara så att när var och en av oss försöker handla bäst utan att ta hänsyn till vad vi tillsammans kan göra så kan handlingskonsekventialismen säga vi bör låta bli att bidra om vi tror att för få skulle hjälpa till. Den skulle med andra ord rekommendera att vi i dessa fall väljer en sämre handling än vi hade kunnat välja. Ett exempel från Torbjörn Tännsjö kan illustrera detta:

> En grupp om tre personer har blivit stående med en bil med motorstopp och trasig startmotor i en uppförsbacke. Tillsammans kan de skjuta bilen upp till backens krön, sätta sig i den och i nedförsbacken på andra sidan krönet starta motorn i farten. Om alla gör en insats går allt väl. I annat fall misslyckas man.[4]

Om ingen skjuter på kan var och en säga att de handlat rätt med hänvisning till att ingen annan heller gjorde det och att det därför vore meningslöst att försöka på egen hand. Däremot kan man ju tycka att ”kollektivet”, alltså gruppen om tre personer, har handlat fel.[5] Detta ska Parfits regelkonsekventialism råda bot på. Den säger att vi inte ska välja optimala handlingar, utan att vi alla i stället ska följa optimala regler, det vill säga de regler för vilka det är sant att det inte finns någon annan

4. Tännsjö, ”Kollektiv skuld”, s. 23f.
5. ”Kollektiv skuld”, s. 25.

regel som, om vi accepterade och följde den, skulle göra att saker gick bättre (OWM3, s. 417). På så vis skulle den undvika många Envar-Alla-dilemman.

En annan sak som denna typ av konsekventialism skulle lösa är vår tendens att bortse från effekter som är små eller triviala. Vår uppfattning om vad som gör handlingar felaktiga är uppenbarligen inte korrekt. Var och en av oss handlar på sätt som förstör klimatet. Parfit menar att felaktigheten i dessa handlingar bäst kan förklaras på regelkonsekventialistiskt sätt (OWM3, s. 432). Men varför skulle just regelkonsekventialismen förklara detta bäst?

Parfits tanke är att det regelsystem som har bäst konsekvenser inte skulle tillåta enskilda handlingar som, om de vore en del i en uppsättning handlingar, har dåliga konsekvenser. Dessa regler skulle förbjuda eller reglera själva uppsättningen, och därmed alla de handlingar som utgör uppsättningen. På detta vis är Parfits regelsystem formulerat i enlighet med tanken om kollektiv konsekventialism, vilket samtidigt kommer åt problemet med små effekter.

Sedan har vi frågan om vad som gör att saker ”går bäst”. Vad är egentligen värdefullt? Vilka fakta ger oss skäl? Parfit menar att det finns en mängd saker som vi har skäl att bry oss om – alltifrån relationer till våra nära, till att göra vetenskapliga upptäckter och att lyckas med olika bedrifter. Liksom att skaffa kunskap och att skapa vackra saker. Och naturligtvis att kännande varelser har det bra. Dessutom, som jag varit inne på, så menar vi ju ofta att det finns handlingar som i sig är felaktiga, och som det vore i sig dåligt att utföra. Regelsystemet skulle alltså även behöva förhålla sig till att vissa handlingar har ett negativt intrinsikalt värde i egenskap av att vara moraliskt felaktiga, såsom tortyr och mord. Det bästa regelsystemet måste spegla denna pluralism.

I många avseenden skulle nog ett sådant regelsystem se ut ungefär som vår vardagsmoral (vilket även Sidgwick var inne på) dock med en del viktiga justeringar. Å ena sidan måste regelsystemet kunna redogöra för vissa partiska skäl att bry oss mer om vissa personer. Parfit skriver till exempel att då det vore bäst ”om de flesta av oss älskar våra nära och våra vänner väldigt mycket, så skulle de optimala reglerna spegla detta faktum” (OWM3, s. 420). Men det betyder inte att vi aldrig behöver ta hänsyn till andras intressen. För även om de reglerna ofta skulle kunna säga att vi får handla på sätt som vore sämre än alternativen, opartiskt sett, så framhåller Parfit att en rättvis teori rimligen ändå skulle kräva att vi ger upp relativt mycket för att hjälpa dem som har det sämre. Han skriver att

> dessa regler skulle inte bortse från den stora orättvisan i global ojämlikhet, och det lidande och de förtidiga dödsfallen bland världens fattigaste människor. Enligt alla rimliga versioner av regelkonsekventialismen bör vi rika människor ge bort mycket av det vi ärvt eller tjänat. Vi skulle kunna göra detta utan att ge upp vår starka kärlek till vissa personer. (OWM3, s. 421)

Regelsystemet är alltså att uppfatta som en optimal avvägning mellan den opartiska utgångspunkten – det vill säga att vi ska handla så att konsekvenserna opartiskt sett blir de bästa – och våra partiska mål och skäl. Det är spänningen mellan dessa två ståndpunkter som gör medgivandeprincipen aktuell.

Det är denna konflikt som utgör Parfits grundläggande problem. Finns det verkligen opartiska skäl som väger tyngre än våra partiska skäl?[6] Och finns det verkligen ett konsekventialistiskt regelsystem som var och en av oss har skäl att godta givet våra partiska mål och ståndpunkter? Parfit svarar att det finns det. Och han använder medgivandeprincipen för att visa det.

4. HUR SKA VI ANVÄNDA MEDGIVANDEPRINCIPEN?

Parfit lånar som sagt medgivandeprincipen från Kant och betraktar den som en reviderad version av en i grunden ganska rimlig deontologisk princip. I sin reviderade form är principen tänkt att hjälpa oss att komma fram till hur godtagbara regler i ett optimalt regelsystem måste se ut. Den säger att reglerna måste vara sådana att var och en rationellt kan önska dem, eller ge sitt medgivande till dem. Medgivandeprincipen är alltså en av de omständigheter som talar om varför vissa handlingar är felaktiga.

Som vi såg tidigare säger principen att det är fel att behandla någon på sätt som denna person inte rationellt kan lämna sitt medgivande till (OWM1, s. 181). Nyckelformuleringen här är *rationellt medgivande*. För att rationellt kunna ge sitt medgivande till något måste vi ha åtminstone tillräckliga skäl att godta regler, handlingar, osv. Tillräckliga skäl har vi när vi inte har konklusiva skäl att välja något av alternativen i någon situation. Då skulle vi fullt rationellt kunna välja vilket som helst av alternativen (OWM1, s. 32–33).

6. Jag diskuterar Parfits grundläggande fråga utförligare i *Klimatet, fattigdomen och lidandet*.

Parfits tanke är alltså att vi skulle kunna godta en regel ifall var och en har åtminstone *tillräckliga* skäl att göra det från både sin partiska och en opartisk ståndpunkt. Men om någon skulle ha *konklusiva* partiska skäl att inte godta en regel, så innebär det att den är felaktig. En hel del rent konsekventialistiska principer skulle rimligen underkännas för att många skulle ha konklusiva partiska skäl att underkänna dem.

Samtidigt skulle vi ha tillräckliga skäl att godta en hel del regler som lägger stor börda på vissa personer. Detta med anledning av Parfits opartiskhetsprincip som innebär att det inte måste vara irrationellt att välja att befrämja andras intressen framför våra egna, i vissa lägen.

I och med att medgivandeprincipen är åtminstone en av de principer som talar om för oss vad som gör handlingar felaktiga så måste det optimala regelsystemet förhålla sig till den. Regelsystemet kan inte föreslå att vi handlar på sätt som någon eller några skulle ha konklusiva skäl att underkänna. Jag tror alltså att man härigenom kan säga att det konsekventialistiska regelsystemet, enligt Parfit, rättfärdigas med hjälp av en deontologisk princip. Det innebär att regelsystemet inte kommer att vara strikt konsekventialistiskt, trots att det är optimalt. Det kommer alltid att vara begränsat av vissa restriktioner. Regelsystemet måste därför formuleras så att det har de bästa tänkbara konsekvenserna *givet* de restriktioner som medgivandeprincipen uttrycker.

Detta förklarar hur det kommer sig att Parfit menar att Kant skulle hålla med om att alla har skäl att välja ett optimalt regelsystem. Parfits argument går så här (OWM1, s. 402–3):

(1) Vi måste handla enligt principer som alla rationellt kan välja.

(2) Alla kan rationellt välja vissa *optimala* principer, det vill säga principer vars konsekvenser är de bästa, om alla handlar enligt dem.

(3) Det finns vissa *optimala* principer.

(4) Det finns inga *icke-optimala* principer som alla har skäl att välja.

(5) Inga opartiska skäl att välja de optimala principerna skulle vägas ut av konklusiva partiska skäl.

(6) Alltså bör vi välja dessa *optimala* principer.

Detta leder till vad han kallar *Kantiansk regelkonsekventialism* som säger att alla bör följa de optimala principerna, för dessa är de enda principer som alla rationellt skulle kunna vilja vore universella lagar (OWM1, s. 411).

För att argumentet ska vara giltigt behöver dock Parfit förklara hur vi kan godta framför allt premiss (5). Stämmer det verkligen att vi, från vår egen ståndpunkt, skulle kunna önska att alla handlar enligt ett optimalt regelsystem som i vissa fall skulle kunna påverka just oss negativt på allvarliga sätt? Parfit ställer frågan så här:

> Om vi valde principer från en opartisk synvinkel så skulle det vara de optimala principerna som alla har mest skäl att välja. Men i tankeexperimentet som denna kantianska formel hänvisar till, så skulle vi inte välja principer från en opartisk synvinkel. Våra val skulle påverka våra egna liv liksom livet för de personer vi har nära relationer till … Så vi skulle kunna ha starka personliga och partiska skäl att inte välja de optimala principerna. (OWM1, s. 379)

Om en regel skulle påverka oss eller våra nära negativt, skulle vi inte kunna sägas ha konklusiva invändningar mot denna regel? Vore det inte, åtminstone ibland, irrationellt av oss att välja de principer som är optimala?

Parfit svarar som sagt nej. Han menar att vi *alltid* har skäl att välja de optimala principerna. Det finns inga icke-optimala, det vill säga icke-konsekventialistiska, konklusiva skäl att underkänna de optimala principerna. Alla har alltså åtminstone tillräckliga skäl att godta en viss uppsättning optimala principer. Parfit vill visa detta genom att titta närmare på några av de främsta invändningarna i *On What Matters* volym 1.[7]

Han börjar med invändningen från egenintresse och tillämpar då sin opartiskhetsprincip, det vill säga tanken att vi ofta fullt rationellt skulle kunna bortse från våra egna intressen (OWM1, s. 380f.). Därefter ställer han våra partiska skäl mot olika perfektionistiska skäl eller målsättningar och förklarar att vi mycket väl skulle kunna förhålla oss opartiska även till olika perfektionistiska målsättningar. Vi skulle till exempel, fullt rationellt, kunna önska att fem andra personers litterära mästerverk räddas hellre än vårt eget mästerverk (OWM1, s. 389).

En annan viktig fråga är ifall regelsystemet skulle kunna släppa igenom vissa i sig felaktiga handlingar, som mord, tortyr och liknande.

7. Han gör det på ett ställe i *On What Matters* volym 1, s. 380–403. Denna passage förtjänar att läsas noga när man vill förstå medgivandeprincipen närmare. Man bör även vända sig till *On What Matters*, volym 3, del 10, för att se hur han menar att många av de främsta deontologiska restriktionerna inte är så starka som man vanligen tänker sig och därmed inte hamnar i konflikt med konsekventialistiska principer så ofta. Även detta skulle tala för att vi har anledning att önska att alla handlar enligt ett optimalt regelsystem.

Detta är handlingar de flesta av oss menar är moraliskt felaktiga, och som vi inte får utföra. Det talar alltså för att vi skulle ha rent moraliska skäl att invända mot regelsystemet (OWM1, s. 390f.).

Just denna invändning är lite knepig. Parfit menar nämligen att medgivandeprincipen inte kan säga att vissa handlingar är fel på grund av sina ”deontiska” egenskaper, alltså på grund av att de är i sig felaktiga. Detta för att medgivandeprincipen just ska förklara vad som gör en handling felaktig och då kan man inte underkänna handlingen med hänvisning till medgivandeprincipen för att den är i sig felaktig. Att någon inte har skäl att lämna sitt medgivande är ju enligt principen vad som gör handlingen fel, och man hamnar så att säga i en rundgång om man hänvisar till dess intrinsikala felaktighet. Medgivandeprincipen måste i stället hänvisa till andra omständigheter, som att handlingar av detta slag har effekter som vi har anledning att vilja att regelsystemet förbjuder, som död och lidande, utan att hänvisa till deras intrinsikala felaktighet.

Det innebär dock inte att regelsystemet inte kan ta hänsyn till handlingars intrinsikala felaktighet, men vi kan inte hänvisa till medgivandeprincipen för att förklara det. Det faktum att sådana handlingar i sig är felaktiga skulle dock vara ytterligare, oberoende, omständigheter som regelsystemet behöver ta hänsyn till (OWM1, s. 396).

Parfit kommer hur som helst fram till att regelsystemet inte skulle rekommendera handlingar som vi skulle ha konklusiva skäl att invända mot. I de fall någon skulle drabbas negativt av vissa regler så skulle det handla om sådana regler som alla har tillräckliga skäl att godta.

På detta sätt tittar Parfit närmare på vad han menar är de främsta invändningarna mot att det skulle gå att rättfärdiga ett optimalt regelsystem. Ingen av invändningarna håller, menar han, och vi har därmed anledning att försöka formulera ett sådant regelsystem.

För att kunna formulera ett regelsystem som alla har skäl att önska att alla handlar enligt är det alltså nödvändigt att man uppfattar konsekventialismen pluralistiskt, då det finns en mängd saker som spelar roll, eller som kan göra saker bäst, enligt Parfit. Därmed måste regelsystemet ta hänsyn både till allas partiska skäl och till olika slags deontologiska restriktioner, olika ideal och plikter. Dessutom bör man uppfatta konsekventialismen på ett ”handlingsinkluderande” sätt, vilket innebär att det ofta i sig vore dåligt att behandla människor på olika sätt, som att bedra dem eller tvinga dem, eller att bryta löften vi gett dem (OWM3, s. 396). Att förhålla sig på rätt sätt till sådana principer *kan* göra saker bäst, vilket

således det optimala systemet måste redogöra för.[8] En optimal princip kan alltså säga att vi ska utföra generösa handlingar, och förbjuda tortyr, och prioritera vår familjs intressen eller vår egen lycka, liksom att vi måste respektera rättigheter eller liknande, trots att det skulle kunna ha bättre konsekvenser att bortse från dem. Härav, är tanken, skulle alla ha åtminstone tillräckliga skäl att godta de optimala principerna.

5. AVSLUTNING

Parfits teori är ett förslag på hur vi kan komma tillrätta med våra mest akuta problem: klimatet, fattigdomen och lidandet i världen. Upphoven till dessa problem analyserade han i sin första bok. Lösningsförslagen ger han i sin andra. Han menar att vi måste inse att vi resonerar fel i avgörande frågor. Vi bortser från sådant som är relevant och tar hänsyn till sådant som inte är det. Därmed misslyckas vi med att tillsammans göra gott för världens fattiga, men paradoxalt nog lyckas vi att tillsammans förstöra klimatet, utan att riktigt inse det. Vi måste alltså försöka handla bättre än vi gör i dag.

Som regelkonsekventialister, säger han, skulle vi kunna undvika dessa problem och handla bättre. Som sådana hävdar vi att vissa handlingar är fel för att de är otillåtna av optimala regler. Huruvida en regel är optimal ”beror på om saker på det hela går bättre om de flesta av oss eller många av oss, accepterade och försökte följa en sådan regel” (OWM3, s. 432).

Men för att var och en av oss ska kunna godta regelsystemet så måste det testas med hjälp av medgivandeprincipen. Det lägger restriktioner på systemet så att var och en av oss skulle kunna vilja att alla handlade enligt detta system. Så man kan nog ändå säga att Kant har gjort avtryck på Parfits teori. Jag är dock inte säker på att Kant skulle ha känt igen sitt imperativ efter Parfits justeringar.

LITTERATUR

Hansson, Anders. 2018. *Tre slags etiska teorier? Aristoteles, Kant och Sidgwick om vad vi har skäl att göra*. Stockholm: Thales.

———. Kommande. *Klimatet, fattigdomen och lidandet: Derek Parfit om vår tids största moraliska problem*. Göteborg: Daidalos.

Parfit, Derek. 1984. *Reasons and Persons*. Oxford: Clarendon Press.

8. För mer om hur Parfit tänkte sig en fungerande konsekventialism, se kapitlet ”Parfits konsekventialism” i min *Klimatet, fattigdomen och lidandet*.

