Filosofisk tidskrift

Årgång 40 Nr 4 November 2019

3 Människans mörka materia:
Om implicit bias och moraliskt ansvar
KATHARINA BERNDT RASMUSSEN OCH ÅSA BURMAN

27 Giltiga slutledningar
LARS BERGSTRÖM

33 Har vi *verkligen* en fri vilja?
RUBEN WIWE

38 Evolutionär kärlek
BERTIL KRISTERSON

RECENSIONER

42 Nicolas Olsson Yaouzis om *Moraliska provokationer* av Marcus Agnafors

47 Lars Bergström om *Liv 3.0* av Max Tegmark

53 William Bülow om *Historisk rättvisa: Gottgörelse i en postkolonial tid* av Göran Collste

57 Lars Bergström om *Förnuft & fördom: Varför vi tänker som vi gör* av Hugo Mercier och Dan Sperber

61 NOTISER

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, c/o Filosofiska institutionen,
Stockholms universitet, 106 91 Stockholm
info@bokforlagetthales.se www.bokforlagetthales.se

Copyright © Respektive författare 2019

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

KATHARINA BERNDT RASMUSSEN OCH ÅSA BURMAN

MÄNNISKANS MÖRKA MATERIA: OM IMPLICIT BIAS OCH MORALISKT ANSVAR

1. INLEDNING

Denna artikel är skriven av två övertygade feminister och antirasister. Vi anser till exempel att kvinnor i allmänhet är lika lämpade att utbilda sig inom, och jobba med, naturvetenskapliga ämnen som män i allmänhet. Likaså är det självklart för oss att våra medmänniskors hudfärg inte spelar någon roll för hur vi bemöter dem.

Ett sådant tillkännagivande kan verka lite onödigt och överdrivet präktigt. Men våra övertygelser kommer i ett lite annat ljus när vi genomför ett par webbaserade så kallade implicita associationstester (IAT).[1] Testresultaten avslöjar att vi båda – var för sig – har en ”[m] edelstark association mellan Manligt och Naturvetenskap samt mellan Kvinnligt och Humanistiska ämnen”, och likaså har vi en ”[m] edelstark automatisk preferens för Ljushyade människor framför Mörkhyade”. Med andra ord: vi verkar härbärgera sexistiska associationsmönster avseende naturvetenskap och humaniora, och rasistiska tendenser i vår syn på mörkhyade respektive ljushyade människor. När det gäller kategorierna kön och ras eller etnicitet uppvisar vi vad som har kommit att kallas för *implicit bias*.

Det tar emot att skriva detta. Diskrepansen mellan IAT-resultaten och de övertygelser och preferenser som vi vill, och tror oss, stå för gör ont. Men framför allt väcks en oro i oss: tänk om dessa mönster och tendenser även avspeglas i vårt vardagliga och professionella handlande. Tänk om vi diskriminerar kvinnor och rasifierade. Tänk om vi bidrar till de sociala orättvisorna som drabbar dessa grupper, utan att ens lägga märke till det. Skälet till att fenomenet implicit bias väcker vårt intresse är alltså slutligen normativt. Det är fel att diskriminera kvinnor och rasifierade. Vi bör avskaffa de sociala orättvisorna. För att ta reda på huruvida implicit bias motverkar dessa normativa mål, och hur vi i så fall kan förändra vår bias, behövs en bättre förståelse för fenomenet och dess mekanismer.

1. För en närmare beskrivning av IAT, se avsnitt två nedan.

Filosofisk tidskrift 2019 nr 4, 3–26

Den här artikeln är ett försök att förstå den empiriska forskningen kring implicit bias, sätta dess resultat i ett filosofiskt ramverk, och skissa dess implikationer för moraliskt ansvar. I nästa avsnitt kommer vi att närmare beskriva hur ett IAT går till och vad det egentligen är som mäts. Avsnitt tre skissar några medvetande- och handlingsfilosofiska teorier om hur det som antas ligga bakom mätresultaten bäst kan konceptualiseras. De efterföljande två avsnitten söker svar på frågan: Vilken praktisk relevans har allt detta? I avsnitt fyra gör vi en övergripande kartläggning av den empiriska forskningen kring hur implicit bias påverkar individers varseblivning, omdömen och handlingar. I avsnitt fem fokuserar vi i stället på vilken relevans implicita fördomar har för samhällsnivån och för kollektivt moraliskt ansvar. Utifrån dessa slutsatser diskuterar vi, i avsnitt sex, vilka idéer om moraliskt ansvar som blir relevanta för dem som bär på, och påverkas av, implicit bias. Avsnitt sju summerar artikelns viktigaste slutsatser.

2. IMPLICIT BIAS

Vad är då ett IAT, och vad mäter det? Enkelt uttryckt bedömer testet styrkan på försökspersonens fördomar. Exempelvis mäter IAT för Kön-Naturvetenskap personens associationer mellan begreppen i distinktionsparen ”man/kvinna” och ”naturvetenskap/humaniora”. Själva uppgiften är att sortera enskilda ord som dyker upp mitt på datorskärmen i rätt kategori i det övre högra eller vänstra hörnet på skärmen. Orden mitt på skärmen som ska sorteras är dels könskodade ord (t.ex. tjej, moder, fästman eller herre), och dels utbildningsämnen (t.ex. litteraturvetenskap, historia, matematik eller astronomi). Kategorierna i de övre hörnen på skärmen är dels ”humaniora” respektive ”naturvetenskap” och dels ”man” respektive ”kvinna”. Tricket är att dessa kategorier sätts ihop som disjunktioner. Ordet mitt på skärmen ska då klassas som exempelvis ”kvinna *eller* humaniora” eller ”man *eller* naturvetenskap”. Sorteringen sker med enkla tangenttryckningar och ska gå så fort och felfritt som möjligt. Det visar sig att de allra flesta försökspersoner är snabbare och gör färre misstag, när de ihopsatta kategorierna följer ovannämnda mönster än när de avser de ombytta disjunktionerna ”man *eller* humaniora” eller ”kvinna *eller* naturvetenskap”. Det här tas som ett tecken på att dessa försökspersoner associerar kvinnor med humaniora och män med naturvetenskap. Hela 72 procent av försökspersonerna uppvisade en sådan – svag, medelstark

eller stark – association.[2] Vi som skrivit denna artikel är alltså i gott (?) sällskap.

På ett liknande sätt fungerar IAT för Hudfärg. Mitt på skärmen dyker det upp dels bilder på ansikten med mörk eller ljus hy, och dels positivt eller negativt laddade ord (t.ex. skratt, fred, smärta eller misslyckande). Dessa ska sorteras in i ihopsatta kategorier som ”mörkhyad *eller* bra” eller ”ljushyad *eller* dålig”. De flesta försökspersoner visar sig vara långsammare eller gör fler misstag, när de ihopsatta kategorierna följer detta mönster än när de avser ”ljushyad *eller* bra” eller ”mörkhyad *eller* dålig”. Det tas som ett tecken på att de föredrar ljushyade människor framför mörkhyade. Sammanlagt 65 procent av försökspersonerna uppvisade en sådan – svag, medelstark eller stark – preferens.[3] Återigen är vi inte ensamma om våra resultat.

IAT mäter och sammanväger alltså reaktionstider och felfrekvenser vid sortering av dessa ord eller bilder. Resultatet blir ett mönster som typiskt sett inte kan förklaras utifrån de övertygelser och värderingar som försökspersonerna ger uttryck för – precis som i vårt fall. Vad ger då upphov till mätresultaten? Hur kan de förklaras?

Ett sätt att tolka resultaten är att de flesta av oss, bakom en tunn civilisationens fernissa, *är* sexister och rasister. Detta får vi dock sällan syn på, eftersom normer kring vad som är socialt önskvärt påverkar vilka attityder vi uttrycker öppet. Empiriska studier visar att många tvekar att avvika från den etablerade normen i politiskt eller moraliskt känsliga frågor, även när deras anonymitet garanteras.[4] Får vi explicita frågor om våra fördomar och tid att formulera oss, så skyler vi över dem, exempelvis med präktiga tillkännagivanden om att vara övertygade feminister och antirasister. Även inför oss själva vill vi inte kännas vid våra fördomar. Men tas möjligheten att reflektera och anpassa oss bort så uppenbaras vad vi egentligen anser, ”innerst inne” – åtminstone enligt denna tolkning.

Ett annat sätt att tolka resultaten – som inte står i direkt konflikt med vad vi genuint tror om oss själva – är att anta att det finns inre mentala processer som står i någon sorts motsättning till våra medvetna

2. Siffran bygger på 299 298 IAT-värden för ”IAT för Kön-Naturvetenskap” som gjordes mellan juli 2000 och maj 2006 (https://implicit.harvard.edu/implicit/sweden/). Det bör noteras att detta IAT räknar ämnet filosofi, som brukligt, som ett humanistiskt ämne. Filosofi skiljer sig dock från andra humanistiska ämnen genom att snarare associeras till det manliga könet, se Saul 2013b och Di Bella m.fl. 2016.

3. Siffran bygger på 122 988 IAT-värden för ”IAT för Hudfärg” som gjordes mellan juli 2000 och maj 2006 (https://implicit.harvard.edu/implicit/sweden/).

4. Se t.ex. Greenwald m.fl. 1998; Krumpal 2013.

och uttalade övertygelser och preferenser, och som (ibland) påverkar vårt beteende – bland annat det beteende som mäts av IAT. En vanlig, funktionell definition av *implicit bias* fångar denna tanke: implicit bias är (mestadels) omedvetna processer som påverkar ens varseblivning, omdöme och handlande gentemot medlemmar i vissa sociala grupper.[5]

En belysande jämförelse kan göras med astrofysikens mörka materia. På 1970-talet upptäckte Vera Rubin och hennes kollegor att de yttre stjärnorna i ett antal avlägsna spiralgalaxer roterade med samma hastighet som stjärnorna närmare galaxernas centrum. Den observationen stred mot den Newtonska gravitationsteorin, givet vad man visste om galaxerna i övrigt: antalet stjärnor, deras massa etc. Utifrån dessa faktorer borde de yttre stjärnorna rotera mycket långsammare runt centrum. En möjlig förklaring av diskrepansen kunde vara att det var fel på (forskarnas användning av) gravitationsteorin. Den rimligaste förklaringen ansågs dock vara att dessa galaxers massa är mycket större och annorlunda fördelad än vad de observerbara delarna ger oss skäl att tro. Slutsatsen blev att det måste finnas mer materia – icke-observerbar sådan – vars distribution och sammantagna massa får ekvationerna att gå ihop enligt gravitationsteorin. Allt vi hittills vet om denna *fysikens mörka materia* är hur den, genom gravitation, påverkar sin omgivning.[6]

På ett liknande sätt strider de IAT-resultat som vi, artikelns författare, fick mot det vi tror om oss själva. Vi är säkra på att vi inte kopplar män till naturvetenskap och kvinnor till humanistiska ämnen. Vi är säkra på att vi inte föredrar ljushyade framför mörkhyade människor. De systematiska skillnader i vårt beteende som mäts av testen kan därför inte förklaras med hänvisning till våra explicita – introspektionstillgängliga och artikulerbara – övertygelser och preferenser.