———. 2011. *On What Matters*, volym 1. Oxford: Oxford University Press.
———. 2017. *On What Matters*, volym 3. Oxford: Oxford University Press.
Sidgwick, Henry. 1907. *The Methods of Ethics*, sjunde utgåvan. London: Macmillan.
Tännsjö, Torbjörn. 2003. ”Kollektiv skuld”. *Tidskrift för politisk filosofi*, nr 3.

MARCO TIOZZO

OM BETYDELSEN AV EVIDENS AV HÖGRE ORDNING

1. INLEDNING

Evidens är underlaget på vilket vi baserar våra trosföreställningar. En vanlig uppfattning är att om man har evidens som ger stöd åt en viss trosföreställning så är man också berättigad att ha denna trosföreställning. På senare tid har man inom epistemologi intresserat sig för vad man kallar ”evidens av högre ordning”.[1] Vad som avses är i grova drag evidens om evidens. Framförallt har man intresserat sig för situationer där evidens av högre ordning tycks *underminera* välgrundade trosföreställningar. Här är ett exempel:

> Du är en läkare som diagnostiserar patienter och föreskriver lämplig behandling. Efter att ha diagnostiserat en patient och ordinerat en viss medicinering, får du veta av en sjuksköterska att du har varit vaken i 36 timmar. Med tanke på vad du vet om folks benägenhet att göra missbedömningar som ett resultat av sömnbrist blir du betydligt osäkrare på om diagnosen och ordinationen av medicin som du har utfärdat är korrekta.[2]

I detta fall utgör informationen att du har varit vaken i 36 timmar evidens av högre ordning. Märk väl att det inte rör sig om någon evidens i vanlig bemärkelse då den inte tillför ny information som har att göra med det aktuella fallet, t.ex. en upptäckt i patientjournalen. Evidens av högre ordning antyder i stället något om kvaliteten hos den bedömning du redan har gjort. I detta fall att du p.g.a. av sömnbrist kan ha föreskrivit en felaktig behandling av patienten.

Vissa epistemologer anser att evidens av högre ordning har förmågan att underminera våra trosföreställningar medan andra anser att så inte behöver vara fallet. Det har skrivits åtskilligt i ämnet och debatten har även kommit att få konsekvenser för andra frågor, t.ex. den epistemiska betydelsen av oenighet.[3] Diskussionen om betydelsen av evidens av högre ordning har dock ännu inte blivit särskilt uppmärksammad i svensk

1. Christensen (2010), Feldman (2005), Horowitz (2014), Kelly (2005, 2010), Lasonen-Aarnio (2014).

2. Detta är en adaption av ett exempel som förekommer i Christensen (2010, s. 186)

3. För en introduktion av debatten om den epistemiska betydelsen av oenighet på svenska se Tiozzo (2016).

Filosofisk tidskrift 2019 nr 1, 16–24

filosofisk debatt. Det huvudsakliga syftet med denna artikel är därför att introducera debatten för den svenska publiken. Vid sidan av detta kommer jag även att försvara den något omstridda uppfattningen att evidens av högre ordning inte måste underminera välgrundade uppfattningar.

2. EN VATTENDELARE: EPISTEMISK AKRASIA

Vad som verkar ligga till grund för konflikten mellan de som anser att evidens av högre ordning är underminerande och de som inte håller med är hur man ställer sig till ett fenomen som kallas för *epistemisk akrasia*. Filosofer har sedan länge intresserat sig för en typ av praktisk irrationalitet (akrasia) som innebär att man misslyckas med att avse att göra det man anser att man borde göra. Till exempel, om jag anser att jag inte borde röka måste jag också ha avsikten att inte röka för att vara rationell. Den epistemiska motsvarigheten innebär att man misslyckas med att anamma en viss trosföreställning eller attityd trots att man är medveten om att ens evidens ger stöd åt denna trosföreställning eller attityd.[4]

Man skulle t.ex. kunna tänka sig att man är medveten om att det finns gott om evidens som talar för att det är farligt att röka men att man för den skull inte bildar trosföreställningen att ”rökning är farligt”. I så fall skulle man tro något i stil med ”evidensen ger stöd åt att det är farligt att röka, men jag tror inte att det är farligt att röka”. En sådan kombination av trosföreställningar är minst sagt underlig och verkar ge uttryck för en typ av *Moore-paradox*. G. E. Moore påpekade att det finns något absurt i att påstå att ”det regnar, men jag tror inte att det regnar”. Enligt den epistemiska versionen av Moores paradox är det även absurt att påstå att ”det regnar, men jag vet inte att det regnar”.[5] På samma sätt verkar det också märkligt att påstå att ”det regnar” om man inte tror att evidensen stödjer detta påstående. Det intuitivt underliga i att hysa akratiska kombinationer av trosföreställningar har fått många att stödja följande princip:

> *Det enkratiska villkoret*. För att vara rationell måste S (i) tro att *p* om S tror att evidensen stödjer *p* och (ii) inte tro *p* om S tror att evidensen inte stödjer *p*.[6]

4. Här kan det röra sig om olika typer av attityder, som att tro att *p* är sann, att tro att *p* är falsk, eller att suspendera omdömet om huruvida *p* är sann.

5. Se Huemer (2007) för mer om den epistemiska versionen av Moores paradox.

6. Det enkratiska villkoret har fått sitt namn från John Broome. Se Broome (2013) för mer bakgrund.

Eftersom epistemisk akrasia anses vara paradigmatiskt irrationellt är det enkratiska villkoret också betraktat som ett nödvändigt villkor för *epistemisk rationalitet*.[7] Enligt vissa (t.ex. Adler 2002) anses det t.o.m. vara *omöjligt* att vara epistemiskt akratisk. Detta då det enkratiska villkoret anses vara konstitutivt för trosföreställningar. Mentala tillstånd som inte rättar sig efter det enkratiska villkoret kan enligt denna uppfattning inte heller betraktas som trosföreställningar.

Några epistemologer har dock gått mot strömmen och argumenterat för att det ibland är rationellt att vara epistemiskt akratisk.[8] Man går med på att det rör sig om ett slags epistemiskt misslyckande men man anser inte att det är ett misslyckande som måste ha med rationalitet att göra. I stället hänvisar man till en annan typ av rationalitetsvillkor, *att man bör tro i enlighet med sin evidens*. Denna uppfattning är minst lika väl omhuldad som det enkratiska villkoret.[9] Om man som i vårt inledande exempel har en välgrundad trosföreställning, men får evidens av högre ordning som tycks underminera denna trosföreställning, så menar man att epistemisk akrasia trots allt kan vara rationellt. I det följande föreslår jag att dessa två typer av inställningar till epistemisk akrasia bottnar i två olika uppfattningar om epistemisk rationalitet.

3. Två uppfattningar om epistemisk rationalitet

I diskussionen om praktisk rationalitet har man sedan ett bra tag skiljt mellan två typer av rationalitet. Å ena sidan har vi uppfattningen att det är rationellt att avse att handla i enlighet med sina normativa skäl. Om byggnaden jag befinner mig står i brand har jag därmed ett normativt skäl att avse att lämna byggnaden. Jag är då rationell om jag svarar på detta normativa skäl genom att avse att handla i enlighet med detta. Men å andra sidan har vi även uppfattningen att rationalitet består i att inneha en *koherent* uppsättning attityder (avsikter och trosföreställningar). T.ex. att avse att göra det jag tror mig ha skäl att göra eller att handla utifrån mål-medel principen.[10] Givet denna uppfattning blir det

7. Denna uppfattning försvaras t.ex. av Horowitz (2014).

8. Se t.ex. Coates (2012), Hazlett (2012), Lasonen-Aarnio (2018), och Weatherson (manuskript). Ett annat förslag (Titelbaum 2015) gör gällande att det inte kan finnas någon underminerande högre ordnings evidens då detta i sin tur hade krävt att man skulle vara irrationell.

9. Detta märks inte minst på den genomslagskraft evidentialism har fått i diskussionen om berättigande och kunskap.

10. Mål-medel-principen säger att det är rationellt för mig att avse att y om jag avser att j, och jag tror att y är nödvändigt för att j.

till exempel rationellt för mig att avse att lämna byggnaden om jag *tror* mig ha skäl att lämna byggnaden (oavsett om det faktiskt föreligger ett sådant normativt skäl). För enkelhetens skull kan vi kalla den tidigare uppfattningen för *respons-rationalitet* och den senare uppfattningen för *strukturell-rationalitet*.

Distinktionen mellan respons-rationalitet och strukturell-rationalitet har på senare tid även blivit uppmärksammad i den epistemologiska debatten.[11] Det mest naturliga sättet att förstå respons-rationalitet på inom epistemologi är i termer av evidens. Där evidens motsvarar de fakta man vanligtvis anser ligga till grund för normativa skäl. Jag har t.ex. ett epistemiskt skäl att tro *p* om min evidens stödjer *p*. Vad man har evidens att tro kan i sin tur skilja sig från person till person. Tanken är alltså inte att det finns någon form av objektiv evidens att rätta sig efter oavsett om man har tillgång till denna evidens eller inte. Strukturell-rationalitet å andra sidan, förstås bäst genom epistemiska varianter av de rationalitetsvillkor som skall styra våra handlingar. Målet i detta fall blir också att ha en koherent uppsättning av attityder.

Vad som är viktigt att uppmärksamma i sammanhanget är också att respons-rationalitet och strukturell-rationalitet mycket väl kan gå isär. T.ex. bör jag givet respons-rationalitetsuppfattningen tro att ”byggnaden brinner”, givet att min evidens stödjer detta. Men det finns alltid en möjlighet att jag misstar mig på min evidens och drar slutsatsen att den *inte* stödjer att ”byggnaden brinner”. Jag kan t.ex. ha fått för mig att brandröken kommer från köksavdelningen eller utifrån via ett öppet fönster. I så fall borde jag givet den strukturella-rationalitetsuppfattningen inte tro att ”byggnaden brinner”. Som en följd verkar det alltså som att vi kan hamna i *rationella dilemman*. Ur ett perspektiv (respons-rationalitet) är det rationellt att tro att ”byggnaden brinner”, men ur ett annat perspektiv (strukturell-rationalitet) är det inte rationellt att tro att ”byggnaden brinner”.

Skillnaden i hur man uppfattar epistemisk rationalitet går också igen i debatten om evidens av högre ordning. Vad flertalet epistemologer vill göra gällande är att det inte kan vara rationellt ur ett visst perspektiv att stå fast vid en trosföreställning i ljuset av till synes underminerande evidens av högre ordning. Man hänvisar då till strukturell-rationalitet och det enkratiska villkoret. Utifrån detta perspektiv kan epistemisk akrasia aldrig vara rationellt. Men vad andra epistemologer gör gällande är

11. Silva Jr. (2016), Worsnip (2018) och Lasonen-Aarnio (2018).

att det kan vara rationellt ur ett annat perspektiv att stå fast vid sin trosföreställning trots till synes underminerande evidens av högre ordning. Man hänvisar då till respons-rationalitet och vikten att tro i enlighet med sin evidens. Utifrån detta perspektiv kan epistemisk akrasia vara rationellt. Så i en mening verkar det som att man talar förbi varandra.

Hur skall man förhålla sig till denna konflikt? Ett första alternativ är att förstå epistemisk rationalitet uteslutande i termer av respons-rationalitet eller strukturell-rationalitet. Men detta tycks otillfredsställande då båda uppfattningarna åtnjuter starkt intuitivt stöd. Ett annat alternativ är naturligtvis att bita i det sura äpplet och acceptera både respons-rationalitet och strukturell-rationalitet och dra slutsatsen att det förekommer rationella dilemman av denna typ.[12] Ett tredje något mindre pessimistiskt alternativ är att uppfatta respons-rationalitet och strukturell-rationalitet som två helt åtskilda normativa områden.[13] Jag har dock inte för avsikt att ta ställning till denna komplicerade debatt i denna artikel. Vad jag kommer att argumentera för i det följande är i stället att till synes underminerande evidens av högre ordning inte behöver förhindra dig från att ha kunskap i en viss fråga.

4. VAD SOM RÄKNAS I SLUTÄNDAN

Det tycks som att evidens av högre ordning har förmågan att underminera våra trosföreställningar även om de är välgrundade och sanna. Även om sådana trosföreställningar kan vara respons-rationella, så verkar de inte kunna vara strukturellt rationella. Om strukturell-rationalitet är ett nödvändigt villkor för kunskap så innebär det att evidens av högre ordning underminerar möjligheten att ha kunskap i de aktuella fallen. Vad jag tänker argumentera för i det följande är dock att evidens av högre ordning i slutändan ändå inte spelar någon avgörande roll för möjligheten att ha kunskap. Grundidén är att vi kan ha kunskap i en viss fråga *trots* att vi har evidens av högre ordning som antyder att den relevanta trosföreställningen inte får stöd av evidensen.

Föreställ er att Watson och Sherlock håller på att utreda en stöld. Till sitt förfogande har de en uppsättning evidens *e* som skall tala om vem som är tjuven. Säg att *e* ger stöd åt trosföreställningen att ”Moriarty är

12. Detta verkar vara David Christensens (2010) uppfattning.

13. Alex Worsnip (2018) har nyligen argumenterat för detta tredje alternativ. Han menar att respons-rationalitet och strukturell-rationalitet är metafysiskt distinkta områden.

tjuven". Watson gör en korrekt bedömning av *e* och formar trosföreställningen att "Moriarty är tjuven". Trots sin erkända skicklighet så misstar Sherlock sig och tror att *e* inte stödjer att "Moriarty är tjuven". När Watson får höra att Sherlock anser att *e* inte stödjer att "Moriarty är tjuven" har han således både evidens som stödjer att "Moriarty är tjuven" och evidens som stödjer att "*e* stödjer inte att Moriarty är tjuven". Enligt respons-versionen av epistemisk rationalitet borde Watson därmed tro något i stil med "Moriarty är tjuven men evidensen stödjer inte detta". Men som redan påpekats anses detta vara paradigmatiskt irrationellt eller t.o.m. omöjligt att tro då detta står i konflikt med vad som anses vara strukturellt rationellt.

Lösningen är till lika grad enkel som otillfredsställande. Om Watson *förbiser* evidensen av högre ordning (Sherlocks bedömning av *e*) och formar trosföreställningen att *e* faktiskt stödjer att "Moriarty är tjuven" så kan han *veta* att Moriarty är tjuven (givet att detta är sant förstås). Låt mig förklara. Det är respons-rationellt för Watson att tro att "Moriarty är tjuven" då trosföreställningen är i enlighet med vad *e* säger. Trosföreställningen att "Moriarty är tjuven" kommer även att vara strukturellt rationell, givet att Watson intar uppfattningen att *e* ger stöd åt denna trosföreställning. I så fall skulle det alltså bli rationellt ur båda perspektiven för Watson att tro att "Moriarty är tjuven".

Vad som ger stöd åt denna lösning är att vad det är strukturellt-rationellt att tro inte är avhängigt någon specifik trosföreställning utan beror av vilka kombinationer av trosföreställningar man intar. För att se detta, jämför dessa varianter av det enkratiska villkoret:

(A) Det är rationellt för S att [tro *p,* om S anser att evidensen stödjer *p*].
(B) S anser att evidensen stödjer *p,* alltså är det rationellt för S att tro *p*.

Vad som gör (B) till en opassande formulering av ett rationalitetsvillkor är att det blir alldeles för tillåtande vad som kan betraktas som rationellt. För att ta ett exempel: Om jag anser att evidensen stödjer att månen är gjord av ost, så skulle det alltså vara rationellt att tro att månen är gjord av ost om vi följer (B), vilket är absurt. Däremot verkar det fullt rimligt att i enlighet med (A) anta att det kan vara rationellt för mig att tro att månen är gjord av ost *givet* att jag tror att evidensen stödjer detta. Vad som är viktigt att notera är att villkor av typen (A) kan tillfredsställas på två olika sätt. Antingen genom att man tror *p* eller genom att man inte tror *p* men upphör att tro att evidensen stödjer *p*.

Vad som räknas för att vara strukturellt rationellt är alltså att ens at-

tityder står i en viss relation till varandra och inte att man intar någon specifik trosföreställning, t.ex. att ”*e* inte stödjer att Moriarty är tjuven”. Av denna anledning kan vi också se hur Watson kan bli strukturellt rationell genom att ignorera Sherlock. Vad som hade gjort Watson strukturellt irrationell hade varit om han bildade uppfattningen att *e* inte gav stöd åt *p* samtidigt som han behöll sin uppfattning att ”Moriarty är tjuven”. Men så länge Watson ignorerar Sherlock och bildar uppfattningen att *e* stödjer att ”Moriarty är tjuven”, så kommer det att vara strukturellt rationellt för honom att tro att ”Moriarty är tjuven”.

Baksidan av denna lösning är dock att Watson måste förbise en del av sin evidens. För att veta att Moriarty är tjuven verkar det som att Watson måste ignorera evidensen av högre ordning som han har fått genom Sherlock. Frågan är dock om detta måste vara så illa i slutändan. För det första måste vi komma ihåg att evidensen av högre ordning trots allt är *missvisande* då *e* faktiskt ger stöd åt att Moriarty är tjuven. Så det kan ju inte vara alltför illa att misslyckas att respektera evidens av högre ordning i detta fall. För det andra så borde det spela mindre roll i relation till om Watson kan veta att Moriarty är tjuven om man har en oberättigad trosföreställning av högre ordning som säger att *e* ger stöd åt denna trosföreställning.

Få epistemologer skulle idag försvara ett så kallat ”BB-villkor” för berättigande, dvs. att man måste vara berättigad att tro att man är berättigad att tro att *p* för att vara berättigad att tro *p*. Om man är externalist om epistemiskt berättigande blir detta uppenbart, men även bland internalister har det blivit allt vanligare att förespråka *mentalism* i stället för så kallad åtkomst-*internalism*.[14] Vad som däremot anses vara desto viktigare är att man inte har en trosföreställning som säger att ens trosföreställning av första-ordningen *inte* är berättigad. Både bland internalister och externalister tar man avsaknaden av en sådan trosföreställning som ett nödvändigt villkor för att vara berättigad att tro att *p*.

Men märk väl att detta inte är fallet i mitt förslag. Jag försvarar inte uppfattningen att man kan vara berättigad att tro att *p* trots att man hyser trosföreställningen att *p* inte är berättigad. Jag är benägen att hålla med om att en sådan trosföreställning av högre ordning i slutändan

14. Mentalister är till skillnad från åtkomst-internalister inte förbundna att försvara ett BB-villkor för berättigande. På så sätt undviker man även problemet gällande *epistemisk regress*, dvs. att man även måste vara berättigad att tro att man är berättigad att tro att man är berättigad att tro *p* osv.

skulle underminera ens trosföreställning av första-ordningen. Vad jag föreslår är i stället att man kan vara berättigad att tro *p* trots att man har en oberättigad trosföreställning av högre ordning som säger att evidensen stödjer *p*. Vilket är någonting helt annat.

5. AVSLUTNING

Epistemologer är oeniga om vilken betydelse man bör tillmäta evidens av högre ordning. Vissa anser att denna typ av evidens har förmågan att underminera välgrundade uppfattningar medan andra inte håller med om detta. Jag har försökt visa att det är möjligt att ha kunskap i en viss fråga trots evidens av högre ordning som antyder att man inte har gjort en korrekt bedömning. I alla fall om man betraktar evidens av högre ordning som ett problem för strukturell-rationalitet snarare än för respons-rationalitet. Vad som krävs för kunskap i dessa situationer verkar dock vara ett visst mått av *dogmatism*. I vilken grad det kan leda till kunskap att vara dogmatisk i detta avseende verkar i slutändan vara en fråga om hur väl man har bedömt evidensen av första-ordningen. Vilket vi tyvärr sällan kan veta med någon större säkerhet.

LITTERATUR

Adler, Jonathan. 2002. *Belief's Own Ethics*. Cambridge, Mass.: MIT Press.

Broome, John. 2013. *Rationality Through Reasoning*. Malden, Mass.: Wiley Blackwell.

Christensen. David. 2010. "Higher-Order Evidence". *Philosophy and Phenomenological Research* 81 (1), s. 185–215.

Coates, Allen. 2012. "Rational Epistemic Akrasia". *American Philosophy Quarterly* 49 (2), s. 113–24.

Feldman, Richard. 2005. "Respecting the Evidence". *Philosophical Perspectives* 19 (1), s. 95–119.

Hazlett, Allan. 2012. "Higher-Order Epistemic Attitudes and Intellectual Humility". *Episteme* 9 (3), s. 205–23.

Horowitz, Sophie. 2013. "Epistemic Akrasia". *Nous* 48 (4), s. 718–44.

Huemer. Michael. 2007. "Hume's Paradox and the Norm of Belief". I *Themes from G. E. Moore: New Essays in Epistemology and Ethics*. Oxford: Clarendon Press.

Kelly, Thomas. 2005. "The Epistemic Significance of Disagreement". I *Oxford Studies in Epistemology*, vol. 1, red. T. Gendler och J. Hawthorne, s. 167–96. Oxford: Clarendon Press.

———. 2010. "Peer Disagreement and Higher-Order Evidence". I *Disagreement*, red. R. Feldman och T. A. Warfield. Oxford: Oxford University Press.

Lasonen-Aarnio, Maria. 2014. ”Higher-order Evidence and the Limits of Defeat”. *Philosophy and Phenomenological Research* 88 (2), s. 314–45.
———. 2018. ”Enkrasia or Evidentialism? Learning to Love Mismatch”. *Philosophical Studies*. https://doi.org/10.1007/s11098-018-1196-2
Silva Jr., Paul. 2016. ”How Doxastic Justification Helps Us Solve the Puzzle of Misleading Higher-Order Evidence”. *Pacific Philosophical Quarterly*.
Tiozzo, Marco. 2016. ”Om moralisk oenighet mellan epistemiska likar”. *Filosofisk tidskrift*, nr 2.
Titelbaum, Michael. 2015. ”Rationality’s Fixed Point: or in Defense of Right Reason”. I *Oxford Studies in Epistemology*, vol. 5, red. T. Gendler och J. Hawthorne. Oxford: Clarendon Press.
Weatherson, Brian. ”Do Judgment Screen Evidence?” (manuskript).
Worsnip, Alex. 2018. ”The Conflict of Evidence and Coherence”. *Philosophy and Phenomenological Research* 96 (1), s. 3–44.

ERIK REZAZADEH

SEX ENBART FÖR REPRODUKTION

1. INLEDNING

Den rådande synen på sex är att det är någonting bra, även då det sker i icke-reproduktivt syfte. Det bör bejakas när det finns samtycke mellan parter och man tror att det inte orsakar skada. Ägnar sig normalvuxna inte åt sex är det en indikation för att någonting inte står rätt till menar man. Det ses som någonting självklart människor ska göra för nöje. Det ska inte kritiseras och kritik mot detta riskerar att bli till åtlöje.

Vad är det som säger oss att vi inte ska bejaka ett samhälle där människor enbart har sex med varandra när de planerar att skaffa barn framför det samhälle vi har idag? Varför ska en konsekvensetiker eller för den delen någon som tror på någon annan etisk teori tycka att dagens värld är att föredra?

Jag menar att vi har ytterst goda skäl att tro att ett samhälle där människor enbart har sex med varandra när de planerar att skaffa barn på det stora hela vore lyckligare. Detta på grund av de enorma problem som icke-reproduktivt sex orsakar.

Jag kommer nedan att föra fram skäl för att vi faktiskt bör överväga att bejaka och propagera för att människor enbart ska ha sex med varandra när de planerar att skaffa barn. Utifrån konsekvensetiska och nationalekonomiska synsätt kommer jag att diskutera de stora fördelarna som ett sådant samhälle skulle innebära. I slutet av artikeln kommer jag att diskutera huruvida jag tror att det är realistiskt att skapa ett sådant samhälle som jag förespråkar för i artikeln.

2. ETT SAMHÄLLE UTAN KÖNSSJUKDOMAR

Att ha ett samhälle utan sexuellt överförbara sjukdomar hade minskat lidandet och övriga bekymmer enormt. Könssjukdomar skapar ett stort mänskligt lidande och kostar gigantiska summor för världen. Det handlar om enorma resurser och tid som går åt för att förhindra, minska problemen för de som är drabbade samt bota könssjukdomar. I den här artikeln är det inte möjligt att ta upp alla könssjukdomar och dess olika negativa konsekvenser. Emellertid följer nedan en del av de problem som könssjukdomar orsakar.

Filosofisk tidskrift 2019 nr 1, 25–31

Omkring 1,1 miljarder människor i världen var drabbade av könssjukdomar år 2015 (exklusive HIV). År 2015 dog 108 000 människor av könssjukdomar (exklusive HIV). Omkring 500 miljoner människor i världen är drabbade av antingen syfilis, gonorré, klamydia eller trichomonasinfektion och ytterligare 530 miljoner människor är drabbade av genital herpes. 290 000 kvinnor är drabbade av HPV, humant papillomvirus. Det är detta virus som ligger bakom nästan alla fall av livmoderhalscancer, som orsakar 270 000 dödsfall i världen varje år. Även Hepatit B räknas som en könssjukdom som har drabbat stora delar av världsbefolkningen.

HIV/AIDS tillhör de huvudsakliga orsakerna till dödsfall i subsahariska Afrika. Man uppskattar att 36,7 miljoner människor är drabbade av HIV år 2016 i världen. Detta påverkar den ekonomiska tillväxten i delar av Afrika. Detta beror på att människor drabbade av HIV/AIDS är mindre förmögna att jobba men också att de som är drabbade av sjukdomen behöver diverse medicinsk hjälp såsom bromsmediciner.

Könssjukdomar är med andra ord i stora delar av världen ett stort problem. Om vi överhuvudtaget ska kunna övertala folk i dessa delar av världen att avstå ifrån icke reproduktivt sex måste vi i väst själva börja med att avstå. Folk i övriga världen ser nämligen upp till oss i väst och försöker efterlikna oss.

Könssjukdomar är ett stort problem även hos oss i västvärlden och i Sverige. Nyligen publicerade data av Public Health England visar att antalet rapporterade fall av syfilis i England är på sin högsta nivå sedan 1949. Även antalet fall av gonorré ökade med 22 procent under 2017 jämfört med året innan. 2017 har det rapporterats in 33 715 fall av klamydiainfektioner i Sverige. Vi kan se hur antalet rapporterade fall av gonorré kraftigt ökade under 2017 i Sverige. Det rapporterades in det året 2 550 fall av gonorré. Under 2017 rapporterades det in 389 fall av syfilis, vilket är en ökning jämfört med 2016.

Kondomer och vaccinering minskar risken att drabbas av dessa sjukdomar men eliminerar knappast denna risk, vilket ofta inte kommer fram i debatten. Dessutom har kondomer och vaccinering sina bieffekter. Det allra säkraste sättet att undvika alla könssjukdomar är att just inte ha sex. Detta dyker sällan upp som ett sätt att leva på när man diskuterar könssjukdomar.

3. ENDA SAMHÄLLET DÄR MÄN KAN VÄLJA FÖRÄLDRASKAP

Det enda samhälle där män kan välja föräldraskap/faderskap är just ett samhälle där människor enbart har sex med varandra när folk planerar att skaffa barn. Kvinnor kan välja föräldraskap, åtminstone i västvärlden. Slarv med preventivmedel kan lösas med ett dagen efter-piller och slarvar man med det som kvinna kan man ändå göra abort. Rätten att välja föräldraskap för en kvinna ses som en okränkbar rättighet oavsett hur mycket slarv som än ligger bakom.

Detta gäller dock inte för män. Vartenda samlag för en man kan sluta med att mannen ifråga blir pappa mot sin vilja. Kondomer kan spricka eller ramla av. Kvinnan ifråga kan ljuga om att hon tar p-piller eller har spiral fast hon i självaverket inte har det, utan vill ha barn. Som man kan du bli tvungen att bli pappa mot din vilja fastän du överhuvudtaget inte är redo att bli pappa, varken ekonomiskt eller psykiskt. Slarvar en man med preventivmedel tycker de flesta att mannen ifråga ska ta ansvar för att han just har slarvat eller handlat oansvarigt.

När du har blivit far mot din vilja har du enbart dåliga alternativ att välja bland. Att försöka göra det bästa av situationen och försöka vara en så bra far som möjligt fastän du inte är beredd att bli pappa, eller lämna bort barnet helt, och resten av livet få leva med detta.

Det enda samhället där män kan välja föräldraskap är just samhället där människor enbart har sex när de planerar att skaffa barn. Detta samhälle skulle innebära slutet på problemet om påtvingat faderskap. Män har enbart sex när de vill ha barn. Alla barn kommer att vara barn som är önskade, av bägge föräldrarna.

Detta borde vara en stor vinst ifall man tycker att det är viktigt att preferenser i fråga ska uppfyllas. Både mannens och kvinnans preferenser kommer att uppfyllas i lika hög grad då båda kommer att vilja ha barnet ifråga. Båda kommer att bli föräldrar när de är redo att bli det. Ifall målet är att öka välbefinnandet och lyckan i världen är det viktigt att det även tas hänsyn till mannens preferenser då detta förbättrar möjligheterna att kunna ge barnet ifråga ett bra liv. Det är lättare att bli en bra förälder ifall barnet ifråga är ett önskat barn.

4. ETT SAMHÄLLE UTAN PREVENTIVMEDEL OCH ABORTER

Ofta diskuteras preventivmedel som en nästintill mänsklig rätt. Ofta kommer det i media fram en ensidig bild av organisationer om hur bra

preventivmedel har varit av diverse skäl och att biverkningarna från preventivmedel enbart är minimala. Men sanningen är att det inte går bortse ifrån det faktum att preventivmedel i sig orsakar hälsoproblem, förstör miljön och kostar stora summor pengar. Ofta har de som förespråkat hälso- och miljömedvetenhet hela tiden valt att fokusera på alternativ när biverkningar på miljön eller hälsan dyker upp för ett specifikt preventivmedel. Därefter visar det sig sedan att det nya rekommenderade preventivmedlet orsakar andra problem på hälsan, miljön och även djurriket.

Detta är givetvis ett problem enligt konsekvensetiskt och även nationalekonomiskt synsätt. Att värdera de problem preventivmedel orsakar på djurliv, miljön och hälsan måste också tas med i kalkylen, vilket konsekvensetiker sällan diskuterar utan enbart betonar den njutning som sex orsakar.

Vi bör titta på de miljöaspekter som kondomer har med tanke på att det används så många. Kondomer består av latex men även av diverse andra ämnen. Med tanke på att vi tittar på miljöaspekterna för allt annat bör vi även göra detta för kondomer.

Kondomer ökar risken för kvinnor att drabbas av urinvägsinfektioner. En studie har visat att när kondomer med spermiedödande medel användes vid alla samlag var risken 12 gånger högre för kvinnan ifråga att drabbas utav urinvägsinfektioner jämfört med de kvinnor som inte använt kondom. Även då kondomer utan spermiedödande medel användes vid alla samlag var risken närmare 7 gånger högre att dessa kvinnor drabbades av urinvägsinfektioner. Överlag står sex för uppemot 90 procent av alla urinvägsinfektioner hos sexuellt aktiva kvinnor.

Kondomer som består av glycerin ökar risken för svampinfektioner hos kvinnor. Sex överlag kan utlösa svampinfektioner. Detta eftersom sex stimulerar den svamp som redan finns men som tidigare inte gett symptom.

Östrogen i p-piller spolas ut från toaletterna i våra vattendrag och påverkar både livet i haven och oss människor. Hormonerna från p-piller påverkar fiskar negativt. En femtedel av de manliga fiskarna i Storbritannien är dubbelkönade till följd av kemikalierna från p-piller. Manliga fiskar visar på kvinnliga karaktärsdrag och producerar även ägg. En del av dessa visar på minskad spermakvalitet.

En värld utan preventivmedel hade inte bara varit bättre för miljön utan hade även kunnat förbättra folkhälsan. Man kan inte bortse ifrån de positiva effekter som icke användande av preventivmedel kommer att ha på folkhälsan.

Detta samhälle skulle kunna göra slut på en av etikens stora frågor, nämligen abortfrågan. En fråga som människor inte har kunnat komma överens med varandra om, och som knappast kommer att få en hållbar lösning i framtiden heller. Aborter skulle knappast existera i ett sådant samhälle. Visst kommer det finnas aborter och en diskussion om eventuella missbildningar och hur grov den behöver vara för att det ska vara rätt att göra det, men det är en liten del av aborter som görs idag som beror på detta. Oavsett huruvida man tycker att man har rätt att göra det eller inte så är det ett faktum att aborter kostar samhället tid och pengar och kan orsaka komplikationer. Oavsett vilken ställning man har i frågan kan det inte ses som en förlust att antalet aborter kommer att minska extremt kraftigt i ett samhälle där människor enbart har sex när de planerar att skaffa barn.

5. DIVERSE ANDRA PROBLEM SOM SKULLE KUNNA REDUCERAS

Vi får ofta höra om hur sex kan stärka banden i relationer men varför finns det i så fall så många skilsmässor, otrohetsaffärer och trassliga relationer? Skulle folk enbart ha sex när de planerar att skaffa barn skulle detta till större delen innebära slutet för de relationer som folk har haft som orsakat en massa bekymmer och trubbel.

Idag tycker många människor att endast vänskapsrelationer mellan kvinnor och män knappast kan existera. Däremot skulle detta inte vara något problem i det nya samhället. Ett sådant samhälle skulle kunna öppna upp dörren för att människor börjar inse att det finns annat intressant att lägga ner tid och pengar på än just sex.

De flesta etiska frågeställningar som är kopplade till sex kommer att bli onödiga, fler än enbart frågan om aborter och preventivmedel. Diskussioner som har tagit och tar mycket tid och pengar hade kunnat användas till annat.