Enligt implicit bias-forskningen är den rimligaste förklaringen inte heller ovannämnda hypotes om att vi inte svarar ärligt på undersökningar. Snarare antas att det finns andra mentala tillstånd som påverkar vårt observerbara beteende, det vill säga associationernas uppmätta hastighet och felfrekvens. Dessa implicita mentala processer eller tillstånd kännetecknas av att vara otillgängliga för introspektion och självrapportering, men även snabbaktiverade och bortom vår direkta kontroll – och potentiellt oförenliga med våra explicita mentala tillstånd. Det är denna *människans mörka materia* som har kommit att kallas för implicit bias. Vi kommer i det följande använda uttrycken ”implicit

5. Jfr Holroyd och Sweetman 2016, s. 81, Saul 2013b, s. 40.
6. Se Rubin 1983.

bias", "implicita fördomar", "implicita skevheter", och "implicit social kognition"[7] som synonymer.

Men hur ska vi närmare förstå detta fenomen, det vill säga det som ger upphov till de här observerbara utfallen? Inom medvetande- och handlingsfilosofin har det gjorts ett flertal olika försök att integrera fenomenet med befintliga teorier om medvetande och motivation.

3. FILOSOFISKA TOLKNINGAR

En central skiljelinje i den nutida analytisk-filosofiska debatten om implicita fördomars natur är om vi ska förstå dessa som övertygelser (eller övertygelse-lika tillstånd) eller inte. Tamar Gendler förnekar att implicita fördomar är just övertygelser och introducerar begreppet "alief" i sitt banbrytande arbete med att konceptualisera implicita fördomar.[8] En alief är ett mentalt tillstånd med intentionalt innehåll av representativa, affektiva och beteendeaktiverings-komponenter (*R–A–B*), som aktiveras automatiskt. En aliefs intentionala innehåll kan vara oförenligt med det intentionala innehållet hos simultant hållna explicita övertygelser och värderingar. Till exempel kan en person skrika (B-komponenten), utifrån en känslan av rädsla (A-komponenten), när hen tittar på filmen *Psycho* (R-komponenten), trots att personen är helt medveten om att hen inte är i någon omedelbar fara. På samma sätt kan en person, i mötet med någon av mörkare hudfärg, ha följande alief: *Mörkhyad person!–Läskigt!–Undvik!* Detta trots att personen har icke-rasistiska explicita övertygelser och attityder.[9] En viktig poäng med Gendlers syn på implicita fördomar är att de kan figurera som en kausal mekanism för att förklara diskrepansen mellan en persons explicita övertygelser och hens beteende.

Denna syn på implicita fördomar har dock kritiserats av exempelvis Eric Mandelbaum som hävdar att om implicita fördomar vore en rent associativ mekanism, så skulle det inte vara möjligt att minska deras förekomst genom rationell argumentation. Empiriska studier indikerar dock att detta är möjligt.[10] Mandelbaum kritiserar också begreppet "alief" för att vara en icke-enhetlig "sammelsurium"-kategori.[11] Tim Bayne och Anandi Hattiangadi hävdar att begreppet "alief" står inför ett

7. Jfr Greenwald och Banaji 1995.
8. Gendler 2008a, Gendler 2008b.
9. Jfr Brownstein 2017, paragraf 2.1.
10. Se t.ex. Briñol m.fl. 2009.
11. Mandelbaum 2016.

kausal-explanatoriskt dilemma. Som en enhetlig, icke-”sammelsurium”-kategori kan den inte ta hänsyn till heterogeniteten av handlingar eller beteenden hos individer med till exempel samma perceptuella erfarenhet (R-komponenten). Men för att kausal-förklara en sådan existerande heterogenitet så måste den upphöra att vara en enhetlig kategori, och således införs en kostsam irregularitiet i förklaringen.[12] Även Kathrin Glüer och Åsa Wikforss har argumenterat emot Gendlers syn på implicit bias som ”alief” genom ett så kallat ”alief-dilemma”. Om alief förstås distinkt från en irrationell övertygelse så kan begreppet inte längre användas i förklaringar av avsiktliga handlingar (i motsats till förklaringar om rent beteende). Men när alief används i avsiktliga handlingsförklaringar, så går det inte längre att skilja en alief från en (irrationell) övertygelse.[13]

Som ett alternativ till Gendlers syn har andra teoretiker föreslagit att utvidga den Humeanska ”övertygelse-önsknings”-modellen för att kunna förstå och integrera implicit bias. Mandelbaum argumenterar för en ”strukturerad övertygelse”-hypotes: implicita fördomar uppstår ur omedvetna, propositionellt strukturerade mentala representationer; och vi människor står i ett slags övertygelse-förhållande till dessa mentala representationer.[14] Andra relaterar implicita fördomar till mer kända begrepp, som generiska övertygelser: Sally Haslanger föreslår att de är ett slags kognitiva scheman, som består av ”kluster av kulturellt gemensamma begrepp, övertygelser och andra attityder som möjliggör att vi tolkar och organiserar information och koordinerar handling, tanke och affekt”.[15] Sarah-Jane Leslie menar i stället att det handlar om ”slående egenskaps-generaliseringar”, en primitiv form av generalisering av typen ”den vita hajen attackerar surfare”, eller (mera relevant i detta sammanhang) ”killar är bra på matte”.[16]

Dessa modeller måste dock kompletteras med en förklaringsmodell för hur en diskrepans kan uppstå mellan dessa övertygelse-liknande tillstånd och våra explicita övertygelser. En generell sådan förklaringsansats är idén att medvetandet är fragmenterat, exempelvis i medvetna och omedvetna övertygelser som aktiveras i olika kontexter. En annan idé är Eric Schwitzgebels gradualistiska dispositionalism, enligt vilken predikatet ”tro att *p*” är vagt. Vi kan därmed ”mitt-emellan”-tro att kvin-

12. Bayne och Hattiangadi 2013.
13. Glüer och Wikforss 2013.
14. Mandelbaum 2016.
15. Haslanger 2013, s. 11, vår övers.
16. Leslie 2017.

nor och män är lika lämpade för att utbilda sig inom och jobba med naturvetenskap.[17]

Det saknas med andra ord inte försök att förklara den mörka materian; däremot saknas enighet kring förklaringsmodellernas även mest basala struktur. Men gör det något, egentligen? Vi får dra oss till minnes varför vi är intresserade av att förstå implicit bias från början. Vi skrev inledningsvis att vi oroar oss för att våra IAT-resultat härrör från sexistiska och rasistiska associationsmönster, som kanske även påverkar vårt vardagliga och professionella handlande. Grundidén bakom denna oro är att implicit bias *orsakar* diskriminering. För att illustrera: låt oss anta att en av artikelförfattarna, som är övertygad om att det inte finns skillnader i intelligens och flit mellan studenter med svenskklingande namn och studenter med icke-svenskklingande namn, ändå ger lägre poäng till den senare gruppen under tentamensrättningen. Idén är att hade hon inte haft sina (genom IAT:et påvisbara) implicita fördomar, så hade det inte funnits någon poängskillnad mellan grupperna.

Denna form av poängdiskriminering är såklart oacceptabel, både moraliskt och utifrån meritokratiska principer. För att kunna förstå och motarbeta sådan här, och annan, diskriminering behöver vi veta mer om mekanismerna bakom, vilket förutsätter att vi har någon plausibel modell av den förmodat bakomliggande mörka materian. Vår bedömning är dock att vi hittills inte har tillräckligt starka skäl att acceptera någon modell framför dess alternativ. Vi måste då suspendera vårt omdöme.

En sak vet vi dock: testresultaten visar på systematiska skillnader i vårt beteende, det vill säga våra tangenttryckningar i en laboratoriesituation. Men kan vi knyta dessa resultat till diskriminering i det verkliga livet? Idén att IAT-resultat korrelerar med diskriminerande beteende, får visst stöd av den empiriska forskningen.[18] Men även här är bilden komplicerad.

4. IMPLICIT BIAS OCH DISKRIMINERING PÅ INDIVIDNIVÅ

Ett flertal studier pekar på att IAT-resultat reliabelt förutsäger olika typer av, speciellt icke-verbala, diskriminerande beteenden: exempelvis huruvida man röstar på den svarta kandidaten,[19] kallar någon med ara-

17. Schwitzgebel 2013.

18. För en genomgripande och koncis sammanfattning av den vetenskapshistoriska bakgrunden till dagens implicit bias-forskning, se Jost m.fl. 2009. Artikeln listar också tio arbetslivsrelevanta studier ”som ingen manager borde ignorera” (ibid. s. 47).

19. Greenwald m.fl. 2009. Deras resultat motsägs dock i viss mån av Ditonto m.fl. 2013.

bisktklingande namn till anställningsintervju,[20] uppför sig hövligt mot kvinnor respektive mörkhyade på arbetsplatsen,[21] etc. En stor metastudie från 2009 visar att just i socialt känsliga lägen, och specifikt i kontakter mellan människor med olika hudfärg, så är IAT-resultaten en signifikant bättre prediktor än självrapporterade (explicita) övertygelser.[22]

Men det finns också studier som motsäger att IAT-resultat är särskilt användbara för att förutsäga beteende, det vill säga studier som underminerar de påståenden om korrelation och prediktion som de förstnämnda studierna gör. En nyare metastudie från 2013 nyanserar och motsäger emellertid i viss mån den tidigares resultat.[23] Andra studier indikerar att IAT-resultaten speglar kulturella stereotyper och ojämlikhet som föreligger i det omgivande samhället.[24] Det innebär att även om vissa korrelationer mellan IAT-resultat och diskriminering kan påvisas, så kan orsakssambandet i själva verket gå från den senare till den förra (snarare än att båda dessa orsakas på något sätt av den okända mörka materian).

Gendler ifrågasätter relevansen i invändningen att testresultaten bör tolkas som avspeglande av social kunskap, snarare än som en reflektion av något inre tillstånd:

> Utifrån alief-perspektivet spelar det ingen roll huruvida IAT mäter ens grad av tillgång till information som man omfamnar (som hävdas av ”evaluativ bias”-läsningen) eller ens grad av tillgång till information som man avvisar (som hävdas av ”social kunskap”-läsningen). Det som IAT otvivelaktigt avslöjar – som benämningen antyder – är implicita associationer. […] den sociala kunskapen i sig involverar implicita associationer mellan vissa ras-/etniska kategorier och ett starkt affektivt innehåll.[25]

Ett sådant gestalt-skifte i konceptualiseringen av implicit bias gör därmed inte fenomenet i sig mindre intressant.