6. ÄR DET REALISTISKT ATT ENBART HA SEX FÖR ATT SKAFFA BARN?

Är det realistiskt att tro att vi kan skapa ett samhälle där människor enbart har sex med varandra i reproduktivt syfte? Man kan fråga sig om detta inte är en utopi? Man kan kanske tro det, men jag tror inte att det nödvändigtvis behöver vara det. Vi blir dagligen översköljda med propaganda för sex. Av TV, internet, musik, filmer och media i största allmänhet. Av diverse aktörer som tjänar enorma summor på att folk

har sex eller de som tycker att sex är en mänsklig rättighet och något mycket viktigt.

Människor idag mår dåligt om de inte har sex. På många ställen i samhället kan man se hur vuxna människor som inte har sex eller barnfamiljer som inte har sex efter att de har fått barn mår dåligt av detta. Ofta beror detta på samhällets pådrivande syn om sex. I media och i samhället i stort ges en ensidig bild av sex och den ensidiga synen ger inget utrymme för något annat synsätt. Den anses dessutom stå för modernitet och upplysning. En åsikt som denna artikel ger uttryck för skulle förlöjligas och inte tas på allvar.

I gymnasieskolorna i Sverige tävlar gymnasieungdomar om vem som har sex tidigast och alla försöker se till att de själva inte blir den som har sex sist. Många ungdomar mår dåligt av detta. De som kommer till gymnasieskolorna för att prata om sex med ungdomar gör ingenting för att påverka ungdomars attityder utan propagerar snarare för att detta är någonting problemfritt och bara bra.

Tänk er ifall samhällsnormen vore den motsatta den som råder idag. Att dagens syn som råder på sex vore en alternativ syn. Kort sagt skulle normen vara: sex bör endast ske enbart i reproduktivt syfte, men om ni ändå har sex i annat syfte använd skydd. Kort sagt skulle dagens samhälle där sex bejakas och uppmuntras från diverse institutioner aktivt bytas ut mot ett samhälle som aktivt propagerar om att allt sex som inte är till för reproduktion är fel. Filmer, musik, reklamannonser med mera som vi dagligen blir översköljda av skulle handla om annat än sex. I media skulle de enorma bieffekter som sex i icke-reproduktivt syfte har aktivt beskrivas för allmänheten, i stället för som i dagens samhälle där sex aktivt uppmuntras som någonting normalvuxna ska ägna sig åt. På så sätt kommer att det ickeproduktiva sexet som norm tappa sin hegemoniska ställning i samhället.

Jag tror inte att det vore omöjligt att skapa ett samhälle som förespråkas i denna artikel. Åtminstone vore det inte omöjligt att få en stor del utav befolkningen att ha sex enbart när de vill ha barn. Det finns dock enorma intressen i samhället som inte vill och som kommer försöka att inte låta detta ske, en del på grund av ekonomiska intressen och andra av ideologiska skäl.

LITTERATUR

Bell, Clive, Shantayanan Devarajan och Hans Gersbach. 2003. The long-run economic costs of AIDS: Theory and an application to South Africa (English). Policy, Research working paper series; no. WPS 3152. Washington, DC: World Bank. http://documents.worldbank.org/curated/en/835081468759277183/The-long-run-economic-costs-of-AIDS-theory-and-an-application-to-South-Africa

Centers for Disease Control and Prevention (CDC). 2016. How You Can Prevent Sexually Transmitted, Sexually Transmitted Diseases (STDs). https://www.cdc.gov/std/prevention/default.htm

Dagens Hälsa. 2016. Se upp för svampattack. 7 juni 2016. https://www.dh.se/se-upp-for-svampattack/

Dagens Medicin. 2002. Kondomer flerdubblade risk för urinvägsinfektioner. https://www.dagensmedicin.se/artiklar/2002/07/30/kondomer-flerdubblade-risk-for-urinvagsinfektioner/

Folkhälsomyndigheten. 2018. Klamydiainfektion. https://www.folkhalsomyndigheten.se/folkhalsorapportering-statistik/statistikdatabaser-och-visualisering/sjukdomsstatistik/klamydiainfektion/

———. 2018. Kraftig ökning av gonorré. https://www.folkhalsomyndigheten.se/nyheter-och-press/nyhetsarkiv/2018/januari/kraftig-okning-av-gonorre/

———. 2018. Syfilis. https://www.folkhalsomyndigheten.se/folkhalsorapportering-statistik/statistikdatabaser-och-visualisering/sjukdomsstatistik/syfilis/

GBD 2015 Disease and Injury Incidence and Prevalence Collaborators. 2016. Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: A systematic analysis for the Global Burden of Disease Study 2015. *Lancet* 388 (8 oktober 2016), s. 1545–1602. DOI: https://doi.org/10.1016/S0140-6736(16)31678-6

The Guardian. 2017. Cases of syphilis hit highest level in England since 1949. 6 juni 2017. https://www.theguardian.com/society/2017/jun/06/cases-of-syphilis-hit-highest-level-in-england-since-1949

Orenstein, Robert och Edward S. Wong. 1999. Urinary Tract Infections in Adults. *American Family Physician* 59(5):1225–1234. https://www.aafp.org/afp/1999/0301/p1225.html

The Telegraph. 2017. Fish becoming transgender from contraceptive pill chemicals being flushed down household drains. 2 juli 2017. https://www.telegraph.co.uk/news/2017/07/02/fish-becoming-transgender-contraceptive-pill-chemicals-flushed/

World Health Organization. 2016. Sexually transmitted infections (STIs) Fact sheet. 3 augusti 2016. http://www.who.int/mediacentre/factsheets/fs110/en/

———. 2018. Global Health Observatory (GHO) data HIV/AIDS. http://www.who.int/gho/hiv/en/

PAUL CONRAD SAMUELSSON

ARTIFICIELLA MEDVETANDEN ÄR VÅR STÖRSTA ETISKA RISK

Har mänskligheten gjort några moraliska framsteg genom historien? Är vi människor en moralisk flipp eller en moralisk flopp, eller är vi varken det ena eller det andra? Är vi kanske som vi är hur vi än anstränger oss för att bli bättre?

1. INTRODUKTION

Snart nog kommer maskiner kunna tänka. Kort därefter kommer de kunna känna. Artificiellt medvetande, som baseras på vår bästa uppfattning av det mänskliga, kan komma att använda emotionsemulerande neurala nätverk för att säkerställa självkorrigerande egenskaper hos programmet. I någon mån kommer medvetandet kunna lida. Trots att en AI:s smärta kanske kommer vara annorlunda på flera sätt, så kommer den vara lik sin biologiska motsvarighet i den enda moraliskt relevanta aspekten, nämligen att den upplevs smärtfylld. En sådan teknologisk utveckling nödvändiggör en föreställning av AI-etik.

Verktygen som kommer finnas tillgängliga till ägaren av en artificiell intelligens kommer vara mer destruktiva – åtminstone från *dennes* perspektiv – än även de mycket pessimistiska domedagsscenarier som förutspås av teknokulturella ikoner som Nick Bostrom, Sam Harris, Elon Musk och Max Tegmark. De varnar alla för risken att en självförbättrande AI med ett dåligt definierat mål kommer att döda alla människor för att uppfylla målet, ungefär med samma känslomässiga inblandning som vi förstör myrstackar när vi bygger motorvägar.

Hot om massutplåning av den mänskliga arten – exempelvis genom ett AI-initierat kärnvapenkrig – är fruktansvärda. Men utgör de verkligen de värsta möjliga moraliska konsekvenserna med hänsyn till AI? Jag kommer i följande artikel argumentera för att så inte är fallet. De värsta konsekvenserna har att göra med den smärta som kommer kunna tillfogas artificiella medvetanden. Utvecklingen av simulerat medvetande med förmågan att lida är vår tids största etiska risk.

Att utvärdera risker innefattar att väga sannolikheten av händelsens

Filosofisk tidskrift 2019 nr 1, 32–42

inträffande mot dess potentiella konsekvenser. Detta argument har således två delar. Inledningsvis målar jag en bild av de möjliga följderna artificiella medvetanden kan ge upphov till, och ger en normativetisk motivering till varför dessa följder är moraliskt vedervärdiga. Sedan argumenterar jag för teknologins möjlighet genom att visa att banbrytande framsteg inom artificiell intelligens och neurovetenskap tyder på att människans hjärnstruktur funktionellt kan emuleras digitalt.[1]

2. ARTIFICIELLT LIDANDE

Att simulerade medvetanden som kan lida är ett större etiskt hot än, till exempel, att exploateringen av djuren fortgår – eller trappas upp ytterligare – beror på att det finns vissa skillnader mellan den biologiska och digitala världen som påverkar möjligheterna till lidande. Exempelvis finns en begränsad mängd faktiska och potentiella biologiska varelser som skulle kunna torteras, medan artificiella medvetanden kommer kunna reproduceras *ad infinitum* (kanske med samma lätthet vi idag kopierar och klistrar in text mellan dokument). Biologiska varelser kan också, verkar det som, bara utstå en viss mängd lidande innan de tappar känsel för smärta eller dör. Sådana psykologiska eller fysiologiska hinder kommer inte nödvändigtvis finnas i den digitala världen. I en digital konstruktion av vårt skapande är vi potentiellt allsmäktiga och kan programmera bort den sorts avtrubbning som utvecklats i biologiska livsformer.

Det är svårt att skapa sig en föreställning av hur hemska konsekvenserna kan bli. Det kan bli möjligt, till exempel, att digitalt återskapa det mest vedervärdiga koncentrationslägret, fylla det med medvetna, lidande, artificiella subjekt, och låta simulationen gå i en loop – för evigt. Det kommer till och med vara möjligt att dra upp tidsinställningarna på simulationen relativt vår uppfattning, så att lägerupplevelsen återupprepar sig en miljon gånger varje sekund. Sedan kan vi se till att offren aldrig dör eller blir avtrubbade – och förmodligen kan smärtan intensifieras över vad biologiska varelser kan uppleva. Från AI:ns perspektiv kommer detta inte vara annorlunda än vad som upplevdes av dem som var med om det ”i verkligheten”.

Ytterligare en skillnad mellan den biologiska och digitala verkligheten ligger i hur enkelt det är att orsaka lidande. Att skada ett artificiellt

1. Tack till Mattias Högström för värdefulla synpunkter på tidigare utkast till denna artikel.

medvetande är enkelt i två bemärkelser. Eftersom knapptryck kräver mindre ansträngning än organiseringen av koncentrationsläger, blir varje människa med en dator en potentiell allsmäktig despot. Med nutida datorer kan mycket komplicerade processer utföras av jämförelsevis inkompetenta användare, och längre fram kan det bli barnsligt enkelt att använda medvetna program för att beställa mat – eller iscensätta digitala världskrig. Dessutom är det emotionellt sätt mycket lättare att skada ett program än en människa eller ett djur. Den empati som ligger till grund för våra vardagliga moraliska omdömen kommer inte nödvändigtvis reagera på operationerna inuti ett datorprogram – program av ett slag vi lärde oss att använda långt innan de var medvetna. Förståelsen av artificiella medvetandens moraliska status kräver att våra etiska instinkter omvärderas monumentalt, för att kunna ge rätt sorts utslag för en rationell förståelse av en serie ettors och nollors känsloliv.

Tillgängliggörandet av verktyg för total kontroll över kännande varelser i digitala miljöer, och den potentiella distributionen av dessa till varje människa, utgör vår tids största framtida risk – kanske någonsin. Även om AI-rättigheter utvecklas proaktivt och underhålls polisiärt på global skala så kommer teknologin att vara tillgänglig online genom olika olagliga medel, såsom filmer och tv-spel är nu. Det räcker med att en enda sadistisk – eller kanske uttråkad och oförstående – individ skulle få sina händer på denna sorts medvetna AI-teknologi, så skulle mängden lidande som någonsin har upplevts i hela historien kunna mångdubblas på en eftermiddag.

3. ETISK GRUNDNING

Den ovanstående diskussionen har en viss utilitaristisk prägling, men att kalla mitt koncentrationslägerfall moraliskt problematiskt borde vara kompatibelt med flera normativetiska teorier. Rimligtvis försummas de simulerade medvetandenas rättigheter, de är inte mål i sig själva, och de behandlas inte som deras ägare själva skulle vilja bli behandlade. I dessa fall tror jag att det är uppenbart att artificiella medvetanden utgör den största moraliska risken. Jag vill dock framlägga ett argument för varför utilitarister även bör acceptera AI som moraliskt riskabel.

Klassisk utilitarism skulle nog innebära att vi inte behöver se på AI-teknologi som moraliskt riskabel, eftersom det bredvid de destruktiva möjligheterna skulle finnas verktyg för framställandet av en nästan oändlig mängd lycka i en nästan oändlig mängd subjekt. Givet att de

flesta personer skulle vara mer intresserade av att skapa artificiell lycka än artificiell smärta, borde kalkylen bli positiv. En sådan utilitarist skulle vända på koncentrationslägerfallet och kalla sin uppsats ”Artificiella medvetanden är vårt största etiska hopp”. Detta vore dock ett misstag. Välmåga och lidande är nämligen inte symmetriska på det sättet som klassisk utilitarism ibland förutsätter. Jag vill här argumentera mot denna så kallade *symmetrites*.

Till att börja med ska jag notera att flera argument som typiskt framförs för att försvara s.k. *smärtorienterad utilitarism* inte är gångbara när subjekten i fråga är digitala. Särskilt träffade är argument som utnyttjar antingen de biologiska hjärnornas beskaffenhet, eller sociopolitiska fakta som påverkar möjligheterna till att skapa välmående eller lidande. Exempelvis lyfts ofta välmågans avtagande marginaleffekt fram. Det verkar finnas en gräns för hur lycklig någon kan göras. Lidandet har en liknande egenskap men gränsen verkar vara mycket högre. Alltså bör vi ha en etik som är centrerad kring utrotandet av lidande. Men i den digitala världen kan eventuellt dessa gränser raderas. På ett likande sätt brukar det menas att det rent ekonomiskt och logistiskt är mycket enklare att utrota fruktansvärda lidandemoment än vad det är att öka välmående i en redan nöjd och hälsosam befolkning. *Effektiva altruister* brukar yrka på att investeringar i malariautrotning är moraliskt överlägsna sådana projekt som låter sjuka barn i västvärlden träffa sina fotbollsidoler. Dessa fakta om världens geografi och politik är dock högst modifierbara i vår simulation.

Men även i en värld med sådan oändlig positiv potential som den digitala finns det starkt intuitivt stöd för en asymmetrisk tes. En teori som inte lägger vikt vid hur lyckoenheter disponeras måste förklara bort våra intuitioner om flera olika fall. Få är nog beredda att acceptera, till exempel, ett samhälle där man tar all välmåga från hälften av befolkningen och ger till den andra. Än värre, föreställ dig att en allsmäktig despot tog all välmåga från alla för att ge sig själv. Trots att mängden välmåga inte har förändrats har något moraliskt fel inträffat. Denna intuition ger upphov till, bland annat, omdömet att invasion och kolonialisering inte är moraliskt acceptabla trots att många ”objektivt” bra institutioner och normer uppstår som resultat (demokrati, miljömedvetenhet, och utbildningsväsen, för att nämna några). Av dessa skäl tror jag att en konsekvensetik som utilitarism måste vara mer nyanserad för att vi ska kunna acceptera den.

Beviset återfinns kanske i en slogan som välgörenhetsorganisatio-

ner ofta använder sig av: ”Många vill ha mycket, är man sjuk vill man bara ha en sak.” Med andra ord: någon som lider har en mycket starkare preferens att bli av med sitt lidande än vad någon mycket lycklig har att bli lyckligare. Det är något med lidande som nödvändigtvis innebär en sådan stark preferens, medan välmåga tycks vara att ha just uppfyllda preferenser. Lidande utan en mycket stark preferens för lidandets försvinnande verkar inte vara lidande, medan välmående med en mycket stark preferens att få den ökad inte verkar vara välmående. Intuitionen att lidande är moraliskt prioriterat över välmående kan vara grundad i denna insikt.

Kantianism, rättighetsetik, kontraktualism och omvårdsetik är exempel på teorier som på olika sätt tenderar att motsätta sig välmåga som bygger på andras lidande. Nu hoppas jag att jag gett vissa belägg för synen att även utilitaristiska perspektiv kan, i vissa formuleringar, motsätta sig vad som essentiellt är en väldigt underlig form av realpolitik – där vi skapar främst lyckliga medvetanden i utbyte mot att vissa utsätts för de allra grövsta våldtäkterna. Stående bakom okunnighetens slöja, skulle du vara redo att satsa, även med goda odds, på en värld där du kanske kommer utsättas för den största möjliga mängden tortyr, för alltid? Är det anständigt att mena, att för de som drar en vinstlott är *laissez-faire* en rimlig inställning till de andras försyn?