Sammanfattningsvis är dock problemet att det är oklart dels vad IAT egentligen mäter, och dels hur starka korrelationerna egentligen är mel-

20. Rooth 2010.
21. Cortina m.fl. 2013, Dovidio m.fl. 2002.
22. Greenwald m.fl. 2009.
23. Oswald m.fl. 2013. Se även Forscher m.fl. 2018, som visar att implicit bias kan påverkas av en rad interventioner, men att de uppmätta förändringarna i implicit bias inte överförs till uppmätta förändringar i beteenden. Detta kastar tvivel på idén om ett direkt orsakssamband mellan implicit bias och diskriminerande beteende. Den senare studien har uppmärksammats stort i olika medier, men det bör tilläggas att den ännu inte har publicerats i någon fackgranskad tidskrift.
24. Se t.ex. Karpinski och Hilton 2001, Arkes och Tetlock 2004.
25. Gendler 2008b, s. 577, vår övers.

lan IAT-resultaten och diskriminerande beteenden. I brist på trovärdiga förklaringsmodeller av de mentala tillstånden saknas dessutom plausibla mekanismer för att sammanlänka det som IAT mäter och diskriminering. Mycket av kritiken och oklarheterna kretsar alltså kring just själva mätverktyget: IAT.

Det bör påpekas att det finns andra tester som inte drabbas av samma invändningar som IAT. Dessa tester mäter andra (misstänkta) effekter av den mörka materian än hastigheten och felfrekvensen i abstrakta sorteringsövningar som vi känner igen från IAT. I detta sammanhang är det särskilt intressant med de tester där det som mäts i hög grad liknar det som utgör reellt diskriminerande beteende.

Ett sådant test kallas för ”shooter bias”-testet, och i det får försökspersonen spela ett slags dataspel där hen möter en rad motspelare som håller i olika svårtydda föremål. Föremålet kan vara ett vapen, en flaska eller en mobiltelefon, och spelaren behöver under tidspress fatta ett aktivt beslut om att antingen inte skjuta eller skjuta på motspelaren. En metastudie från 2015 visar att försökspersonerna – både poliser och civilpersoner – är snabbare med att skjuta beväpnade svarta motspelare och långsammare att inte skjuta obeväpnade svarta motspelare, jämfört med respektive kategori av vita motspelare.[26] Liknande resultat har visats för shooter bias-test med ”icke-muslimska” och ”muslimska” motspelare (där de senare kännetecknas av att bära en turban eller hijab).[27] Dessa testresultat framstår som högrelevanta för att belysa reell diskriminering (såsom polisskjutningar och övervåld) – oavsett hur motivationsmekanismerna bakom utfallen ter sig.

En liknande typ av test, där uppgiften är att *uppfatta* farligheten hos svårtolkade föremål under så kallad priming, kan även anses belysa en del av de mentala processerna bakom dessa handlingar (att skjuta respektive inte skjuta). Testerna består av att ett moment av priming, det vill säga en snabb exponering med en bild på ett mörk- eller ljushyat ansikte, åtföljt av ett moment där en bild på föremålet presenteras och ska bedömas som föreställande ett vapen eller ej. Testdeltagarna identifierade vapen snabbare och misstog oftare icke-vapen för vapen, när de prime:ats med mörkhyade ansikten än när de prime:ats med ljushyade ansikten.[28] Detta indikerar att det föreligger en skevhet redan i vår *varseblivning* – oavsett vilken typ av implicit skevhet som antas ligga

26. Mekawi och Bresin 2015.
27. Unkelbach m.fl. 2008.
28. Payne 2001, jfr Saul 2013a.

till grund för den. Jennifer Saul drar slutsatsen att dessa experiment ”visar att implicit bias får fatt i oss redan innan vi hinner reflektera kring världen – det påverkar vår varseblivning av denna värld [...] på ett oroväckande sätt”.[29]

Ett annat exempel på verkliga omständigheter där människors omdömen blir direkt relevanta för utfallet är anställningsprocessen. I så kallade ”CV-studier” skickas ett antal identiska CV:n till arbetsgivare, med den enda skillnaden att namnet byts ut (t.ex. från ett förmodat svenskt namn till ett arabiskklingande namn eller från ett typiskt mansnamn till ett kvinnonamn). Den första kategorin visar sig ha betydligt större sannolikhet att få komma till en jobbintervju, trots att innehållet i CV:n är exakt detsamma.[30]

Ytterligare ett exempel utgörs av studier som utvärderar bedömningar av kvinnliga respektive manliga sökanden till chefspositioner. Kvinnor visades där stå inför en så kallad dubbelbindnings-situation: när de presenterades (genom en filminspelning) som självsäkra framstod de som mindre socialt kompetenta än män som presenterades på exakt samma självsäkra sätt, och förlorade därmed i meritvärde. När de däremot presenterades som mer socialt kompetenta (och därmed mindre självsäkra) ansågs de inte kvalificerade för chefspositionen.[31]

De här testerna och studierna har alltså gemensamt att det de mäter i labbet liknar det som är själva problemet i verkligheten: diskriminering. Exempelvis diskriminering i form av en större risk (att bli skjuten trots att man är oskyldig), eller en sämre chans (att få komma på anställningsintervju och därmed få jobbet), endast på grund av hudfärg, etnicitet eller kön. Idén är alltså att det som mäts inte är någon (potentiellt bristfällig) proxy för något dunkelt fenomen som i sin tur måste kopplas till verklig diskriminering, utan studierna simulerar diskriminering direkt i labbet.[32]

Även om den idén håller, kan det invändas att det inte är säkert att det som shooter bias-tester och CV-studier mäter härrör från samma

29. Saul 2013a, s. 245, vår övers. Saul menar att den nya kunskapen om implicit bias ger upphov till ett särskilt slags tvivel, som hon kallar bias-relaterat tvivel (”bias-related doubt”).

30. Se t.ex. Rooth 2010. Jfr även Arai m.fl. 2016, som visar att när information om den sökandes etnicitet kombineras med information om dennes kön, så vänds diskrimineringsmönstret för de med arabiskklingande namn (så att männen kallas mer sällan på intervju än kvinnorna), medan för de med svenskklingande namn följer mönstret vanlig könsdiskriminering (där män kallas oftare på intervju än kvinnor).

31. Rudman och Glick 1999.

32. Jfr Holroyd 2012, s. 276.

inre mörka materia som det som IAT mäter. Testerna kanske spårar helt skilda mentala fenomen?[33] Här kan vi dock svara att forskare inte heller vet vad IAT egentligen mäter. Det vi har antagit är att IAT-resultaten *inte* kan förklaras med hänvisning till explicita övertygelser och preferenser. *Om* vi kan anta samma sak även för de andra testerna, så faller även deras (dolda) orsaker under samma, ovannämnda, funktionella definition: implicit bias är (mestadels) omedvetna processer som påverkar ens varseblivning, omdöme och handlande gentemot medlemmar i vissa sociala grupper. Det här antagandet verkar hålla för shooter biastester,[34] och även för testerna där självsäkra kvinnors lämplighet för chefspositioner skulle utvärderas.[35] Det är svårare att underbygga det för CV-studier, då dessa typiskt utgörs av fältstudier där CV:n sänds ut, utan möjlighet till uppföljning för att kunna utesluta *explicita* fördomar. Men även här finns vissa belägg för att den lägre svarsfrekvensen för exempelvis muslimska arbetssökande inte prediceras av arbetsgivarens explicita, men väl hens implicita fördomar.[36]

En mer besvärande invändning är att det endast finns ett fåtal tester som direkt simulerar själva grundproblemet: reell diskriminering. Testerna pekar på potentiella problem med våra sociala (polisiära eller meritokratiska) praktiker. Men dessa problem framstår som mycket mer begränsade än den omfattning av diskriminering som vår allmänna kunskap om implicit bias låter oss misstänka. I ljuset av detta glapp vill vi föreslå en alternativ strategi som sammanlänkar resultaten av IAT och liknande tester med storskalig diskriminering och ojämlikhet på samhällsnivå. Denna strategi innebär att vi skiftar fokus från individnivå till samhällsnivå.

5. EN SAMHÄLLELIG FÖRKLARINGSMODELL

Vår strategi utgår ifrån de sociala ojämlikheter på samhällsnivå som exempelvis svenska myndigheter sedan länge mäter och följer. Statistiska data visar att trots att kvinnor i högre utsträckning än män har en högre utbildning, så tjänar de bara runt 80 procent av männens sammanlagda inkomst; de tar ut tre utav fyra föräldraledighetsdagar; de är systematiskt underrepresenterade i sociala och ekonomiska maktpositioner;

33. Jfr Holroyd och Sweetman 2016.
34. Se t.ex. Correll m.fl. 2002, Banks m.fl. 2006.
35. Rudman och Glick 2001.
36. Rooth 2010.

dessutom jobbar kvinnor och män inom en till stora delar könssegregerad arbetsmarknad, och så vidare.[37] När det kommer till ras och etnicitet så vet vi att invandrare, jämfört med människor födda i Sverige, har statistiskt sett lägre sannolikhet att avsluta sin grundskoleutbildning; att arbetslöshet och relativ fattigdom drabbar ett oproportionerligt större antal utomeuropeiska invandrare (speciellt från Mellanöstern och Nordafrika); att människor med utomeuropeiska namn har lägre chans att få komma på anställningsintervju eller att få lägenhetskontrakt än människor med traditionellt svenska namn; att människor bor i tydligt etniskt segregerade områden, etc.[38]

Hur kan denna uppmätta ojämlikhet bäst förklaras? De förklaringsförsök som hänvisar till skevheter i de formella lagarna och reglerna eller till explicita fördomar hos svenskar i allmänhet, verkar inte fungera fullt ut. Sverige är en liberal demokrati med lagar som inte i sig tar hänsyn till kön, ras eller etnicitet. Det finns därutöver en stark antidiskrimineringslagstiftning och en antidiskrimineringsmyndighet (DO). Dessutom visar undersökningar att personer med sexistiska och rasistiska attityder är i minoritetsställning i Sverige: en tydlig majoritet rapporterar positiva attityder till diversitet,[39] respektive anser att jämställdhet är ett politiskt viktigt mål.[40] Med andra ord har vi även här en diskrepans mellan explicita regler och värderingar, och de beteenden eller utfall vi kan observera – denna gång på en kollektiv eller samhällelig nivå.

En möjlig förklaring är att svenskar i allmänhet, men även beslutsfattare och ”grindvakter” såsom rekryterare, har en statistisk tendens att nedvärdera exempelvis mörkhyade personers kompetens och tillförlitlighet, respektive att koppla ihop kvinnor med vissa yrken och färdigheter mer än andra. Denna förklaring hänvisar med andra ord till de kunskaper som vi har kring implicit bias, genom studier på individnivå. När vi lyfter upp dessa kunskaper till en samhällelig nivå kan även de väldigt svaga uppmätta korrelationerna aggregeras till märkbara skillnader i utfall. Många små skevheter, eller blott statistiska tendenser till sådana skevheter, blir tillsammans till ett tydligt skevt mönster.