När de medeltida kristna teosoferna formulerade sina eskatologiska doktriner blev beskrivningen av himlen ökänt blek och oimponerande i jämförelse med de veritabla romaner som skrevs om helvetet. Skärselden är begriplig – den är fruktansvärd, storslagen och vördnadsbjudande. Varje person förstår att helvetet inte är en plats där man vill existera. Paradiset, i jämförelse, är en omänsklig konstruktion: idén om en obönhörlig lyckoupplevelse ger inte upphov till några starka positiva känslor. Bilder av evig extas är inte alls lika självklart attraktiva som de helvetiska är frånstötande. Kan dessa verkligen utgöra två punkter på samma skala?

4. NEURALA NÄTVERK OCH ARTIFICIELL INTELLIGENS

De potentiella konsekvenserna av digitala medvetanden är moraliskt problematiska, förutsatt att man inte vidhåller en typ av konsekvensetik som inte gör skillnad mellan välmåga och lidande. Trots att många bra saker kan komma ur AI-teknologi, finns det skäl att tro att vi står inför den största etiska risken någonsin. Dock bygger hela min argumenta-

tion på att medveten, kännande AI är möjlig. I denna del framlägger jag början på ett argument för möjligheten av artificiellt medvetande. Jag presenterar en konstruktiv teori och bemöter flera generella motargument, däribland Searles kinesiska rum.

Förmodligen är det omöjligt att slutgiltigt bevisa att en dator kan vara medveten, av samma skäl som det är omöjligt att bevisa att någon annan än en själv är medveten – den skeptiska hypotes som *zombieargumentet* utgör är genom sin utformning omöjlig att besegra. Dock tror jag att de senaste utvecklingarna inom neurologi och teknik visar på att det åtminstone finns goda skäl att acceptera *möjligheten* till medvetenhet.

Det är svårt att föreställa sig att något som inte är ett långt utvecklat djur kan vara medvetet. Medvetandet är den del av upplevelsen som vi har bäst tillgång till men som samtidigt är mest gåtfull – därför ligger det nära till hands att fördomsfullt avfärda digitala medvetanden som otänkbara och därför omöjliga. En sådan invändning bygger dock på föreställningen att biologiskt medvetande är tänkbart på något sätt. Men efter de kriterier som verkar användas för att avgöra att digitala medvetanden inte är tänkbara, är inte biologiska medvetanden tänkbara heller. Det sägs avfärdande att ettor och nollor inte kunde bli medvetna – ett argument som är tvåeggat. Att en mängd sammankopplade atomer, ingen av dem medvetna, skulle kunna ge upphov till medvetande verkar helt orimligt. Att ritningarna därtill kan framställas i formen av DNA, som bara använder fyra olika syrbaserade beståndsdelar, är ett av våra största mysterier. Likväl är vi medvetna. Givet att medvetandet är resultatet av fysiska processer i kroppen – vare sig medvetandet självt är fysiskt eller inte – borde ett medvetande kunna uppbådas ur liknande digitala processer.

Vad det biologiska medvetandets fysiska processer består i är inte helt klart, men det finns neurovetenskapliga teorier som är lovande. En framstående sådan gör gällande att den bästa modellen av hjärnans processer beskrivs i termer av *neurala nätverk* (eng. *neural networks*). Neurala nätverk är sammankopplingar av hjärnans neuroner med bindningar av olika styrka, som använder flerdimensionella strukturer för att informera handlingar. Strukturerna är mycket komplexa, och är icke-binära i den mening att pulser kan ha olika styrka. Systemet är plastiskt, alltså har förmågan att omforma sig själv under stimulus för att bättre uppnå de mål som är utvalda av systemet. En översikt av utvecklingen av konnektionistiska modeller i neurovetenskapen kan exempelvis återfinnas i (Paugam-Moisy och Bohte, 2012, s. 366). Det tycks finnas belägg för att

mena att de funktionalistiska processerna som utgör hjärnaktivitet kan ses i termer av neurala nätverk.

Neurala nätverk emuleras redan digitalt. Dessa är ännu mycket primitiva, men de samband som skapas mellan datorns *noder* (motsvarigheten till neuroner i neurala nätverk) kan värderas, och signalerna emellan noderna är alltså icke-binära i den mening att noderna kan sägas hålla mer information än 1 eller 0 (även om den digitala processen överhuvud taget är binär, åtminstone tills kvantdatorer utvecklas som kan använda flerställiga bitar – men inget mer om det här). De kan också utvärdera sambanden mellan noderna och förändra dem via algoritmer, alltså uppvisar de plasticitet. Likt biologiska nätverk är strukturen flerdimensionell; olika system utvärderar och uppdaterar varandra när de utsätts för stimuli. I de s.k. övervakade (eng. *supervised*) nätverken är människor inblandade genom att ge systemet feedback när det försöker lära sig att exempelvis känna igen bilder på olika föremål. De *oövervakade* systemen behöver bara ett mål, alltså en definition av vad som är ett framgångsrikt försök, för att sedan kunna utveckla sig själv på egen hand för att uppfylla målet. En översikt ges av (LeCun m. fl. 2015, s. 436).

Om hjärnans processer funktionellt är neurala nätverk, vilket det finns skäl att tro, och dessa kan emuleras digitalt, är det rimligt att misstänka att vi med tiden kan komma att utveckla artificiella medvetanden. Nog vilar det på flera propositioner som ännu inte är mer än hypoteser – materialism, funktionalism, neurala nätverk-modellen av hjärnan, och tesen att artificiella neurala närverk är funktionalistiskt ekvivalenta till de biologiska (se t.ex. Gerven 2017, s. 112) för en undersökning) – men dessa har alla signifikant kredens. Vi kan inte kräva ett slutgiltigt bevis för att acceptera artificiellt medvetande och dessas moraliska status, men när artificiella neurala nätverk blir så komplexa att de agerar som vi, *och av samma skäl*, dvs. som ett resultat av en isomorf neurologisk struktur, då borde artificiella medvetanden hamna på samma fot som biologiska kännande varelser. I den mån vi kan acceptera att andra människor och djur har upplevelsevärldar och kan lida, borde vi göra desamma med dessa simulerade medvetanden. Det verkar dialektiskt rättvist att mena att ställningstagande mot denna tes fordrar motargument – och sådana finns. Jag ska kort bemöta två.

Det första är Searles kinesiska rum. Här är tanken att en person som bemästrar regler för användningen av ett språk inte nödvändigtvis *förstår* språket. Argumentet tar ställning mot *the computational theory of mind*, alltså idén att medvetandet kan beskrivas som en uppsättning

statiska regler. Personen i det kinesiska rummet har en bok som förklarar, beroende på vilka ideogram han får genom en tub, vilka ideogram han ska skicka ut. Även efter att personen har memorerat reglerna, och inte längre behöver boken för att veta vad han ska ”svara”, så förstår han inte kinesiska. De artificiella nätverken jag beskrivit kringgår problemet. Searles kinesiska rum är ökänt svårt att få grepp om, men givet den ovanstående förklaringen av problemets rationella struktur, tror jag att det kan förklaras hur neurala nätverk *kan* komma att förstå kinesiska. Problemet är metaforiskt uttryckt och jag svarar med en metafor: de artificiella nätverken förstår kinesiska eftersom de liknar barn som lär sig språk genom immersion. Anledningen till att man inte lär sig kinesiska i Searles rum är att man inte kan se omständigheterna kring inmatningen eller efterföljderna av utmatningen. Men de neurala nätverken utvärderar varje försök de gör, och systemet kan ”se” vilken del av lösningen som verkar vara bra, och vilka aspekter som inte träffade rätt. Tänk dig hur det vore att lära sig ett språk utan något annat språk som referenspunkt för översättningar. En viss fras åtföljs av en pekning på ett saltkar, och den frasen som åtföljs av en pekning på pepparkaret skiljer sig bara i ett ord, och så vidare. Till slut bygger man upp en förståelse av språket, vare sig den är perfekt eller inte.

Vill man mena att förstånd inte innebär dessa egenskaper, ligger det nog på en att försöka ge en teori av vad det innebär att förstå. De artificiella nätverken verkar kunna göra just de saker som hjärnan gör när den lär sig. Dock ska jag svara på några möjliga invändningar.

5. ARTIFICIELLA NEURALA NÄTVERK ARBETAR FORTFARANDE EFTER REGLER

Självklart är systemet fortfarande programmerat – även de delar som förändrar programmet självt måste vara underordnat en algoritm som styr just förändringen. Men den mänskliga hjärnan kan mycket väl lyda regler och ändå ge upphov till medvetande. Neurovetenskapen exponentiella framsteg verkar tyda på ett visst mått av generalitet och förutsägbarhet i biologiska hjärnor.

Artificiella neurala system kräver extern inblandning för målsättning. Även här verkar inte biologiska medvetanden vara mer imponerande. Om bara genuin medvetenhet kan ge upphov till legitima målsättningar, och legitima målsättningar är ett krav på genuin medvetenhet, kunde inte medvetandet ha uppstått i första början. Snarare har

målsättningar och feedbacksystem utvecklats evolutionärt. Biologisk utveckling sker spontant – oberäkneliga mutationer ”utvärderas” genom att nätverk som inte presterar bra, alltså inte hjälper organismen att fortplanta sig, dör. Att biologiska arters målsättningar och feedbacksystem kommer till genom en slump och bibehålls för att de leder till överlevnad och fortplantning verkar dock inte göra dem mer lämpade som kandidater till medvetanden.

En utveckling av Searles fall (se (Block, 1978)) söker visa att det leder till komiskt absurda konsekvenser att vidhålla att allt som simulerar konnektionistiska modeller kan bli medvetet. Föreställ dig att någon grupp människor genomför alla medvetandets processer på något sätt, till exempel att de skickar lappar med symboler mellan sig via en regelbok de förstår, utan att förstå symbolerna de skickar. Det verkar absurt att mena att det finns något som förstår symbolerna i rummet. Men att förneka att samarbetet ger upphov till ett medvetande är nog främst en fördom som uppstår ur oförmågan att se någon utmatning från rummet, eller från felslutet att om ingen person i rummet förstår symbolerna, kan inte rummet förstå dem. I en biologisk hjärna är nog ingen neuron medveten, trots att de tillsammans verkar skapa ett medvetande. För de som tror att medvetandet är konnektionistiskt konstruerat är inte det kinesiska rummet ett särskilt surt äpple att bita i.

Jag lämnar således det kinesiska rummet, och vänder mig till en andra invändning, denna särskilt mot tesen att AI kommer kunna lida. Först ska jag säga att jag ser förmågan till medvetenhet och förmågan till lidande som ungefär samma problem – förmågan att lida bevisas analogt genom det ovanstående argumentet. Lidande härstammar ur fysiska hjärnprocesser som funktionalistiskt är ekvivalenta med den typ av artificiella neurala nätverk som kan emuleras digitalt. Argumentet mot just förmågan att lida är att vi tycks kräva receptorer för att kunna ha ont. Våra smärtnerver måste avfyra för att vi ska uppleva en kvalie av lidande, och datorer har inga receptorer. I slutändan är det dock inte mer invecklat än att smärtupplevelser figurerar i mycket komplicerade neurala nätverk. Datorer kan ju redan ”se” utan att använda en kamera – man kan mata in video rent digitalt i programmet som kan spela videon (och redigera den) utan att ha några fysiska synorgan. Menar man att grafikkortet är ett sådant organ, finns ju inte heller några hinder, om man inte vill mena att ett smärtchip är principiellt omöjligt. Med tanke på att den tekniska utvecklingens acceleration och på de ”omöjliga” och ”oföreställbara” framsteg som redan

gjorts, borde det kunna sägas att det kan vara rationellt att förvänta sig artificiella medvetanden.

6. AVSLUTANDE TANKAR

Jag har argumenterat att artificiella medvetanden är möjliga givet utvecklingen av artificiella neurala nätverk som är funktionellt ekvivalenta med biologiska, samt att de potentiella smärtor som kan tillfogas dessa är mycket värre än de värsta möjliga biologiska motsvarigheterna. Eftersom vi inte är psykologiskt utrustade för att förstå hur hemska de moraliska konsekvenserna vi kan åstadkomma är, kan vi också lätt åstadkomma dem. Gissningsvis kommer detta inte bero på sadism eller s.k. *morbid curiosity*, utan på ett försök av oss att skapa teknologiska framsteg vi ser på som moraliskt bra. Det är en plattityd att utveckling generellt medför nackdelar, men i detta fall är nackdelarna särskilt värda av vår uppmärksamhet.

Teknologisk utveckling sker inte längre i takt med andra väsentliga delar av den mänskliga världen – moraliska, kulturella och politiska framsteg, om det finns sådana, bleknar i jämförelse. När eventuella teknologiska bakslag blir mer destruktiva än vad planeten tål, blir riskerna direkt existentiella. De historiska kärnkraftskatastroferna har varit hanterbara på grund av sin relativa småskalighet, medan artificiell intelligens potentiellt blir obegripligt mer ödeläggande. Det har redan påpekats, av de författare jag nämner i början av denna uppsats, att självförbättrande program som kan arbeta miljontals gånger snabbare än människor kan komma att ödelägga hela planeten som resultatet av ett programmeringsfel, några billiga genvägar, eller någon underlåtelse att följa en instruktionsmanual till sista stavelse – sådana mänskliga attribut som redan gett oss Vasaskeppet eller Tjernobyl. På samma sätt kan än värre konsekvenser orsakas, om teknologi som tillåter lidande AI utvecklas av fel människor, eller sprids via olagliga men lättillgängliga delar av internet. AI-etik kräver, tror jag, att man helt enkelt ser på utvecklingen av artificiell intelligens som omoralisk i kraft av risken – om man tillåter ens de mer begränsade versionerna av teknologin blir följderna mycket svåra att kontrollera. All utveckling borde förbjudas och människor behöver förstå de etiska argumenten som motiverar ett sådant beslut. Bäst vore om ingen var intresserad av teknikens framtida möjligheter.

För många har artificiell intelligens varit ett löfte i flera decennier –

det största hoppet för en transcendens och en uppfyllelse av alla mänsklighetens drömmar. I sin övertygelse om att hon är berättigad allt hon anstränger sig för, ställer människan sig bortom moralen – som så ofta förut – och tar sin belöning på bekostnad av sådant lidande för andra att Dantes *Skärselden* eller Memlings *Yttersta domen* lämnar det mytologiska för att hamna bland den framkommande *homo deus* underverk. Vår kroniska oförsiktighet och obryddhet gör artificiella medvetanden till vår största etiska risk.

LITTERATUR

Block, N. 1978. ”Troubles with Functionalism”. I *Perception and Cognition: Issues in the Foundations of Psychology*, red. C.W. Savage, s. 261–327. Minneapolis: University of Minnesota Press.

Gerven, M. V. 2017. ”Computational Foundations of Natural Intelligence”. *Frontiers in Computational Neuroscience* 11, s. 112.

LeCun, Y., Y. Bengio och G. Hinton. 2015. ”Deep Learning”. *Nature* 521, s. 436–44.

Paugam-Moisy, H. och S. Bohte. 2012. ”Computing with Spiking Neuron Networks”. I *Handbook of Natural Computing*, red. G. Rozenberg, s. 335–76. Berlin och Heidelberg: Springer.

Searle, J. 1980. ”Minds, Brains, and Programs”. *Behavioral and Brain Sciences* 3, s. 417–57.

KAJ BØRGE HANSEN

OM THORILD DAHLQUISTS ”BIDRAG TILL TEORIN FÖR DE KLASSLOGISKA ANTINOMIERNA”

Thorild Dahlquist blev fil. lic. 1949. Det gick till på följande sätt. Han hade vid tiden strax före de händelser, som kommer att beskrivas på de följande raderna, bedrivit filosofistudier vid Uppsala universitet i ett decennium. Det fanns allmän enighet bland lärarna vid institutionen om att han var mycket beläst, lärd och kunnig inom ämnesområdet, men trots detta hade han inte tenterat någon kurs. Orsaken till det anges vara att han led av tentamensskräck. Dåvarande professorn i teoretisk filosofi Konrad Marc-Wogau beslöt sig för att ta tag i problemet. Han kontaktade sina professorskolleger Ingemar Hedenius i praktisk filosofi och Torgny Segerstedt i sociologi och föreslog en gemensam aktion. Alla tre professorer kände Thorild väl och visste om både hans psykiska besvär och hans imponerande kunskaper i deras ämnesområden. Marc-Wogau gav utan vidare Thorild 3 betyg (60 poäng/90 högskolepoäng) och dessutom Godkänd på den muntliga tentamen för licentiatexamen, också den omfattande 3 betyg. Segerstedt gav Thorild 1½ betyg (30 poäng/45 högskolepoäng) i sociologi, även det utan examination. Enligt hörsägen tog Hedenius Thorild ut på en promenad i Engelska Parken där de konverserade om olika praktisk-filosofiska frågor varefter Hedenius fattade beslut om att Thorild hade kunskaper som räckte för 2½ betyg (50 poäng/75 högskolepoäng) i praktisk filosofi. Därmed hade Thorild alla kurser som krävdes för en fil. kand. och för en fil. lic. examen i filosofi. Han hade dock inte skrivit någon licentiatavhandling. Däremot hade han en del anteckningar. Thorilds fru Ann-Mari Henschen-Dahlquist, som själv var filosof, åtog sig då att på maskin renskriva och ordna ett urval av anteckningarna som handlade om mängdteoretiska antinomier. Resultatet blev ett manuskript på 20 ganska glest skrivna sidor som ingavs som licentiatavhandling med titeln ”Bidrag till teorin för de klasslogiska antinomierna”. Efter bedömning godkändes det som licentiatavhandling och Thorild blev fil. kand. och fil. lic. samma dag. Enligt uppgift i Stig Kangers doktorsavhandling *Provability in Logic* publicerad 1957 ingavs licentiatmanuskriptet i 1949; men i den version jag har en kopia av finns mot slutet ett kort appendix daterat ”Tisd. den 9 maj 1950”.

Filosofisk tidskrift 2019 nr 1, 43–50

Troligen är appendixet bifogat i efterhand som komplettering, efter att avhandlingen hade lämnats in för bedömning.

Ann-Maris kunskaper i logik begränsade sig till elementär logik. Trots det är manuskriptet klart och välgjort. Det finns några få smärre fel och egendomligheter som inte är meningsstörande. Det innebär att de anteckningar från Thorilds hand som avhandlingen är baserad på har varit väl genomarbetade och utarbetade och att innehållet helt är resultat av Thorilds tankemöda och att Ann-Mari inte har bidragit med resultat eller idéer, fast hon givetvis har gjort ett bra och intelligent jobb med att göra en avhandling av anteckningarna. I många år misstänkte jag att manuskriptet hade bedömts milt och kanske egentligen inte helt motsvarade de då gällande kraven för en licentiatavhandling. Efter att noggrant ha penetrerat avhandlingen har jag ändrat uppfattning på den punkten. En jämförelse med ett par andra logisk-filosofiska licentiatavhandlingar från den tiden, som jag tidigare har sett närmare på, faller enligt min bedömning ut till Thorilds fördel. Hans licentiatarbete är ”a creditable intellectual achievement”.

Thorild publicerade aldrig sin licentiatavhandling. För flera år sedan försökte jag att hitta avhandlingen på Carolina och i Filosofiska institutionens arkiv, men utan resultat. Mot slutet av Thorilds liv, när han var sängliggande dygnet runt, nämnde jag för honom per telefon att jag hade letat efter avhandlingen men inte hittat den. Han sade att han hade ett exemplar någonstans i lägenheten men att han på grund av sitt hälsotillstånd inte var i stånd att leta fram det till mig. När Wlodek Rabinowicz och Svante Nordin efter Thorilds död hade planer om att ge ut ett urval av hans manuskript och brev, gjorde jag en ny sökning efter hans licentiatavhandling. Bland annat skrev jag email till Carolina Rediviva och till Kungliga biblioteket i Stockholm samt kollade om ett exemplar skulle finnas i Filosofiska institutionens arkiv. Jag vände mig även till Uppsala universitets centrala administration om det där skulle finnas ett exemplar arkiverat. Svaren var negativa överallt. När jag för några år sedan, på uppdrag av Ann-Maris syster Märta Henschen och Thorilds brorsdotter Gun Dahlquist, grovsorterade Thorilds och Ann-Maris brev och högar av andra papper, hittade jag licentiatavhandlingen. Innan jag lämnade originalet till Filosofiska institutionens arkiv, gjorde jag två kopior av det, en lämnade jag till Filosofiska institutionen och en behöll jag själv. Originalet och en av kopiorna finns numera i Uppsala universitetsbibliotek Carolina Redivivas Dahlquist-arkiv.

1. SAMMANFATTNING AV AVHANDLINGENS INNEHÅLL

Thorild opererar i det som kallas ”naiv klassteori”, först föreslagen av Cantor och av Frege. Han formulerar två villkor

(1) $x \in x \rightarrow \varphi x$
(2) $\varphi x \rightarrow \exists y\,(y \in x \wedge \varphi y)$

och bevisar ett lemma:

Påstående. Varje egenskap φ som satisfierar (1) och (2) ger upphov till en antinomi.

Därnäst undersöker han en oändlig serie av egenskaper $E_1(x)$, $E_2(x)$, $E_3(x)$ etc. som alla var för sig uttrycker att elementrelationen för x är cirkulär. Enligt E_1 är $x \in x$, enligt E_2 är $x \in y \in x$ för någon klass y, enligt E_3 är $x \in z \in y \in x$ för klasser y och z, och så vidare. Med hjälp av lemmat bevisar Thorild att om vi sätter $\varphi = E_k$ för $k = 1, 2, \ldots$, så ger varje sådan E_k upphov till en antinomi. Resultaten är generaliseringar av Russells paradox. De var kända sedan tidigare, men Thorilds eleganta och enhetliga sätt att bevisa dem är, så vitt jag vet, originellt.

Den nästa antinomin har Thorild däremot upptäckt själv. (Flera år senare visade det sig dock att den hade formulerats och bevisats av Mirimanoff redan 1917; men det kände Thorild inte till i 1949.) Thorild definierar en egenskap B(x), x är bottenlös, för en klass x. Klassen x är *bottenlös*, om det finns en oändlig sekvens av klasser $y_1, y_2, y_3, \ldots$ så att

$$\ldots\, \in y_3 \in y_2 \in y_1 \in x$$

Bottenlöshet är en generellare egenskap än cirkularitet uttryckt av Ek. Således representeras t.ex. $E_2(x)$ av bottenlöshetssekvensen

$$\ldots\, \in y \in x \in y \in x$$

Däremot kan elementrelationen för en klass x vara bottenlös utan att vara cirkulär. Thorild bevisar att $\varphi x = B(x)$ satisfierar villkoren (1) och (2) ovan vilket enligt lemmat implicerar en antinomi, *bottenlöshetsantinomin*. Han utvecklar ett flertal varianter av sina resultat; men bottenlöshetsantinomin är självklart avhandlingens huvudresultat. Thorild påpekar korrekt att ordningen av typerna i Russells typteori blockerar härledningen av cirkularitets- och bottenlöshetsantinomierna. I ett appendix försöker han härleda en bottenlöshetsantinomi i Quines hybrid NF av mängdteori och typteori. Som han själv påpekar är den definie-

rande formel han använder oförenlig med Quines stratifieringsvillkor. Således ger både Russells typteori och Quines system lösningar till de antinomier i naiv klassteori som Thorild har undersökt i sitt arbete.

2. KOMMENTARER

Thorild var i viss mån okunnig om den relevanta litteraturen. Den armenisk-rysk-schweiziske matematikern Mirimanoff var troligen den första som tog möjligheten av en teori för icke-välgrundade mängder (som Thorild kallar bottenlösa klasser) på allvar. (Familjenamnet var ursprungligen det armeniska Mirimanian vilket förryskades till Mirimanoff.) Redan 1917 hade han publicerat en antinomi som väsentligen är identisk med Thorilds bottenlöshetsantinomi. Tydligen kände Thorild inte till Mirimanoffs arbete, när han skrev sin licentiatavhandling. Några få år efter Thorilds avhandling publicerade Shen (1953), ovetande om Mirimanoffs och Dahlquists tidigare bidrag, samma antinomi, varefter kunskapen om antinomin spreds även i engelsktalande länder.

Thorild analyserar antinomin enbart inom ramarna för Russells typteori och Quines stratifieringsteori. Han påpekar korrekt att välordningen av Russells typer och Quines stratifieringsmetod var för sig blockerar härledningen av bottenlöshetsantinomin. Formler som

(1) $x \in x$, $(x \in y \wedge y \in x)$ och $(\ldots \wedge y_3 \in y_2 \wedge y_2 \in y_1 \wedge y_1 \in x)$

är meningslösa i typteorin och kan inte stratifieras i Quines system. Thorild nämner inte alls Zermelos mängdteori Z och dess utvidgningar Zermelo-Fraenkel mängdteori utan urvalsaxiomet ZF eller med samma axiom ZFC. Redan på 1940-talet ansågs ZF och ZFC bland logiker och matematiker att vara överlägsna typteorin och stratifieringsteorin som lösning till mängd- och klassantinomierna. I ZF blockeras härledningen av bottenlöshetsantinomin av två axiom: Axiom of Foundations och Aussonderungsaxiomet. När Mirimanoff härledde sin version av bottenlöshetsantinomin 1917, hade Axiom of Foundations ännu inte formulerats och lagts till Zermelos mängdteori. Dessoaktat implicerar Mirimanoffs bottenlöshetsantinomi, och därför också Thorilds resultat, att Axiom of Foundations inte ensamt räcker för att skydda alla typer av mängdteori mot antinomier; men härledningen av bottenlöshetsantinomierna kan fortfarande blockeras i ZF utan Foundations, nämligen genom närvaran av Aussonderungsaxiomet. I sin härledning av lemmat på sidan I i avhandlingen definierar Thorild mängden

$$K_\varphi = \{x| \neg\varphi x\}$$

Den använder han sedan i alla härledningar. Enligt Axiom of Foundations finns ingen mängd x som satisfierar φx. Därför är $K_\varphi = \{x| \neg\varphi x\}$ identisk med den äkta klassen som innehåller alla mängder så att K_φ inte är en mängd. Existensen av K_φ i mängduniversum är följaktligen inte förenlig med Zermelos Aussonderungsaxiom. Detta är det vedertagna sättet att i Z, ZF och ZFC undvika alla de antinomier som Thorild härleder i sin uppsats, inklusive bottenlöshetsantinomin.

Russells åsikt var att alla logiska paradoxer (som mängd- och klassantinomierna) och semantiska paradoxer (som lögnarparadoxen) har det gemensamt att de bryter mot ”The Vicious Circle Principle”. I en formulering lyder den: ”Whatever involves all of a collection must not be one of the collection.” Russells ”Vicious Circle Principle” utesluter klasser baserade på cirkulära uttryck som

$$x \in x \text{ och } (x \in y \wedge y \in x)$$

Men den uteslutar inte klasser definierade genom bottenlöshetsegenskaper som

$$(\ldots . \wedge y_3 \in y_2 \wedge y_2 \in y_1 \wedge y_1 \in x)$$

där mängderna x, y_1, y_2, ... är parvis olika. Med andra ord har Thorild hittat ett motexempel till Russells ”Vicious Circle Principle” vilket han dock inte påpekar och möjligen inte har observerat själv.

Dessa anmärkningar ändrar inte min uppfattning att Thorilds (åter) upptäckt och utveckling av bottenlöshetsantinomin är en utmärkt intellektuell prestation. Thorilds uppsats godkändes som licentiatavhandling av Konrad Marc-Wogau. Marc-Wogau var inte logiker och hade endast tämligen elementära kunskaper i logik. Jag undrar om han skulle ha godkänt Thorilds uppsats som licentiatavhandling, om han hade känt till Mirimanoffs härledning från 1917.

3. HISTORISKT SAMMANHANG

Jag känner bara till en referens till Thorilds bottenlöshetsantinomi i ett seriöst forskningssammanhang. Stig Kanger skriver i en fotnot på sidan 26 i sin doktorsavhandling *Provability in Logic*:

> See SHEN 1953. The paradox of the class of all grounded classes was stated by Mr. Thorild Dahlquist in a paper on some new set-theoretical antinomies, which was submitted in 1949 in partial fulfillment of the requirements for the fil. lic. degree at the University of Uppsala.

Den effekt som Thorilds bidrag fick på senare forskning blev därför begränsad. En orsak är givetvis att Thorild aldrig publicerade sin licentiatavhandling.

En forskningslinje där Thorilds härledning av bottenlöshetsantinomin har en plats är i utvecklingen av mängd- och klassantinomier. I den linjen återfinns Cantors, Burali-Fortis, Russells och Zermelos paradoxer samt Mirimanoffs, Dahlquists och Shens versioner av bottenlöshetsantinomin. En annan linje där hans arbete har en plats är i strävandena att härleda de logiska och semantiska paradoxerna utan självreferens och cirkularitet och i stället ersätta dessa med bottenlöshetsegenskaper. Därmed fås motexempel till Russells ”vicious circle principle”. För de logiska paradoxerna (mängd- och klassantinomierna) har detta gjorts av Mirimanoff (1917), Dahlquist (1949) och Shen (1953). För de semantiska paradoxerna (lögnarparadoxen) har det gjorts av Yablo (1993).

En axiomatiserad version av teorin för icke-välgrundade mängder publicerades av Peter Aczel (1988). I den är Axiom of Foundations negerad. Inspirationen fann Aczel bl.a. i axiomatiseringen av ZF av Zermelo med tillägg av von Neumann, Skolem och Fraenkel och i Mirimanoffs arbeten om icke-välgrundade mängder, inklusive dennes tidiga version av bottenlöshetsantinomin. Givetvis har Aczel inte känt till Thorilds arbete. (Däremot ser det ut som Aczel vid utarbetningen av boken från 1988 kände till Kangers användning av icke-välgrundade mängder i *Provability of Logic*.)

4. TVÅ ANEKDOTER

Denna anekdot är självupplevd. Vid ett samtal på tumanhand med Thorild någon gång på 1980-talet framförde jag min åsikt att Zermelo-Fraenkels mängdteori är överlägsen Russells typteori som lösning till mängdparadoxerna och som grundvalar för matematiken. Thorild svarade då i en lätt irriterad ton: ”Ja, men fördelen med typteorin är att den har intressanta filosofiska tillämpningar” – uppenbarligen med undermeningen att det har ZF inte.

Kommentar: (1) Thorild har fel i denna synpunkt. Både den enkla typteorin och den förgrenade typteorin har modeller i ZFC. Härav följer att

allt som kan göras i typteorierna kan göras i ZFC, oavsett om det är matematiskt eller filosofiskt. (2) En annan konklusion jag drar är att Thorild, möjligen av den anledning han själv anger, aldrig fann det värt mödan att sätta sig ordentligt in i ZF. Detta märks i licentiatavhandlingen och det märktes i hans kollokvier helt in på 1990-talet. (3) En tredje reflektion är att Thorild betraktade Russell som en av sina tre stora idoler inom den analytiska filosofin vars storhet inte fick ifrågasättas. (De andra två var Carnap och Quine.)

Den andra anekdoten har berättats av Krister Segerberg. När Shen publicerade sin version av bottenlöshetsantinomin i 1953, informerades Thorild om detta vilket uppenbarligen tyngde honom. Den vidare utvecklingen har Thorild själv berättat i ett brev till Bo Petersson, Linköping, daterad 4 september 2007: ”[M]ånga år senare fick jag genom Quines bok *Set Theory and its Logic* reda på att mitt lilla fynd hade föregripits av [Mirimanoff] i en fransk matematisk tidskrift [så tidigt som 1917].” När Thorild berättade om denna logikhistoriska upptäckt för Krister Segerberg, avslutade han med att utbrista: ”Vilken lättnad!” – en reaktion Krister fann lätt komisk.

Kommentar: Krister Segerberg har, tror jag, missförstått Thorild. Krister har korrekt tolkat Thorilds reaktion på Shens publicering som uttryck för att det var negativt för honom själv (Thorild) att en annan hade kommit före med publiceringen. Krister har sedan dragit slutsatsen att då kan det inte vara bättre för Thorild att en tredje person (Mirimanoff) har publicerat samma upptäckt ännu tidigare. Den korrekta tolkningen av att Thorild blev ledsen, när han hörde om Shens publicering, är givetvis att han förebrådde sig själv för att ha förlorat möjligheten att göra anspråk på prioritet till bottenlöshetsantinomin genom att ha underlåtit att publicera den i tid. Den lättnad han kände, när han informerades om Mirimanoffs ännu tidigare publicering, kommer sig av att eftersom antinomin publicerades 1917, tre år innan Thorild själv föddes, hade han hursomhelst inte haft möjlighet att genom rättidig publicering göra anspråk på att själv vara först med bottenlöshetsantinomin.

LITTERATUR

Aczel, P. 1988. *Non-Well-Founded Sets*. Stanford: Center for the Study of Language and Information.

Carlson, E. och R. Sliwinski, utg. 2001. *Omnium-gatherum: Philosophical Essays Dedicated to Jan Österberg on the Occasion of his Sixtieth Birthday.* Uppsala Philosophical Studies 50. Uppsala: Uppsala universitet.

Dahlquist, T. 1949–50. ”Bidrag till teorin för de klasslogiska antinomierna”. Opublicerad licentiatavhandling. Filosofiska institutionen, Uppsala universitet.