För att ta bostadssegregationen som exempel: Thomas Schellings dy-

37. SCB 2016.

38. Socialstyrelsen 2010. För undersökningar kring etnisk diskriminering på den svenska arbets- resp. bostadsmarknaden, se Bursell 2014 resp. Ahmed och Hammarstedt 2008.

39. Ahmadi m.fl. 2016. Procentandelen har dock sjunkit från 74 till 64 procent mellan 2014 och 2016. Jfr Mella m.fl. 2014. Jfr även World Values Survey Association 2014.

40. SIFO 2014. Jfr även World Values Survey Association 2014.

namiska modeller undersöker de systemiska effekterna av exempelvis en samling individers mycket svaga aversion mot att bo i ”blandade” bostadsområden – eller till och med enbart mot att själv tillhöra en minoritet inom sitt grannskap. Utfallet blir mycket segregerade bostadsområden.[41] Betänk ytterligare ett exempel: Richard Martell och kollegor studerar de kumulativa effekterna av köns-bias i en företagsintern befordringsprocess. Även en mycket svag bias för manliga anställda (uttryckt till exempel som en systematisk 1- eller 5-procentig övervärdering av männens prestationer, jämfört med kvinnornas) leder till att de högre nivåerna inom organisationen efter en period kommer att starkt domineras av män, medan de ”kvarlämnade” kvinnorna dominerar de lägre nivåerna – även om utgångsläget är helt jämställt.[42]

Visserligen kan det vara så att dessa modellers ingångsvärden – aversionerna mot ”blandade” områden, respektive bias för manliga anställda – skulle kunna ligga på en explicit snarare än på en implicit nivå. Men modellerna visar att *om* implicit bias finns och korrelerar med människors beteende, så kan den slå igenom på en kollektiv nivå och leda till tydliga ojämlikheter, *även om* korrelationen är mycket svag på individnivå.

Utöver de slutsatser som kan dras av sådana teoretiska modeller har det nu också påbörjats ett nytt sätt att empiriskt testa korrelationen mellan testresultat och diskriminering på kollektiv nivå. Genom att exempelvis använda sig av individuella IAT-resultat, aggregerade på nationsnivå för 34 olika länder, kan forskare påvisa en tydlig korrelation mellan Köns-IAT-resultat och uppmätta könsskillnader i åttondeklassares prestationer i NO och matte. Korrelationen mellan dessa könsskillnader och aggregerade explicita (självrapporterade) sexistiska attityder var jämförelsevis betydligt svagare.[43] Det har också visats att i regioner i USA med högre genomsnittliga associationsnivåer mellan svarta människor och vapen (enligt ovan diskuterade shooter bias- eller priming-tester) så löper också obeväpnade svarta en högre risk att bli skjutna av polisen. Även här är korrelationen starkare mellan svartas skjutningsrisker och genomsnittliga implicita mått, än mellan de förstnämnda och mått på explicita attityder – eller mått på regionala skillnader i exempelvis bostadssegregation eller brottsligt beteende.[44]

41. Schelling 1971. För en enkel webbaserad simulering av Schellings modeller, se www.ncase.me/polygons/.

42. Martell m.fl. 1996. Jfr även Martell m.fl. 2012. För en enkel webbaserad simulering av deras modell, se www.doesgenderbiasmatter.com.

43. Nosek m.fl. 2009.

44. Hehman m.fl. 2018.

Ett problem med en sammanlänkning av sådana implicita mått och strukturell diskriminering är att det moraliska ansvaret – klandervärdheten – för det sistnämnda fördelas bland *för många* individer med, var för sig, *för otydlig* delaktighet. Det verkar göra det svårt att på ett korrekt sätt *tillskriva* ansvar – tillskriva klandervärdhet – och att rättmätigt *utkräva* ansvar – klandra eller straffa. Men just möjligheten att kunna utkräva ansvar, för att kunna motverka diskriminering, var ju själva grunden till varför vi började undersöka fenomenet implicit bias i denna artikel. Har vi hamnat i en återvändsgränd?

I nästa avsnitt visar vi på några implikationer för moraliskt ansvar utifrån det som har sagts hittills. Mer specifikt diskuterar vi moraliskt ansvar för implicit bias på individnivå och på kollektiv nivå. Kan *jag* vara moraliskt ansvarig för *mina* implicita fördomar? Och kan *vi* som företag eller nation vara ansvariga för *våra* implicita fördomar?

6. MORALISKT ANSVAR

Individnivå

Jennifer Saul anför tre skäl mot att tillskriva ansvar – klandra människor – för deras implicita fördomar. Dessa skäl grundar sig i tre, som hon anser nödvändiga, villkor för att kunna tillskriva moraliskt ansvar, vilka enligt det vi vet om implicit bias inte är uppfyllda:

> En person bör inte klandras för implicit bias som hen är [i] helt omedveten om, som [ii] härstammar enbart från faktumet att hen lever i en sexistisk kultur. Även när hen blir medveten om att hen sannolikt har implicit bias, blir hen [iii] inte omedelbart förmögen att kontrollera sina bias, och därför bör hen inte klandras för dem.[45]

Innan dessa villkor – och de argument de är tänkta att ge upphov till – kan diskuteras behöver vi uppmärksamma en otydlighet i Sauls resonemang. Vad exakt vi har ansvar *för* kan förstås på minst tre olika sätt. Jules Holroyd skiljer mellan ansvar (1) för att *ha* en implicit fördom; (2) för att denna *manifesteras* i beteende; och (3) för att inte *agera* på kunskapen om att man själv sannolikt har implicita fördomar.[46] Saul skiljer inte mellan (1) och (2) medan hon är tydlig med att (3) medför klandervärdhet.

Utifrån vår horisont framstår Holroyds alternativ (1) som föga relevant. Vårt intresse av fenomenet implicit bias grundar sig, som sagt, i

45. Saul 2013b, s. 55, vår övers.
46. Holroyd 2012.

vår oro att det påverkar vårt handlande, det vill säga orsakar diskriminering. Fenomenet blir normativt intressant – även när det gäller ansvarsfrågan – bara i den mån det manifesteras i det vi gör. I debatten om implicit bias och ansvar är det också vanligast att man fokuserar på (2). Vi kommer därför här att begränsa oss till att undersöka de utmaningar detta ger upphov till, och slutligen återkomma till (3) i nästa avsnitt.

Medvetenhetsargumentet (baserat på villkor (i) i Sauls citat) innebär i vårt sammanhang i korthet att (a) individer inte är medvetna om att implicita fördomar påverkar deras handlande, och att (b) de kan klandras endast för handlingar om de är medvetna om de kognitiva tillstånd som påverkar dessa, varför de (c) inte kan klandras för de handlingar som påverkas av implicit bias.

Holroyd invänder mot (a) genom att hänvisa till studier som visar att en del försökspersoner – även om de är omedvetna om sina implicita fördomar – upplever en diskrepans mellan sina handlingar och sina värderingar och övertygelser, såsom att handlingarna ”svek idealen”.[47] De är alltså medvetna om att *något* påverkar det de gör i fel riktning. Denna invändning verkar dock enbart framkalla en ny källa till oenighet: hur diffus kan en sådan ”medvetenhet” tillåtas vara för att utgöra ett nödvändigt villkor för moraliskt ansvar?

Holroyd framför dock även en invändning mot premiss (b). Att individer måste vara medvetna om de kognitiva tillstånd som influerar deras handlingar är ett för starkt villkor för moraliskt ansvar, eftersom det leder till en orimlig slutsats: global eller allmän skepticism gällande moraliskt ansvar. Vi är aldrig medvetna om alla eller ens de mest avgörande kognitiva tillstånd som påverkar vårt handlande. Om vi vill undvika denna globala skepticism – och specifikt, om vi vill kunna säga att ansvarstillskrivningen ser *olika* ut, beroende på om vi handlar utifrån implicit bias eller ”klassiska” Humeanska skäl – bör vi förkasta (b).

Ursprungsargumentet (baserat på villkor (ii) i Sauls citat) går ut på att (a) individer inte kan klandras för handlingar som manifesterar kognitiva tillstånd som orsakas av faktorer som helt ligger utanför deras kontroll, att (b) implicita fördomar orsakas uteslutande av att leva i ett rasistiskt eller sexistiskt samhälle och att (c) det är bortom en individs kontroll huruvida hen lever i ett rasistiskt eller sexistiskt samhälle. Slutsatsen blir att (d) i våra sexistiska och rasistiska samhällen kan ingen klandras för att manifestera implicit bias.

47. Ibid., s. 294, vår övers.

Holroyd ifrågasätter att människors implicita fördomar uteslutande beror på det samhälle de lever i. Olika individer inom samma kultur kan ha olika grader av implicita fördomar, och vissa individer har inga alls. Denna variation kan förklaras, med att individuella skillnader – exempelvis i explicita, självrapporterade attityder – spelar roll. Vissa studier visar att individer som har värderingen att det är intrinsikalt viktigt att agera fördomsfritt uppvisar mindre bias i sitt beteende än individer som anser att detta är viktigt på grund av sociala normer och sanktioner.[48] Om detta stämmer så faller ursprungsargumentets premiss (b). Men är inte grundtanken bakom detta argument att kontroll är nödvändigt för ansvar – och att vi saknar kontroll över huruvida vi manifesterar våra implicita fördomar i handlingar?

Kontrollargumentet (baserat på villkor (iii) i Sauls citat) är det starkaste och mest intressanta. I vårt sammanhang går det ut på att (a) individer inte kan klandras för handlingar som manifesterar kognitiva tillstånd som de inte har omedelbar och direkt kontroll över, och att (b) implicita fördomar inte är under agentens omedelbara och direkta kontroll, varför (c) individer inte kan klandras för handlingar som manifesterar deras implicita fördomar.[49]

Holroyd framhåller att empiriska studier pekar i olika riktning när det gäller premiss (b). I vissa studier verkar individer kunna begränsa sina implicita fördomar under en kort stund, men de riskerar senare att återkomma med större kraft. I stället kretsar den fundamentala oenigheten mellan Holroyd och Saul, det vill säga mellan de som menar att moraliskt ansvar kan tillskrivas och de som förnekar detta, kring premiss (a). Holroyd menar att kontroll – förstått som möjligheten att kunna handla annorlunda – är ett nödvändigt villkor för moraliskt ansvar, men att direkt och omedelbar kontroll inte är nödvändig. Hon skiljer mellan denna typ av kontroll och långsiktig kontroll. Vi har långsiktigt kontroll över aktiviteter som att lära oss att spela piano eller att gå upp eller ner i vikt. Detta eftersom vi har direkt kontroll över en serie av mellanliggande steg: att placera händerna på tangenterna, att öka eller minska vårt dagliga matintag med mera. Vi kan ses som ansvariga när det gäller dessa mellanliggande steg och därmed även när det gäller aktiviteterna i sin helhet. Likaså vad gäller exempelvis våra övertygelser: vi kan inte välja huruvida vi ska tro på *p*. Men vi kan genom en serie mellanliggande steg – att söka, pröva och väga evidens av olika slag – komma fram till

48. Se t.ex. Devine m.fl. 2002.
49. Holroyd 2012, s. 282.

övertygelsen att *p*. Om vi antar att individer har långsiktigt kontroll över sina implicita fördomar blir frågan: finns det saker vi kan göra som bidrar till att på sikt minska manifestationen av implicit bias i vårt handlande, likt de steg vi tar för att lära oss spela piano, förändra vår vikt eller våra övertygelser? Det vill säga, kan vi utöva långsiktig kontroll genom att utöva omedelbar och direkt kontroll över exempelvis serien av steg för att reducera implicita fördomar?