Hansen, K. B. 2001a. ”Kanger’s Ideas on Non-Well-Founded Sets: Some Remarks”. I Holmström-Hintikka m.fl. 2001b, s. 69–86.

Hansen, K. B. 2001b. “Two Paradoxes Revisited”. I Carlson och Sliwinski 2001.

Holmström-Hintikka, G., S. Lindström och R. Sliwinski, utg. 2001a. *Collected Papers of Stig Kanger with Essays on his Life and Work*, vol. 1. Dordrecht: Kluwer.

——— 2001b. *Collected Papers of Stig Kanger with Essays on his Life and Work*, vol. 2. Dordrecht: Kluwer.

Kanger, S. 1957. *Provability in Logic*. Acta Universitatis Stockholmiensis, 1. Stockholm: Almqvist & Wiksell. Omtryckt i Holmström-Hintikka m.fl. 2001a.

Mirimanoff, D. 1917a. ”Les antinomies de Russell et de Burali-Forti et le problème fondamental de la théorie des ensembles”. *L’enseignement mathématique* 19, s. 37–52.

———. 1917b. ”Remarques sur la théorie des ensembles”. *L’enseignement mathématique* 19, s. 209–17.

Shen, Y. 1953. ”Paradox of the Class of all Grounded Classes”. *Journal of Symbolic Logic* 18, s. 114.

Yablo, S. 1993. ”Paradox Without Self-Reference”. *Analysis* 53, s. 251–52.

RECENSIONER

Fysik
Aristoteles. Översättning Charlotta Weigelt. Thales 2017. 375 s.
ISBN 978-91-7235-102-8

Ibland händer det att jag får i uppdrag att recensera böcker om fysik som översatts till svenska. Det är ofta en trivsam syssla där det ges möjlighet att sväva ut lite grand om fysik i största allmänhet. När det gäller det föreliggande verket är det par saker som skiljer ut sig. Dels handlar det om en översättning från grekiska vilket väl får sägas höra till ovanligheterna. Engelska är ju det som annars dominerar. Men det verkligt anmärkningsvärda är att vi har fått vänta i mer än två tusen år på att ta del av boken på svenska. Det handlar om inget mindre än Aristoteles *Fysik*.

Hur det kunnat dröja så länge innan det äntligen blev av är lite underligt. När Uppsala universitet slog upp portarna för den förste fysikstudenten år 1486 gissar jag att Aristoteles nog fanns högt upp på listan över kurslitteratur. Men på samma sätt som många kursböcker idag med fördel läses på engelska, och man inte behöver göra sig besvär med att översätta till svenska, fanns det väl inte något skäl heller när det gäller Aristoteles. Kunde man inte grekiska gick det bra med latin. Med tiden avklingade intresset för hans bok och ingen verkar ha kommit sig för att översätta.

Vad är det då för en bok? När man skall popularisera fysiken och sätta den i ett historiskt sammanhang är det inte ovanligt att Aristoteles får klä skott och utgöra exempel på hur primitiv den tidiga vetenskapen var. Hur kunde man tro att jorden låg i världsalltets mitt? Hur kunde man ha en så felaktig bild av vad rörelse är? Jag har själv farit ut på detta sätt i olika sammanhang, men med tiden lärt mig behärskning. Och efter att nu haft i uppdrag att skriva en uppriktig recension inser jag skamset hur orättvis jag varit i min förhastade dom.

Det tar sin början redan med titeln som är lika koncis som genialisk: ”Fysik”, kort och gott. Aristoteles tar ett djärvt grepp på hela naturvetenskapen – och delvis mer än så – men kallar sitt verk ändå rätt och slätt fysik trots att den handlar om så mycket mer än det som vi i våra snäv-

synta dagar betecknar som fysik. Om man i stället för att betrakta detta förhållande i något slags historiskt ljus – och skyller på att ord förändrar sin mening – tar det på fullaste allvar får man inblick i en världsbild som fortfarande har en del att ge. Det handlar om en övergripande världsbeskrivning som skall omfatta allt och vad kan passa bättre som titel än just ”Fysik”. För det är väl just fysik som allt i grund och botten handlar om? Frågan är bara vad fysik egentligen är.

Aristoteles låter djuren, växterna och alla kroppars yttersta beståndsdelar höra till den naturliga världen. Det enda som inte hör dit är det som är framställt med hjälp av mänsklig konstfärdighet. Som exempel nämner Aristoteles föremål som en säng eller en mantel. De substanser som de består av hör till naturen men inte föremålen i sig. Om någon skulle gräva ner en säng av trä i jorden, växer det inte upp en ny säng – resonerar Aristoteles – men någon gång kanske ett träd. Aristoteles vill förstå sig på dessa materiens inneboende drivkrafter men tvekar heller inte inför att klassificera de orsaker som ligger bakom mänsklig aktivitet.

Denna grandiosa ambition kräver mer än vad dagens fattiga fysikaliska världsbild mäktar med. Aristoteles laborerar med fyra olika orsaker, där den fjärde för oss känns alldeles främmande. Aristoteles menar att det finns saker som är och händelser som sker ”för någots skull”. För en modern fysiker saknar begrepp som avsikt och mening helt funktion. Ju längre från den mänskliga världen de fenomen man beskriver befinner sig, desto lättare är det att förbli renlärig i detta avseende. Men när levande varelser, och inte minst människor, skall omfattas av beskrivningen blir det lite svårare. När det gäller huruvuda det mänskliga medvetandet verkligen kan inrymmas i en sådan beskrivning går uppfattningarna isär, och det är nog få som konsekvent undviker att i ord tillskriva sig själva och andra människor syften och avsikter. När Aristoteles förklarar existensen av en staty genom att hänvisa inte bara till de materiella grunderna, utan också till att skulptören hade ett syfte med att skapa den, är vi med på noterna. Visst spelar den finala orsaken roll, men inte har väl detta med fysik att göra? Här är det Aristoteles som driver hem poängen. Aristoteles ser världen som en helhet där den mänskliga sfären ingår tillsammans med den naturliga och att samma spelregler i grunden måste gälla. Allt är fysik.

Dessutom är vi alla hycklare, åtminstone i det sätt som vi väljer att i ord beskriva världen. Hur många gånger har jag inte använt ett målande bildspråk där jag talat om att kroppar vill falla hit eller dit för att

få lyssnaren att förstå. Det är i grunden omöjligt att avhålla sig av rent pedagogiska skäl. Detta är också något att ha i åtanke när vi bedömer Aristoteles system. Vad vi i all vetenskap försöker göra är att avbilda, eller översätta, det vi förnimmer av världen till något vi kan hantera. Bilderna kan aldrig bli perfekta men de kan vara olika användbara. Hänvisningar till finalitet blir på så sätt ursäktliga.

Men menar han verkligen att naturen verkar utifrån en avsikt? Innebär inte teleologi i praktiken teologi? En noggrannare läsning av Aristoteles skingrar misstankarna. Inget kunde vara honom mer främmande. Han är en nykter naturvetare, kanske den allra förste, och han är tydlig med att det fortfarande är ett slags fysiklagar som det handlar om. Det är bara det att det inte nödvändigtvis alltid är händelser i det förflutna som förklarar det som sker i en framtid, utan riktningen kan också vara den motsatta. Det vill säga, ett framtida förhållande kan påverka vad som sker idag. Man kan väl se detta som ett uppluckrande av den tidspil som vi annars tar för given.

Det är också begripligt att författaren väljer att utvidga orsaksbegreppet på detta sätt. Biologin skall ju enligt hans synsätt betraktas som en gren av fysiken och allt det som biologin tampas med måste fysiken alltså kunna förklara. Aristoteles refererar till greken Empedokles som i efterhand väl kan ses som ett slags antikens Charles Darwin. Empedokles föreställde sig hur levande varelser komponerats på ett slumpmässigt sätt och hur endast de bäst skickade överlevde. Aristoteles är inte säker på att detta räcker som förklaring och ser ett behov av orsaker som handlar om att något är för någots skull. Och faktum är att det nog är svårt också för en modern evolutionsbiolog att avhålla sig från sådana formuleringar när man skall förklara varför ett djur besitter en särskild färdighet.

Men det är om Aristoteles syn på rörelse som det skojats mest. Och den är väl med våra dagars sätt att se inte helt lyckad, även om den definitivt är ganska intutitiv. Den tar ju sitt avstamp i vad som måste vara en direkt erfarenhet grundad i egna observationer och experiment. En rörelse upphör helt enkelt då det inte längre finns något som kan skjuta på. Nog känns det som en rimligt slutsats? När man slutar trampa, och det inte längre lutar nerför, så stannar cykeln. Det finns några enstaka undantag, men det var först genom Galilei som tanken att det kan vara precis tvärtom slog rot. Den naturliga rörelsen är helt enkelt den evigt pågående och oförändrade så länge inga krafter verkar – i enlighet med det Newton så småningom skulle kalla sin första lag. Detta ideala till-

stånd uppnås dock endast i det absoluta tomrum som Aristoteles av flera skäl såg som omöjligt. Det är intressant att se hur han brottas med begreppet. När han försöker leda i bevis att tomheten är en motsägelsefull omöjlighet hamnar han ibland hisnande nära den korrekta slutsatsen, men backar alltid i sista sekunden. Han inser att en massa som rör sig i ett medium har en fart som beror av dess massa och tätheten hos mediet – han utför till och med en liten enkel beräkning för att illustrera förhållandet. Men om det är alldeles tomt? Vad händer då? Kommer alla kroppar att röra sig oändligt fort, eller kanske lika fort? Han är något på spåren men tar inte det avgörande steget. Resonemangen knyter till och med an till en slags primitiv relativitetsteori. Rörelse är antingen påtvingad eller naturlig, men om kroppen inte har någon påtvingad rörelse, och befinner sig i ett tomrum, kan det heller inte finnas någon naturlig rörelse. Det finns ju inga skillnader mellan här och där, så hur vet kroppen vart den skall bege sig?

Aristoteles sätt att argumentera känns ibland märkligt moderna. Det finns en resonerande röst som eftertänksamt ställer upp tänkta situationer som sedan analyseras. Frågan som ställs är ”vad skulle hända om?” Det handlar om tankeexperiment av ett slag som får en modern teoretisk fysiker känna sig hemma och man påminns om hur själve Einsten frågade sig ”hur skulle det vara att rida på en ljusstråle?” Följeslagaren är inte auktoritär och ställer sig själv i bakgrunden till förmån för det intellektuella argumentets kraft.

När någon frågar mig vad tid är svarar jag alltid ”tid är det man mäter med en klocka”. Jag säger det inte på skämt utan på fullaste allvar och gör på så sätt sällskap med Aristoteles. Han gör en jämförelse mellan tid och rum. Tiden blir till något som kan diskuteras på samma sätt som rummet. De har liknande egenskaper och i detta för en fysiker så självklara påstående hittar man en föraning om vad som skulle komma tusentals år senare. Liksom en kontinuerlig linje inte kan reduceras till en räcka diskreta punkter, kan inte tiden vara bara en samling av ”nu”. Ett ”nu” blir i stället en slags utkikspunkt men i detta nu finns ingen rörelse. Tiden är något mer. ”För detta är tiden: ett tal för rörelse med ett före och ett efter.” Tid är i grunden rörelse, en rörelse som kan bokföras med hjälp av matematiska tal. Ett före och ett efter och man kopplar an till Newtons upptäckt av differentialkalkylen där rörelsen genom ett finurligt gränsvärde kunde sägas existera i ögonblicket vid en viss tidpunkt. Inte genom att i egentlig mening finnas i nuet – detta är som Aristoteles insåg omöjligt – utan genom en jämförelse. Genom

att jämföra positionen vid två tidpunkter och låta dessa komma varandra allt närmare, kan en hastighet beräknas som kvoten mellan en allt mindre sträcka och ett allt mindre tidsintervall. I gränsvärdet kryper de allt närmare nuet utan att egentligen komma dit. Ett före och ett efter som ger ett tal för rörelse.

När jag berättar om universums storhet och vår obetydliga plats får jag ofta frågor om oändligheten. Den fråga man brottas med handlar inte om någon krånglig process i det tidiga universum eller egenskaper hos den mörka energin, utan rätt och slätt om begreppet oändlighet. Man förutsätter att en förståelse av vad oändligheten egentligen är för något – och något sällsamt märkligt måste det förstås vara – är det som är nyckeln till att begripa universum. Här kommer Aristoteles till räddning med sitt nyktra förhållningssätt. Med utgångspunkt inte i en översinnlig matematisk värld utan i ett konkret och praktiskt förhållningssätt till en fysiskt existerande verklighet drar han slutsatsen att oändligheten i någon absolut mening inte existerar annat än som en möjlighet. En möjlighet som aldrig fullt ut kan förverkligas. När skolbarnen tävlar om att säga det högsta talet och någon av dem prövar ”ett mer än du någonsin kan säga” är de samma insikt på spåren. Oändligheten uppnås stegvis: 1, 2, 3, 4.... så långt du bara orkar. Och sedan ett tal till, och ännu ett till. Oändligheten är ingen sak att lägga i en låda, det är en process. Och just så används begreppet i fysiken. Hur stort universum egentligen är vet ingen. För eller senare, om vi blickar tillräckligt långt bort, eller räknar oss fram, kommer vi nog till ett slut, eller åtminstone till regioner som ser helt annorlunda ut. Men så länge denna eventuella gräns inte påverkar vad vi gör här och nu kan vi bortse från dess existens och i våra beräkningar använda oss av den potentiella oändligheten. En slags approximation till något som är alldeles gräsligt stort och minst ett mer än vad vi än så länge har anledning att diskutera. Någon platonsk idévärld där den aktualiserade oändligheten verkligen existerar behövs inte.

Aristoteles har något att tillföra också i vår tid. Han tar ställning för den empiriska verkligheten. Naturen är på riktigt och han är själv en del av den. Världen är inte bara matematik som andra mer världsfrånvända föregångare som Pythagoras kanske skulle ha hävdat, eller mindre verklig än den platonska idévärlden. Aristoteles är empiristen som tar världen på fullaste allvar och vill förstå den så långt det bara är möjligt.

Men mot slutet av *Fysik* tappar Aristoteles lite av sin skärpa och spårar ur. Han konstaterar att rörelsen är evig – det kan inte finnas någon

början, och naturen evig och beständig. Gott så. Men sedan får han för sig att hans världsbygge inte riktigt kan stå för sig själv utan kräver en upprätthållande kraft som utgör början av alla orsakskedjor: den förste röraren. Det handlar inte om något slags skapare som vid ett givet tillfälle får fart på allt – naturen är ju evig. Aristoteles har lite svårt att riktigt veta var denna entitet, som tydligen har någon form av substans, skall hålla till. Det får bli någonstans i utkanten av himlasfären. Det går inte ihop och det hade nog varit bättre om Aristoteles behållit just dessa funderingar för sig själv.

Aristoteles *Fysik* är en förvånansvärt lustfylld läsning. Jag får väl erkänna att vissa partier kan vara lite tungrodda men överallt hittar man skarpsinniga iaktagelser att grunna vidare på. Det är dessutom mycket underhållande att försöka passa in det han skriver i en modern världsbild. En bidragande orsak är det lättflytande språket, som vi nog inte bara har författaren själv att tacka utan också översättaren Charlotta Weigelt. Hon har dessutom tillfogat boken en innehållsrik introduktion på 50 sidor.

Det var sannerligen på tiden att Aristoteles *Fysik* blev översatt till svenska. Mycket av det som nu är allra hetast inom den fundamentala vetenskapen kommer nog snart att vara bortglömt, men jag är övertygad om att Aristoteles fortfarande kommer att läsas med behållning när det gått ytterligare två tusen år. Och vem vet, kanske det då har blivit dags för en nyöversättning.

ULF DANIELSSON

Superintelligens: Vägar, faror, strategier
Nick Bostrom. Översättning Jim Jakobsson. Fri Tanke 2017, 516 s.
ISBN 978-91-87513-08-4

Nick Bostrom är en framgångsrik svenskfödd filosof vid universitetet i Oxford, där han leder en avdelning som heter ”The Future of Humanity Institute”. En annan internationellt framgångsrik svensk, fysikern Max Tegmark vid MIT, anser enligt förlagsreklamen för denna bok att Bostrom är ”en av världens vassaste tänkare”.

Bostrom hävdar här att superintelligenta maskiner kan komma att överglänsa – och i värsta fall utrota – biologiska människor. ”Maskiner har ett antal grundläggande fördelar som kommer att göra dem enormt överlägsna. Biologiska människor, även om de är förstärkta, kommer att bli utklassade” (s. 88).

Denna profetia låter väldigt definitiv, och rätt skrämmande, men i förordet säger Bostrom också att boken ”sannolikt är allvarligt felaktig och vilseledande” (s. 11). Så man vet inte riktigt vad man ska tro. I alla händelser menar han att en maskinell superintelligens ”skulle själv kunna vara en extremt mäktig agent, som med framgång kunde hävda sig mot det projekt som skapat det likaväl som mot resten av världen” (s. 152).