Här finns gott om så kallade interventionsstudier som pekar mot ett jakande svar. Vi verkar kunna minska manifestationen av implicit bias i en beslutssituation genom att, i förväg, tänka på ”kontra-stereotypiska exemplar” eller genom att umgås med medlemmar ur den stigmatiserade gruppen; genom att forma så kallade ”implementeringsavsikter” (”om jag möter en kvinna som är matematiker ska jag bete mig professionellt”); genom att aktivera ovannämnda värdering, eller att det är intrinsikalt viktigt att agera fördomsfritt. Även om effekterna av dessa interventioner är kortlivade kan vi genom att upprepa dessa metoder inför nya beslutssituationer verka för att minska manifestationen av implicit bias i vårt handlande. Holroyd menar därför att vi har långsiktig kontroll – och att denna är tillräcklig för att tillskriva moraliskt ansvar.

En kvarstående utmaning är nu följande: mycket av den diskriminering som manifesterar implicita fördomar tycks utgöras av mikrobeteenden: Holroyd själv diskuterar exempelvis ett ökat flackande med blicken, minskad ögonkontakt, valet av en stol lite längre bort, i mötet med medlemmar av den stigmatiserade gruppen jämfört med andra. Men *hur mycket* klander kan vi egentligen rimligen dela ut när en person flackar med blicken, tittar bort, eller sätter sig lite längre ifrån en annan? Vi riskerar att hamna i ett dilemma: när klandret är tillräckligt starkt för att märkas så är det redan oproportionerligt med avseende på beteendets allvarlighetsgrad.

Vi har visserligen hävdat att i vissa fall, exempelvis när det gäller CV-rankningar eller polisskjutningar, så kan denna diskriminering vara allvarlig, även livsavgörande. Och i de fallen kan – med Holroyds argument – ansvar tillskrivas, tydligt och kraftfullt. Men om samhällets stora ojämlikheter mellan olika grupper har sitt ursprung i många individers var för sig mikroskopiskt små ”mikro-orättvisor” (så som vi föreslog i avsnitt fem), måste inte vi då hela tiden och mycket småskaligt klandra de allra flesta (närmast urskillningslöst), alternativt ibland (och därmed godtyckligt) välja vissa mikro-orättvisor och klandra deras

upphovspersoner bortom alla proportioner? Och blir inte i annat fall ansvarsfrågan här till stora delar praktiskt irrelevant?

Vi kommer i nästa avsnitt att argumentera för att dessa farhågor är missriktade och beror på en för närsynt syn på ansvarstillskrivning som en uteslutande individuell fråga. När vi lyfter blicken till den kollektiva nivån får frågan om moraliskt ansvar både en ny teoretisk ställning och en ny praktisk relevans.

Från individuellt till kollektivt moraliskt ansvar

Den nuvarande debatten om implicit bias och moraliskt ansvar inriktar sig alltså främst på individer och deras beteenden. Är en person moraliskt ansvarig för att ha implicit bias, för att manifestera dessa, eller för att inte agera på kunskapen att hen med stor sannolikhet har implicit bias? Denna individualistiska tendens återfinns även, som vi redan sett, i förslag på lösningar: olika försök att förändra individen är den vanligast föreslagna lösningen. Men detta fokus på individen och moraliskt ansvar är för snävt eftersom det finns andra relevanta frågor gällande implicit bias: Kan vi som kollektiv – säg en nation, klubb eller styrelse för en organisation – vara moraliskt ansvariga för implicit bias hos individer? Och kan jag som gruppmedlem, säg som medlem i en anställningskommitté, hållas moraliskt ansvarig för hur implicita fördomar påverkar utfallet?

Detta individualistiska perspektiv bör alltså kompletteras med ett kollektivt perspektiv. Det finns minst tre anledningar till ett sådant perspektivskifte. För det första, i många av de sammanhang där implicit bias är särskilt problematiskt – exempelvis vid rekryteringar och rankningar av jobbkandidater – agerar vi inte enbart som enskilda agenter, utan också som gruppmedlemmar, säg i en anställningskommitté. Betänk också att HR-avdelningen på ett företag är en grupp som utför den kollektiva handlingen att utesluta vissa CV från vidare granskning och att inte kalla vissa personer ur stigmatiserade grupper till intervju trots att personerna har lämpliga kvalifikationer. Det är därför centralt att även diskutera kollektivt ansvar – både ansvaret hos enskilda gruppmedlemmar och ansvaret hos gruppen som sådan – för implicit bias.

För det andra krävs kollektiv handling, i kontrast till enbart individuell handling, för att påverka orsakerna bakom att individer har implicita fördomar. Exempelvis kan det krävas en förändring av sociala normer, policyer eller lagstiftning för att påverka den långsammare associationen mellan kvinnor och naturvetenskap än den mellan män

och naturvetenskap. Att exempelvis genomföra politiska förändringar som syftar till att öka andelen kvinnor inom vissa vetenskapliga områden kräver ofta kollektiva åtgärder.

För det tredje är implicit bias i en viktig mening ett samhälleligt eller kollektivt fenomen snarare än ett individuellt fenomen: dess förekomst hos individer beror på hur vi har organiserat samhället. Jennifer Saul påpekar exempelvis att den främsta orsaken till att individer har implicita fördomar av sexistiskt slag är att de lever i ett sexistiskt samhälle, och hon menar att den viktigaste lösningen just återfinns på samhällsnivå:

> Den enda sättet att bli helt fri från bias-relaterat tvivel är att skapa en social värld där de stereotyper som nu förvränger våra omdömen inte längre håller oss i sitt grepp. Och sättet att göra det är att eliminera de samhälleliga regelbundenheter som ger upphov till och stödjer dessa stereotyper.[50]

Vår poäng är alltså att debatten om moraliskt ansvar för implicita fördomar behöver utvidgas avseende agenten, eller vem som kan hållas moraliskt ansvarig för implicita fördomar i individer. I samband med detta har vi gjort en åtskillnad mellan de fall där vi agerar som enskilda agenter, de fall där vi agerar som gruppmedlemmar och de fall där gruppen som sådan handlar. De två senare aktörerna har inte behandlats i debatten hittills, samtidigt som just dessa fall ofta är de viktigaste. Sammanfattningsvis har den övergripande frågan om kollektivt moraliskt ansvar för implicit bias hittills blivit förbisedd. Vi menar att det både är fruktbart och nödvändigt att utvidga diskussionen till att även gälla kollektivt moraliskt ansvar (i betydelsen att agera som en gruppmedlem och att agera som en grupp) för implicit bias. En möjlig konsekvens av detta perspektivskifte (beroende på vilken teori om kollektivt moraliskt ansvar som förutsätts) är att vi är moraliskt ansvariga på andra sätt och i större utsträckning än vi tidigare trott. Jag som gruppmedlem kan exempelvis under vissa omständigheter vara (delvis) moraliskt ansvarig för vår kollektiva handling att inte anställa den lämpligaste kandidaten (till exempel om implicita fördomar gjort att vi inte kallat personer ur stigmatiserade grupper till intervju och/eller rankat dem på ett felaktigt sätt). Och vi som grupp kan vara moraliskt ansvariga för att inte agera på kunskapen om att vi med största sannolikhet har implicita fördomar. Exempelvis genom att låta bli att implementera beslutsprocesser

50. Saul 2012, s. 260–61, vår övers.

som gör att implicita fördomar inte ”kickar in” såsom att anonymisera ansökningar.

7. SLUTSATSER

Vi har argumenterat för tre övergripande poänger. För det första att den empiriska och filosofiska forskningen kring implicit bias fortfarande är i ett skede där den kan förstås i analogi med astrofysikens mörka materia – det finns något (implicit bias) som påverkar oss men vi har ännu inte en klar förståelse av dess natur och hur vi kan nå kunskap om den. För det andra att det är fruktbart att skifta från individ- till samhällsnivå avseende förklaringsmodeller. Mer specifikt är vår poäng att även om det enbart finns svaga korrelationer mellan implicita fördomar hos enskilda individer och deras diskriminerande handlingar, så kan dessa korrelationer – när de aggregeras – bli till märkbara och stora skillnader i utfall; exempelvis vilka personer som innehar vilka positioner, vilka personer som arbetar med humaniora respektive naturvetenskap, och vilka som kommer att arbeta med vetenskap överhuvudtaget. Implicita fördomar skulle alltså kunna vara ett sätt att förklara gapet mellan människors självrapporterade explicita positiva attityder till diversitet och den faktiska förekomsten av diskriminering i det svenska samhället. För det tredje att ett skifte från individnivå till kollektiv nivå är både nödvändigt och fruktbart utifrån en moralisk synvinkel: paradigmatiska och viktiga fall av diskriminering utifrån implicita fördomar sker när vi agerar i egenskap av gruppmedlemmar i exempelvis en anställningskommitté eller som en grupp, exempelvis som en myndighets personalavdelning.

LITTERATUR

Ahmadi, Fereshteh, Irving Palm och Nader Ahmadi. 2016. ”Mångfaldsbarometern”. FoU-rapport nr 44. Gävle: Högskolan i Gävle. http://hig.diva-portal.org/smash/get/diva2:1085307/FULLTEXT01.pdf

Ahmed, Ali M. och Mats Hammarstedt. 2008. ”Discrimination in the Rental Housing Market: A Field Experiment on the Internet”. *Journal of Urban Economics* 64 (2), s. 362–72. https://doi.org/10.1016/j.jue.2008.02.004.

Arai, Mahmood, Moa Bursell och Lena Nekby. 2016. ”The Reverse Gender Gap in Ethnic Discrimination: Employer Stereotypes of Men and Women with Arabic Names”. *International Migration Review* 50 (2), s. 385–412. https://doi.org/10.1111/imre.12170.

Arkes, Hal R. och Philip E. Tetlock. 2004. ”Attributions of Implicit Prejudice,

or 'Would Jesse Jackson "Fail" the Implicit Association Test?'". *Psychological Inquiry* 15 (4), s. 257–78.
Banks, R. Richard, Jennifer L. Eberhardt och Lee Ross. 2006. "Discrimination and Implicit Bias in a Racially Unequal Society". *California Law Review* 94 (4), s. 1169–90. https://doi.org/10.2307/20439061.
Bayne, Tim och Anandi Hattiangadi. 2013. "Belief and Its Bedfellows". I *New Essays on Belief*, red. Nikolaj Nottelmann. London: Palgrave Macmillan. https://doi.org/10.1057/9781137026521_7.
Briñol, Pablo, Richard Petty och Michael McCaslin. 2009. "Changing Attitudes on Implicit versus Explicit Measures: What is the Difference?" I *Attitudes: Insights from the New Implicit Measures*, red. Richard E. Petty, Russell H. Fazio, Pablo Briñol. New York: Psychology Press.
Di Bella, Laura, Eleanor Miles och Jennifer Saul. 2016. "Philosophers Explicitly Associate Philosophy with Maleness". I *Implicit Bias and Philosophy*, vol. 1, red. Michael Brownstein och Jennifer Saul, s. 283–308. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198713241.003.0012.
Brownstein, Michael. 2017. "Implicit Bias". I *The Stanford Encyclopedia of Philosophy*, red. Edward N. Zalta. Spring 2017. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2017/entries/implicit-bias/.
Bursell, Moa. 2014. "The Multiple Burdens of Foreign-Named Men – Evidence from a Field Experiment on Gendered Ethnic Hiring Discrimination in Sweden". *European Sociological Review* 30 (3), s. 399–409. https://doi.org/10.1093/esr/jcu047.
Correll, Joshua, Bernadette Park, Charles M. Judd och Bernd Wittenbrink. 2002. "The Police Officer's Dilemma: Using Ethnicity to Disambiguate Potentially Threatening Individuals". *Journal of Personality and Social Psychology* 83 (6), s. 1314–29. https://doi.org/10.1037/0022-3514.83.6.1314.
Cortina, Lilia M., Dana Kabat-Farr, Emily A. Leskinen, Marisela Huerta och Vicki J. Magley. 2013. "Selective Incivility as Modern Discrimination in Organizations: Evidence and Impact". *Journal of Management* 39 (6), s. 1579–1605. https://doi.org/10.1177/0149206311418835.
Devine, Patricia G., E Ashby Plan, David M. Amodio, Eddie Harmon-Jones, Stephanie L. Vance. 2002. "The Regulation of Explicit and Implicit Race Bias: The Role of Motivations to Respond Without Prejudice". *Journal of Personality and Social Psychology* 82(5), s. 835–48.
Ditonto, Tessa M., Richard R. Lau och David O. Sears. 2013. "AMPing Racial Attitudes: Comparing the Power of Explicit and Implicit Racism Measures in 2008". *Political Psychology* 34 (4), s. 487–510. https://doi.org/10.1111/pops.12013.
Dovidio, John F., Kerry Kawakami och Samuel L. Gaertner. 2002. "Implicit and Explicit Prejudice and Interracial Interaction". *Journal of Personality and Social Psychology* 82 (1), s. 62–68.
Forscher, Patrick, Calvin K. Lai, Jordan R. Axt, Charles R. Ebersole, Michelle Herman, Patricia G. Devine och Brian A. Nosek. 2018. "A Meta-Analysis of Procedures to Change Implicit Measures". https://psyarxiv.com/dv8tu/.
Gendler, Tamar Szabó. 2008a. "Alief and Belief". *The Journal of Philosophy*. 105 (10), s. 634–63.

Gendler, Tamar Szabó. 2008b. ”Alief in Action (and Reaction)”. *Mind & Language* 23 (5), s. 552–85. https://doi.org/10.1111/j.1468-0017.2008.00352.x.

Glüer, Kathrin och Åsa Wikforss. 2013. ”Aiming at Truth: On The Role of Belief”. *Teorema: International Journal of Philosophy* 32 (3), s. 137–62.

Greenwald, A. G. och M. R. Banaji. 1995. ”Implicit Social Cognition: Attitudes, Self-Esteem och Stereotypes”. *Psychological Review* 102 (1), s. 4–27.

Greenwald, A., D. McGhee och J. Schwartz. 1998. ”Measuring individual differences in implicit cognition: The Implicit Association Test”. *Journal of Personality and Social Psychology* 74, s. 1464–80.

Greenwald, Anthony G., T. Andrew Poehlman, Eric Luis Uhlmann och Mahzarin R. Banaji. 2009. ”Understanding and Using the Implicit Association Test: III. Meta-Analysis of Predictive Validity”. *Journal of Personality and Social Psychology* 97 (1), s. 17–41. https://doi.org/10.1037/a0015575.

Greenwald, Anthony G., Colin Tucker Smith, N. Sriram, Yoav Bar-Anan och Brian A. Nosek. 2009. ”Implicit Race Attitudes Predicted Vote in the 2008 U.S. Presidential Election”. *Analyses of Social Issues and Public Policy* 9 (1), s. 241–53. https://doi.org/10.1111/j.1530-2415.2009.01195.x.

Haslanger, Sally. 2013. ”Social Meaning and Philosophical Method”. Presidential Address presented at the Eastern Division of the American Philosophical Association. https://www.google.se/url?sa=t&rct=j&q=&esrc=s&source=web&cd=2&ved=0ahUKEwjP-N-KqaDZAhWChSwKHaD8AHAQFgguMAE&url=http%3A%2F%2Fsallyhaslanger.weebly.com%2Fuploads%2F1%2F8%2F2%2F7%2F18272031%2Fhaslangerapapresidentialaddress2013submitted.docx&usg=AOvVaw1Ev4ztu-xpnrlQpjGqwIcs.

Hehman, Eric, Jessica Flake och Jimmy Calanchini. 2018. ”Disproportionate Use of Lethal Force in Policing Is Associated With Regional Racial Biases of Residents”. *Social Psychological and Personality Science* 9 (4), s. 393–401.

Holroyd, Jules och Joseph Sweetman. 2016. ”The Heterogeneity of Implicit Bias”. I *Implicit Bias and Philosophy,* Vol. I, *Metaphysics and Epistemology*. Oxford: Oxford University Press. https://ore.exeter.ac.uk/repository/handle/10871/19147.

Holroyd, Jules. 2012. ”Responsibility for Implicit Bias”. *Journal of Social Philosophy* 43 (3), s. 274–306. https://doi.org/10.1111/j.1467-9833.2012.01565.x.

Jost, John T., Laurie A. Rudman, Irene V. Blair, Dana R. Carney, Nilanjana Dasgupta, Jack Glaser, Curtis D. Hardin. 2009. ”The existence of implicit bias is beyond reasonable doubt: A refutation of ideological and methodological objections and executive summary of ten studies that no manager should ignore”. *Research in Organizational Behavior* 29, s. 39–69.

Karpinski, Andrew och James L. Hilton. 2001. ”Attitudes and the Implicit Association Test”. *Journal of Personality and Social Psychology* 81 (5), s. 774–88.

Krumpal, Ivar. 2013. ”Determinants of social desirability bias in sensitive surveys: a literature review”. *Quality & Quantity* 47 (4), s. 2025–47.

Leslie, Sarah-Jane. 2017. ”The Original Sin of Cognition: Fear, Prejudice, and Generalization”. *Journal of Philosophy* 114 (8), s. 393–421.

Mandelbaum, Eric. 2016. ”Attitude, Inference, Association: On the Propositional Structure of Implicit Bias”. *Noûs* 50 (3), s. 629–58. https://doi.org/10.1111/nous.12089.

Martell, Richard F., Cynthia G. Emrich och James Robison-Cox. 2012. ”From Bias to Exclusion: A Multilevel Emergent Theory of Gender Segregation in Organizations”. *Research in Organizational Behavior* 32 (January), s. 137–62. https://doi.org/10.1016/j.riob.2012.10.001.

Martell, Richard F., David M. Lane och Cynthia G. Emrich. 1996. ”Male-Female Differences: A Computer Simulation”. *American Psychologist* 51 (2), s. 157–58.

Mekawi, Yara och Konrad Bresin. 2015. ”Is the Evidence from Racial Bias Shooting Task Studies a Smoking Gun? Results from a Meta-Analysis”. *Journal of Experimental Social Psychology* 61 (November), s. 120–30. https://doi.org/10.1016/j.jesp.2015.08.002.

Mella, Orlando, Fereshteh Ahmadi och Irving Palm. 2014. ”Mångfaldsbarometern”. Gävle: Högskolan i Gävle. https://www.hig.se/download/18.4db66897149b264e1ba4df24/1417115797086/mangfald.pdf.

Nosek, Brian A., Frederick L. Smyth, N. Sriram, Nicole M. Lindner, Thierry Devos, Alfonso Ayala, Yoav Bar-Anan, m.fl. 2009. ”National Differences in Gender–Science Stereotypes Predict National Sex Differences in Science and Math Achievement”. *Proceedings of the National Academy of Sciences* 106 (26), s. 10593–97. https://doi.org/10.1073/pnas.0809921106.

Oswald, Frederick L., Gregory Mitchell, Hart Blanton, James Jaccard och Philip E. Tetlock. 2013. ”Predicting Ethnic and Racial Discrimination: A Meta-Analysis of IAT Criterion Studies”. *Journal of Personality and Social Psychology* 105 (2), s. 171–92. http://dx.doi.org.ezp.sub.su.se/10.1037/a0032734.

Payne, B. K. 2001. ”Prejudice and Perception: The Role of Automatic and Controlled Processes in Misperceiving a Weapon”. *Journal of Personality and Social Psychology* 81 (2), s. 181–92.

Rooth, Dan-Olof. 2010. ”Automatic Associations and Discrimination in Hiring: Real World Evidence”. *Labour Economics* 17 (3), s. 523–34. https://doi.org/10.1016/j.labeco.2009.04.005.

Rubin, Vera C. 1983. ”Dark Matter in Spiral Galaxies”. *Scientific American* 248 (6), s. 96–109.

Rudman, Laurie A. och Peter Glick. 1999. ”Feminized Management and Backlash toward Agentic Women: The Hidden Costs to Women of a Kinder, Gentler Image of Middle Managers”. *Journal of Personality and Social Psychology* 77 (5), s. 1004–10.

Rudman, Laurie A. och Peter Glick. 2001. ”Prescriptive Gender Stereotypes and Backlash Toward Agentic Women”. *Journal of Social Issues* 57 (4), s. 743–62. https://doi.org/10.1111/0022-4537.00239.

Saul, Jennifer. 2013a. ”Scepticism and Implicit Bias”. *Disputatio* V (37), s. 243–63.