Utvecklingen av maskinintelligens kan dessutom bli explosionsartad. Denna idé om en ”intelligensexplosion” har framförts av många andra, tidigast kanske 1965 av matematikern I. J. Good, som också citeras av Bostrom. När – och om – man uppnår superintelligens kommer den vidare utvecklingen att gå mycket fort, ty då ”äger utveckling och forskning rum i de tidsskalor som kännetecknar maskinell superintelligens – kanske tusentals eller miljontals gånger snabbare än den forskning som bedrivs i en biologisk mänsklig tidsskala” (s. 276).

Enligt Bostrom kan en superintelligens innebära stora fördelar för oss människor. ”Risker som härrör från naturen – exempelvis asteroidnedslag, supervulkaner och naturliga pandemier – skulle närmast elimineras, eftersom superintelligensen kunde vidta motåtgärder mot de flesta sådana risker eller åtminstone flytta ner dem till den icke-existentiella kategorin (till exempel genom kolonisering av rymden)”. Den skulle också ”minska risken för oavsiktlig förstörelse, inklusive risken för olyckor som är relaterade till ny teknologi” (s. 356).

Inte nog med det. ”Om revolutionen i maskinintelligens avlöper väl, skulle den superintelligens som blir resultatet nästan säkert kunna utveckla metoder för att efter behag förlänga livet för de då ännu levande människorna, och inte bara hålla dem vid liv utan återge dem hälsa och ungdomlig vitalitet och öka deras förmåga långt bortom vad vi idag ser som det mänskliga spektrumet” (s. 379).

KONTROLLPROBLEMET

Men även om en superintelligens kan vara oss till stor hjälp – om nu Bostrom har rätt – så kan den som sagt också bli livsfarlig. Även om den håller sig till de ”slutmål” vi har bestämt för den, så kan den komma på helt oförutsedda medel att uppnå dem och dessa medel kan drabba oss. Till exempel så att superintelligensen röjer oss ur vägen för att vi hindrar dess verksamhet – t.ex. genom att hota att ”dra ur kontakten” – eller använder oss som råvara för att fabricera något den anser sig behöva (t.ex. s. 195).

Därmed ställs vi inför det som Bostrom kallar ”kontrollproblemet” (s. 202f). Hur ska vi hindra maskinintelligensen att skada oss? Isaac Asimovs tre lagar för robotar räcker inte som skydd (s. 220).

Men superintelligensen har knappast någon fri vilja, även om Bostroms formuleringar ibland kan tyda på motsatsen (som när han säger att den kan vara ”en extremt mäktig agent”). Såvitt jag förstår gör den bara det den är programmerad att göra (så länge det inte uppstår något mekaniskt fel). Den kan förstås vara programmerad att lära sig saker, så att den till sist vet och kan mycket mer än programmeraren. Och den kan förstås tänka mycket snabbare. Men den gör ändå bara det programmeraren direkt eller indirekt har beordrat den att göra. Kontrollproblemet löser man genom att se till att superintelligensens program förhindrar att den gör något dumt. Gör den något dumt, så är det alltså programmerarens fel. Ytterst är det inte superintelligensen som är livsfarlig, utan programmeraren. Alltså är det egentligen *programmeraren* – eller snarare alla AI-forskare – som måste kontrolleras! Vilket nog är omöjligt. Så även om man skulle hitta ett program som hindrar en superintelligens att åstadkomma skada, så är kontrollproblemet därmed inte löst.

Man kunde kanske tänka sig att överlåta kontrollproblemet till den väldigt smarta superintelligensen. Men Bostrom skulle antagligen invända att superintelligensen då kan lösa problemet genom att utplåna mänskligheten. Om vi inte längre finns löper vi ju ingen risk att skadas av superintelligensen!

GENERELL ARTIFICIELL INTELLIGENS

För att en artificiell superintelligens ska uppstå krävs två saker. För det första att man skapar en generell artificiell intelligens (AI) på minst mänsklig nivå. För det andra att en sådan artificiell intelligens sätter igång en vidare utveckling till allt högre grader av intelligens.

Kan dessa två betingelser uppfyllas?

Det är fortfarande en öppen fråga om man kan skapa generell intelligens på hög mänsklig nivå. Bostrom anser att det är möjligt och att det finns olika vägar till superintelligens, men han säger: ”Sann superintelligens (i kontrast till marginella ökningar i nuvarande intelligensnivåer) kan troligen uppnås först på AI-vägen” (s. 86).

Vi bör alltså fokusera på *artificiell* intelligens. (Observera att en maskin som lyckas imitera en människa i ett Turingtest kan vara lika ointelligent som en ointelligent människa.) Men det är oklart vad som me-

nas med ”mänsklig nivå”. AI-experter talar om ”human-level machine intelligence, HLMI” (s. 39), men man kan ju, som Bostrom själv senare påpekar, skilja mellan t.ex. snabbhet och kvalitet (s. 89–94). Digital intelligens är ju oerhört mycket snabbare än biologisk (s. 99). Ska den ligga på ”mänsklig nivå” måste den alltså vara kvalitativt betydligt sämre, för att kompensera för överlägsenheten i snabbhet (jfr s. 118). Kort sagt: snabb men dum.

Det framgår inte särskilt tydligt i Bostroms bok vad som avses med ”intelligens” (och ordet finns inte heller med i det omfattande registret). På ett ställe definieras det som ”något i stil med förmåga till förutsägelse, planering och resonemang om mål och medel generellt” (s. 171).

Kanske kan man säga att intelligens består i förmåga att *lösa problem*. Människor är intelligentare än maskar och skalbaggar, vilket kunde betyda att vi kan lösa en väldig massa problem som de varken kan eller vill lösa. Och en superintelligens kan förhålla sig till oss som vi till maskar och skalbaggar (s. 149). Bostrom talar dunkelt om ”möjliga men oförverkligade kognitiva talanger” (s. 96), som gör det möjligt att lösa problem som ”inte kan lösas bit för bit och som kanske kräver kvalitativt nya typer av förståelse eller nya representationella ramverk som är för djupa eller för komplicerade för att dödliga varelser av nuvarande snitt ska kunna upptäcka eller använda dem på ett effektivt sätt” (s. 98).

Det är klart att vi kan skapa maskiner som löser vissa problem oerhört mycket snabbare än vi. Redan en miniräknare gör det, för att inte tala om en avancerad schackdator. Men de har inte *generell* intelligens. Det skulle kräva att de kan lösa ungefär samma problem som vi på samma områden som vi.

Men kan de också, på samma sätt som vi, vara *nyfikna*? Att de *kan* lösa problem är en sak, men *vill* de lösa problem? Och i så fall *vilka* problem?

En miniräknare kan lösa aritmetiska problem, men den är inte nyfiken. Den frågar sig ingenting. Den levererar bara lösningar på problem som vi har ställt upp. Detsamma gäller schackprogram. Gäller det även en *generell* intelligens?

Miniräknare och schackprogram kräver (utöver energitillförsel) en viss *input* – eller mer specifikt: en problemformulering eller en fråga – för att leverera en *output,* ett svar. Detsamma gäller såvitt jag förstår en artificiell generell intelligens. Liksom de program som kan klara ett Turingtest. (Det gäller nog också människor, även om många av oss tror att det är ”vi själva” som så att säga ”inifrån” producerar input till den output vi sedan uppvisar i vårt handlande.)

Bostrom menar att en generell intelligens, eller åtminstone en superintelligens, har ett *slutmål*. Som exempel nämner han bl.a. sådant som att ”producera så många gem som möjligt” (s. 195) eller att ”att göra projektets finansiär nöjd” (s. 189) eller ”att göra oss lyckliga” (s. 192). Det verkar väldigt konstigt att en generell superintelligens skulle ha ett sådant slutmål. Såvitt jag förstår kan ”slutmålet” knappast vara något annat än att lösa de problem som matas in som input i systemet. Miniräknare och schackprogram har heller inga slutmål utöver att lösa de problem som de har programmerats att lösa.

Bostrom tänker sig för övrigt också att superintelligensen är en Bayesiansk agent (s. 196), med en viss nyttofunktion och en apriori sannolikhetsfördelning över alla möjliga världar (s. 346). Kanske menar han att detta (eller enbart nyttofunktionen) är dess ”slutmål”? Och att detsamma gäller varje generell intelligens?

Men det gäller i alla fall inte människor. Även om människor skulle ha en nyttofunktion och en sannolikhetsfördelning vid varje given tidpunkt – vilket verkligen kan betvivlas – så skulle de i alla fall inte vara konstanta över tid. Vår hjärna förändras hela tiden och vi utsätts för olika stimuli (input) vid olika tillfällen.

För övrigt tror jag att om en artificiell generell intelligens har en nyttofunktion och en sannolikhetsfördelning över alla möjliga världar, så försvinner distinktionen mellan slutmål och instrumentellalla mål (dvs. medel), som Bostrom lägger så stor vikt vid och som i sin tur ska motivera intelligensexplosion och kontrollproblem. Ty *all* ”motivation” specificeras då av dessa funktioner och allt är lika mycket ett ”slutmål”.

INTELLIGENSEXPLOSION

Vad är det som kan sätta igång en digital intelligensexplosion? Enligt Bostrom är det att en AI får förmåga att förbättra sig själv, speciellt att den får en ”områdesspecifik talang för kodning och AI-forskning”. Den kan då komma in i en process av ”rekursiv självförbättring” (s. 54). Denna process kan bli väldigt snabb och leda fram till en ”stark superintelligens”, dvs. ”en intelligensnivå långt över den samtida mänsklighetens samlade intellektuella resurser” (s. 105).

Kort sagt, superintelligensen kan programmera mycket bättre än människor och den kan därför lösa många svåra problem som ligger långt utanför mänsklig räckvidd. Inte bara enskilda geniala människors räckvidd, utan hela mänsklighetens räckvidd.

Lösningarna på de problem som mänskligheten just nu har kunnat lösa finns, kan man kanske säga, på nätet. Bostrom säger också: ”Googles sökmotor skulle kunna beskrivas som det största AI-system som hittills konstruerats” (s. 35). Men är Googles sökmotor över huvud taget ”intelligent”? Den kan inte lösa problem som människor inte kan lösa. Den kan inte heller programmera. Den är kanske i någon mening en ”superintelligens”, men den kan knappast ”tänka själv” och den kan inte råka in i en process av ”rekursiv självförbättring” och den vill nog inte utrota oss. Bostrom själv menar att Googles sökmotor ”ligger långt under den mänskliga baslinjen för varje rimligt mått på generell intellektuell förmåga” (s. 105).

Så *varför* börjar en superintelligens på ”den mänskliga baslinjen” förbättra sig själv? Bostrom har antagligen ett svar på detta, i kapitel 4, ”En intelligensexplosions kinetik” (s. 104–27), men jag har tyvärr inte lyckats förstå vad detta svar går ut på. Program skapade av människor kan väl förväntas bli bättre och bättre, som hittills, men varför skulle ett program som ligger på den mänskliga baslinjen, om något sådant skulle bli möjligt, fortsätta *på egen hand* att skapa nya program? Det kan lära sig mer och mer, men det är inte detsamma som att skapa nya program, dvs. program som löser nya problem.

VÄRDEFRÅGOR

Bostroms bok innehåller massor av detaljer och spekulationer. För den specialintresserade finns det mycket att hämta. Det finns information om allt möjligt som har samband med människor, hjärnforskning, maskiner, databehandling och den allmänna teknologiska utvecklingen sedan antiken. Noter och litteraturförteckning omfattar ungefär hundra tätskrivna sidor.

Men boken tangerar också rena värdefrågor. En stark superintelligens borde kanske även ha värderingar. Men vilka? ”Det är för närvarande inte känt hur mänskliga värden kan överföras till en digital dator, även givet maskinintelligens på mänsklig nivå” (s. 320). Men dessutom är det kanske inte just *våra* värden som borde överföras. Bostrom anser tydligen att en superintelligens kunde prestera en mycket *bättre* värdelära än vad vi primitiva varelser hittills har lyckats åstadkomma (jfr s. 323f.). Han tycks vara moralisk realist! Han anser att ”vi kan ha fel om moralen” och att filosofernas oenighet dessutom visar att ”de flesta filosofer måste ha fel” (s. 324). Superintelligensen är mer tillförlitlig.

Men i så fall borde vi kanske akta oss väldigt noga för att försöka överföra våra ”mänskliga värden” till en maktfullkomlig digital dator. Å andra sidan kan man undra om vi verkligen bör låta superintelligensen sköta värderingarna. Den kommer kanske fram till att livet är meningslöst och att både den själv och det som eventuellt återstår av mänskligheten därför bör utrotas för gott. Kanske inser den att mänskligheten på det hela taget är av ondo och skadlig för sig själv och för andra levande varelser. Den värderingen har nog inte Bostrom själv, men han anser ju att han kan ha fel och att han bör lita på superintelligensen. Kanske är det vår undermåliga intelligens som gör att så många av oss finner livet uthärdligt och ibland till och med trevligt. Och att vi har fått för oss att det är värdefullt att mänskligheten överlever inom överskådlig framtid. I så fall kan vi strunta i kontrollproblemet.

LARS BERGSTRÖM

NOTISER

NÅGRA KÄNDA FILOSOFER SOM DÖTT UNDER 2018

Adolf Grünbaum, Gene Sharp, Baruch Brody, Stanley Cavell, Albrecht Wellmer, Sylvain Bromberger, Isaac Levi och Mary Midgley (som blev 99 år och därmed överlevde sina berömda jämnåriga kamrater: Iris Murdoch som dog 1999, Elisabeth Anscombe som dog 2001 och Philippa Foot som dog 2010).

ETT MORALISKT LIV

”Att vara moralisk är något annat och mer än att bara anpassa sitt liv efter normer som alltid skulle finnas till hands och lösa problem, som om det främst gällde att rätta sig efter en kommunikationstabell. Att vara moralisk är en form av liv. […] Att vara moralisk – det är att ibland kunna vara – ja, ofta – ansvarslös och bekymmerslös även vad gäller ens eget moraliska liv.” Så skriver Åke Löfgren i en uppsats som nyligen har publicerats i tidskriften *Divan* och där han går igenom en lång rad olika egenskaper som han anser kännetecknа ett moraliskt liv.

Åke Löfgren, som dog 2017, var lektor i Karlstad, men studerade innan dess praktisk filosofi i Stockholm under 1950- och 1960-talen. Han var en omtyckt lärare och han startade 1965 gruppen Unga filosofer, som kom att spela stor roll inom den tidens vänsterdebatt. En av hans närmaste medarbetare var Roger Fjellström, förste redaktör för *Häften för Kritiska Studier*, som nu har redigerat och presenterat Löfgrens uppsats.

Medarbetare i detta nummer av *Filosofisk tidskrift*: *Lars Bergström* är professor i praktisk filosofi, *Ulf Danielsson* är professor i teoretisk fysik i Uppsala, *Kaj Børge Hansen* är docent i teoretisk filosofi i Uppsala, *Anders Hansson* är fil. mag. i praktisk filosofi, *Erik Rezazadeh* studerar matematisk statistik vid Göteborgs universitet och har tidigare studerat praktisk filosofi i Stockholm och Göteborg, *Paul Conrad Samuelsson* studerar filosofi i Stockholm och *Marco Tiozzo* är gymnasielärare och doktorand i praktisk filosofi i Göteborg.

INSTRUKTIONER TILL SKRIBENTER

Filosofisk tidskrift har som syfte att bidra till en allsidig och fruktbar diskussion av filosofiska problem, samt att på ett lättfattligt sätt informera om aktuell filosofisk forskning. Den vänder sig inte enbart till fackfilosofer, utan vill framför allt nå en bredare läsekrets av filosofiskt intresserade personer.

Tidskriften står öppen för olika filosofiska inriktningar, men den vill undvika bidrag som man inte kan tillgodogöra sig utan speciella förkunskaper eller tekniska färdigheter. Utöver längre artiklar på omkring 10–15 sidor, tar tidskriften gärna emot även kortare bidrag och inlägg av notiskaraktär.

MANUSKRIPT TILL FILOSOFISK TIDSKRIFT:

- skickas med e-post till lars.bergstrom@philosophy.su.se
- skall utformas i överensstämmelse med den typografi som är normal i tidskriften, utan onödig formatering
- skall vara försedda med namn och adress
- skall som regel vara skrivna på svenska, och citat från andra språk bör översättas till svenska (ej nödvändigt i fotnoter)
- särskild litteraturförteckning upprättas vid behov i alfabetisk ordning och placeras sist i manus
- för icke beställt material ansvaras ej
- korrektur läses i regel endast av redaktören
- införda bidrag honoreras inte
- i stället för särtryck erhåller författaren, gratis, 10 exemplar – för recensioner 5 exemplar – av det nummer av tidskriften i vilket bidraget varit infört