Saul, Jennifer. 2013b. ”Implicit Bias, Stereotype Threat, and Women in Philosophy”. I *Women in Philosophy: What Needs to Change?*, red. Katrina Hutchison and Fiona Jenkins. Oxford: Oxford University Press. http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780199325603.001.0001/acprof-9780199325603-chapter-3.

SCB, Statistics Sweden. 2018. ”På tal om kvinnor och män – Lathund om jämställdhet 2018”. https://www.scb.se/hitta-statistik/statistik-efter-amne/lev-

nadsforhallanden/jamstalldhet/jamstalldhetsstatistik/pong/publikationer/pa-tal-om-kvinnor-och-man.-lathund-om-jamstalldhet-2018/
Schelling, Thomas C. 1971. ”Dynamic Models of Segregation”. *The Journal of Mathematical Sociology* 1 (2), s. 143–86. https://doi.org/10.1080/002225 0X.1971.9989794.
Schwitzgebel, Eric. 2013. ”A Dispositional Approach to Attitudes: Thinking Outside of the Belief Box”. I *New Essays on Belief*, red. Nikolaj Nottelmann, s. 75–99. London: Palgrave Macmillan. https://doi.org/10.1057/9781137026521_5.
SIFO. 2014. ”Allt Fler Svenskar Är Feminister”. 2014. http://opinion.se/sakfragor/allt-fler-svenskar-ar-feminister.
Unkelbach, Christian, Joseph P. Forgas och Thomas F. Denson. 2008. ”The Turban Effect: The Influence of Muslim Headgear and Induced Affect on Aggressive Responses in the Shooter Bias Paradigm”. *Journal of Experimental Social Psychology* 44 (5), s. 1409–13. https://doi.org/10.1016/j.jesp.2008.04.003.
World Values Survey Association. 2014. ”World Values Survey, Wave 6 (2010-2014): Sweden 2011- 2014”. http://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp.

LARS BERGSTRÖM

GILTIGA SLUTLEDNINGAR

En slutledning ska innehålla en eller flera premisser samt en slutsats. Men ordet ”slutledning” är inte entydigt. Det kan syfta på endera av åtminstone tre olika saker. Det kan stå för en konkret handling, där en bestämd person vid en viss tidpunkt utför en tankegång från premisser till slutsats. Eller för en handlingstyp, som kan utföras vid flera olika tillfällen, där det gemensamma är premisser och slutsats. Eller för enbart själva kombinationen av premisser och slutsats.

Jag ska här utgå från denna tredje innebörd. Den överensstämmer med vad man kan läsa i den svenska *Nationalencyklopedin*, där det inte ställs några särskilda krav på hur man *gör* giltiga slutledningar, utan endast på själva förhållandet mellan premisser och slutsats:

> En slutledning innehåller dels ett antal *premisser*, dels en *slutsats*. Den är *giltig* om slutsatsen *följer av* premisserna, dvs. om slutsatsen måste vara sann ifall alla premisserna är sanna, eller med andra ord om det är omöjligt att premisserna är sanna men slutsatsen falsk. […] Logiken syftar till att dels ge en systematisk framställning av giltiga slutledningar, dels svara på frågan vad det innebär att en slutledning är giltig.

Låt oss för enkelhetens skull säga att en slutledning, som uppfyller kravet om att det är ”omöjligt att premisserna är sanna men slutsatsen falsk”, är *sanningsbevarande*. Sådana slutledningar brukar också kallas ”deduktiva”, till skillnad från andra slutledningar som är ”induktiva”.

Det kan tyckas lite egendomligt att *Nationalencyklopedin* säger att ett av logikens syften är att ”svara på frågan vad det innebär att en slutledning är giltig”. Har den inte just själv besvarat den frågan? Har den inte just sagt att giltiga slutledningar är de som är sanningsbevarande?

Jo, det har den ju sagt. Men den menar antagligen att svaret fortfarande är alltför obestämt, innan man har förklarat vad de centrala begreppen ”omöjligt”, ”sant” och ”falskt” betyder?

Eller menar *Nationalencyklopedin* att även om det gäller att slutledningar är (deduktivt) giltiga om och endast om de är sanningsbevarande, så *betyder* ”giltig” eventuellt inte detsamma som ”sanningsbevarande”?

Filosofisk tidskrift 2019 nr 4, 27–32

Man kunde kanske anse att giltighet, till skillnad från sanningsbevarande, är något *normativt* – något som rent begreppsligt har att göra med hur man bör eller får resonera – vilket gör att de två orden inte är helt synonyma.

Kanske är det så. Eller kanske inte. Min gissning är att bägge tolkningarna förekommer – även bland logiker – även om de inte tydligt hålls isär.

En annan sak man kan undra över är *logikens* roll i sammanhanget. Av *Nationalencyklopedin* kan man ju få intrycket att logiken handlar om deduktiva slutledningar i största allmänhet, men det finns såvitt jag förstår många deduktivt giltiga (sanningsbevarade) slutledningar som inte skulle betraktas som ”logiskt” giltiga och om vilka logiken inte har något att säga.

Ett exempel kunde vara följande:

(1) X är större än Y
Y är större än Z
———————
X är större än Z

Om X är större än Y och Y är större än Z, så *måste* ju X vara större än Z. Men detta beror på betydelsen av orden ”större än”, inte på några logiska principer. (1) är en deduktivt giltig slutledning, men den skulle nog inte räknas som logiskt giltig. Något liknande gäller följande slutledning.

(2) Påven vet att Gud existerar
———————————
Gud existerar

och kanske också:

(3) Påven är ungkarl
———————
Påven är ogift

De flesta skulle nog hålla med om att (1), (2) och (3) är sanningsbevarande. Det är nödvändigt att slutsatserna är sanna om premisserna är sanna. Men man brukar tänka sig att en *logiskt* giltig slutledning är giltig i kraft av sin ”logiska form”, något som i sin tur bestäms av de logiska uttryck – så kallade ”logiska konstanter” – som ingår i premisser och slutsats. Logiska konstanter är t.ex. uttryck som ”icke”, ”eller”, ”om …, så …”, ”alla” och ”någon”. Beträffande (2) kunde kanske någon

vilja säga att ”vet” är en logisk konstant och att det är i kraft av dess betydelse som (2) är sanningsbevarande. Då kunde även (2) anses vara logiskt giltig. Men ingen skulle nog säga att (1) och (3) är logiskt giltiga.

Följande är däremot ett klassiskt läroboksexempel på en *logiskt* giltig slutledning. Dess giltighet beror huvudsakligen på betydelsen hos den logiska konstanten ”alla”:

(4) Alla människor är dödliga
Sokrates är en människa

Sokrates är dödlig

Om man plockar bort den första premissen i (4) får man:

(5) Sokrates är människa

Sokrates är dödlig

vilket fortfarande tycks vara en deduktivt giltig, men inte längre logiskt giltig slutledning. Den som betvivlar att (5) är deduktivt giltig (sanningsbevarande) kan tydligen tänka sig en människa som inte är dödlig, men andra skulle säga att detta är en begreppslig – och kanske biologisk, men i alla fall inte logisk – omöjlighet. En varelse som liknar en människa i alla avseenden, utom i det att vederbörande är odödlig, skulle nog inte anses vara en människa.

Men detta är förstås inte alldeles säkert. Vad som är begreppsligt nödvändigt är ganska oklart och kan diskuteras. (1) är nog okontroversiell, (2) är kanske något mer tveksam och (5) ganska diskutabel.

Någon skulle kanske vilja gå ännu längre och tala om slutledningar som är sanningsbevarande i kraft av andra sorters nödvändighet. Som exempelvis denna:

(6) Den åtalade har torterat små barn för nöjes skull

Den åtalade har handlat moraliskt orätt.

Här är det fråga om någon sorts *moralisk* nödvändighet. Kanske är nödvändigheten även begreppslig? Det är inte lätt att säga. Att slutledningen är giltig skulle i alla fall många gå med på, men den är inte logiskt – och kanske inte ens deduktivt – giltig. (Enligt en tolkning av det som kallas Humes Lag kan slutledningar av detta slag inte vara deduktivt giltiga.) Och logiken brukar inte uppfattas som att den ska

”ge en systematisk framställning av giltiga slutledningar” av bland annat detta slag.

En tänkbar invändning mot att (6) skulle vara sanningsbevarande är att slutsatsen inte *kan* vara sann. Detta kunde då antas bero på den mer generella teorin att moraliska omdömen inte kan vara sanna. Men den teorin är omstridd och jag ska här inte säga mer om detta.

Sedan kan man tänka sig slutledningar som är problematiska av andra skäl. Som till exempel denna:

(7) Denna premiss är falsk

Premissen i denna slutledning är sann.

Att (7) är deduktivt giltig kan verka självklart. Premissen kan inte gärna vara sann, utan att vara sann. Men man kan bli osäker när man betänker att även följande slutledning kan tyckas vara giltig, samtidigt som de båda slutsatserna i någon mening är oförenliga:

(8) Denna premiss är falsk

Premissen i denna slutledning är falsk.

Jag ska här inte ge mig in på något försök att bedöma olika sätt att hantera problem av det slag som (7) och (8) tycks ge upphov till.

Det som här framför allt intresserar mig är hur man ska förstå *Nationalencyklopedins* påstående att logiken syftar till att ”dels ge en systematisk framställning av giltiga slutledningar, dels svara på frågan vad det innebär att en slutledning är giltig”.

Svaret på frågan om innebörden av giltighet kan tyckas vara att ”giltig” betyder ”sanningsbevarande” – vilket ju är en rätt allmänt vedertagen förklaring, som också uttryckligen anges i *Nationalencyklopedin.* Men det strider mot tesen att logiken syftar till att ge en systematisk framställning av giltiga slutledningar. Logiken intresserar sig exempelvis inte för sådant som (1), (3), (5) och (6).

Möjligen kunde man tänka sig att logiker än så länge bara intresserar sig för sådant som har med så kallade logiska konstanter att göra, men att de på längre sikt ska systematisera deduktiva slutledningar i största allmänhet. Kanske är det så logiker brukar uppfatta sitt ämnesområde. Men jag tvivlar på det. Jag tror de håller sig till logiska konstanter – även om de inte är överens om vilka dessa är.

Om en slutledning från A till B är sanningsbevarande i kraft av inne-

börden hos logiska konstanter, så kan man också säga att B är en *logisk konsekvens* av A. Men måste också en slutledning från A till B vara *giltig*, om B är en logisk konsekvens av A?

Vid närmare eftertanke skulle kanske många förneka det. Ty som jag antydde inledningsvis kan giltighet i samband med slutledningar uppfattas som att det har något att göra med hur man bör eller får resonera. Och man kan mycket väl tänka sig att B är en logisk konsekvens av A, trots att man inte bör eller får dra slutsatsen B *direkt* från A. Ett sådant resonemang kan vara kunskapsteoretiskt illegitimt.

För att ett resonemang ska vara giltigt i betydelsen kunskapsteoretiskt legitimt kan man kräva att det innebär en förklaring eller demonstration av *hur* B följer från A. Om en sådan förklaring saknas har man knappast något skäl att tro att B verkligen följer av A, även om det faktiskt råkar förhålla sig så.

I enkla fall behövs ingen närmare förklaring, eftersom vem som helst (med lämplig språklig och övrig kompetens) direkt kan *se* att B följer av A. Men om steget mellan A och B är längre och mer komplicerat, som exempelvis i ett långt matematiskt bevis, så kan det kännas väldigt konstigt att tala om en ”giltig slutledning”, om de mellanliggande stegen har plockats bort. Är det då ens en ”slutledning”?

En förklaring av hur B följer av A, i de fall där man inte direkt drar slutsatsen B från A, kan bestå i att man artikulerar sådana mellanliggande steg, som är tillräckligt enkla för att man ska se att vart och ett av dessa steg är giltigt och att de tillsammans från A leder fram till B.

För att en slutledningen ska vara kunskapsteoretiskt legitim – och i den meningen ”giltig” – skulle man alltså kunna kräva att man ska kunna *se* – eller *inse* – att slutsatsen följer av premisserna. Om man inte ser det, så ”gör” man förmodligen inte heller slutledningen.

Detta är kanske mest en terminologisk fråga. Man kan förstås kalla slutledningen ”giltig”, om man biter sig fast vid den definition som bl.a. *Nationalencyklopedin* anger. Men då behöver man ett annat ord för de giltiga slutledningar som har en mer kunskapsteoretisk relevans. De är fortfarande giltiga, men man kunde kanske säga att de dessutom är ”legitima”.

Själv tycker jag dock är det vore bättre att skilja mellan giltig slutledning och logisk konsekvens.

Giltiga, eller legitima, slutledningar ska då vara sådana att man kan se att slutsatserna följer av premisserna. Man kunde kanske tycka att detta extra villkor inför ett subjektivt element i logiken, som gör att den inte

kan göra anspråk på vetenskaplig objektivitet. Det hänger ju då på vad enskilda personer upplever. Men det ”seende” som krävs kan nog anses vara helt analogt till vetenskapliga observationer i allmänhet: man kan kräva att det ska vara *intersubjektivt* – och i den meningen ”objektivt”. Observationerna, bedömningarna, ska i princip kunna upprepas av vem som helst som är språkligt och intellektuellt kompetent och i övrigt har den uppfostran som krävs. Om slutledningen gäller specialområden som exempelvis högre matematik eller teoretisk fysik måste till exempel den nödvändiga kompetensen innefatta kunskaper på dessa områden. Liksom när det gäller andra observationer kan förstås observationer av sanningsbevarande vara reviderbara. Man kan ta miste. Även en mycket framstående matematiker kan tro att ett bevis är korrekt, trots att det senare visar sig innehålla ett fel eller en lucka.

(Någon skulle kanske invända att det här finns en viktig komplikation. Olika logiker tycks ju faktiskt inte se detsamma; deras logiska observationer är inte intersubjektiva. Anhängare av olika logiska system – t.ex. klassisk och intuitionistisk logik – tycks ju göra olika bedömningar. Men egentligen gör de nog inte olika bedömningar. Såvitt jag förstår är de överens om vilka slutledningar som är klassiskt giltiga och vilka som är intuitionistiskt giltiga. Och om de är oense om huruvida en viss slutledning är giltig i mer ”absolut” mening, så beror det på att de tolkar någon logisk konstant olika – eller på att de har olika uppfattning om vilken sorts sanning som bevaras, bivalent sanning eller bevisbarhet.)

Man kan förstås undvika det ”subjektiva” elementet i logiken, och även undvika uppfattningen att logiken har en normativ sida, om man bestämmer sig för en idé om ”logikens uppgift”, vad den ”syftar till”, som innebär att den enbart ska konstruera formella system som beskriver relationen ”logisk konsekvens” för givna språk. Det kan göras på många olika sätt. Men om man dessutom vill att logiken ska säga något om vad en kunskapsteoretiskt legitim slutledning är, så måste man nog acceptera något i stil med intersubjektiva observationer om sanningsbevarande.

Hur man ska välja mellan dessa två alternativa syften säger *Nationalencyklopedin* ingenting om. Bägge är väl helt respektabla. Men de bör inte blandas ihop. Ty då riskerar man att hamna i ofruktbara diskussioner om vad som kännetecknar giltiga slutledningar och om huruvida logiken är normativ eller inte.

RUBEN WIWE

HAR VI *VERKLIGEN* EN FRI VILJA?

Vi upplever oss som fria, men är vi *verkligen* fria? Så brukar problemformuleringen lyda. Alla som läser denna artikel vet sannolikt vad det innebär att fatta ett beslut mellan olika alternativ. Ibland är det enkelt, ibland svårt, ibland nära nog omöjligt. Vi upplever det dock, i vanliga fall, som om vi har olika alternativ och en möjlighet att välja fritt bland dessa, även om det, som sagt, ofta kan vara svårt att välja. Vissa filosofer menar att detta är en illusion. De menar att det som vi vill och väljer redan är bestämt på förhand. De menar att allt är förutbestämt, även vår vilja och våra val. Även om vi alltså upplever att vi kan välja mellan flera alternativ, så kan vi i själva verket inte välja alls. Valet är redan förutbestämt och kan därför varken vara ett val eller fritt.

Argumentationen för detta synsätt har ungefär följande uppbyggnad. Allt i universum följer lagen om orsak–verkan. Atomer följer denna lag. Vi består av atomer. Vår hjärna består av atomer. Våra tankar och vår vilja uppstår i hjärnan och måste således också vara en följd av lagen om orsak–verkan. Vi kan inte styra hur atomerna rör sig, och vi kan därmed ej styra vilka tankar eller viljor som orsakas hos oss. När vi alltså tror att vi väljer fritt, så är det redan, på atomär eller subatomär nivå, orsakat vad vi ska vilja och välja, eftersom den materia som vi består av, i kombination med all annan materia i universum, tvingande och med nödvändighet har lett fram till just denna situation i kraft av lagen om orsak–verkan. Allt är förutbestämt och kommer att ske med nödvändighet, även våra val. Vi känner inte ännu till det som kommer att ske, men det kommer att ske och det finns endast en väg, ett alternativ, eftersom allt måste följa orsakslagen. Detta gäller även för en subatomär nivå, där slumpmässighet sägs kunna råda, ty slumpen kan inte vara grunden för frihet.

Om ovanstående argumentation är sann, så är allt i universum ett skeende. Ett skeende innebär att något bara händer. Det föreligger ingen vilja eller medvetenhet bakom ett skeende. När ett träd faller i en storm, så är det ett skeende. Ingen gör något, ingen vill något, det finns ingen avsikt; det blåser och trädet faller. Punkt!

Om ovanstående argumentation är sann, så är människor avancerade

Filosofisk tidskrift 2019 nr 4, 33–37

robotar. Det är nämligen skillnad mellan att *utföra en handling* och *ett skeende*. En robot utför aldrig en handling. En robot följer ett protokoll, eller närmare bestämt, den är sitt protokoll. Att säga att en robot *utför en handling*, är liktydigt med att säga att *ett protokoll följs* eller *något sker*. Människor skiljer sig dock i ett viktigt avseende från traditionella robotar; människor har ett medvetande. Enligt ovanstående argumentation är emellertid allt i medvetandet, inklusive vår vilja och våra val, redan bestämt på förhand på atomär eller subatomär nivå. Det innebär att medvetandet är en passiv biprodukt av ett skeende. Det kan jämföras med ett piano och pianots toner. På samma sätt som pianots toner är en passiv biprodukt av att det spelas på pianot, så är allt i medvetandet en passiv biprodukt av att det sker saker på en atomär nivå i hjärnan. *Hjärnan är pianot och medvetandet dess melodi.* När människan väljer fritt, så väljer hon i själva verket inte alls, utan det är ett skådespel, ett protokoll som körs, något som redan skett på atomär nivå i hjärnan och som hon nu åser som passiv åskådare i tron att hon är den som aktivt väljer. Allt är bara skeende. Ett lurendrejeri på hög nivå!

Jag kommer att argumentera mot ovanstående uppfattning. Först kommer jag att visa att ovanstående argument mot den fria viljan vilar på oförenliga premisser. Därefter kommer jag att visa att vi har en fri vilja på det enda sätt som det är möjligt.

Låt oss börja med att visa att ovanstående argument mot den fria viljan vilar på oförenliga premisser.

I analogi med frågan ”Har vi *verkligen* en fri vilja?”, ställer vi nu en annan fråga. Vi ställer frågan: ”Finns det hammare?” och mer specifikt: ”Vi upplever att det finns hammare, men finns det *verkligen* hammare?”

Alla som läser denna artikel har sannolikt upplevt och använt en hammare. Vad är det alltså man vill åt med frågan om det *verkligen* finns hammare? Det troliga svaret är att man undrar hur en hammare ter sig utanför ett medvetande, hammaren i-sig, hammaren som atomer eller hammaren på valfritt sätt utanför medvetandet. Det korta svaret på frågan om hammare verkligen finns är: Det finns inga hammare utanför ett medvetande. Hammare är medvetandeobjekt. Människor har endast tillgång till medvetandeobjekt och medvetandeföreteelser. En människas själva existens konstitueras av att hon ständigt befinner sig i en värld som är full av för henne meningsfulla medvetandeobjekt och medvetandeföreteelser, som t.ex. hammare, pianon, berg, solen, abstrakta begrepp som moral och rättvisa. Att fråga om hammare utanför en medvetandesfär är således redundant, då det är konstitutivt för en

hammare att den är ett medvetandeobjekt. Det vore som att fråga en ungkarl om han är gift.

Men det kanske ändå är lockande att tänka kring vad, om något, utanför medvetandet som, *objektivt,* ger upphov till medvetandeobjekt som hammare osv. Vad är det som, så att säga, ligger till grund för vår uppfattning av hammaren? Det kan vara en spännande tankeexercis, men det förändrar inte hammaren som medvetandeobjekt, den fungerar lika bra eller dåligt som innan. Problemet uppstår dock när man börjar dra slutsatser ur dessa tankeövningar och överföra dessa slutsatser på medvetandeobjekten. Utifrån hammaren som medvetandeobjekt tänker man fram hur den skulle kunna te sig utanför ett medvetande, vilket vi har sett är en omöjlighet, och återför sedan slutsatserna om den bortomempiriska, endast tänkta hammaren på hammaren som medvetandeobjekt. Man säger kanske, till exempel, att en hammare endast är en samling atomer på olika avstånd, vilket måhända är fallet på en annan nivå, men det är ointressant för argumentet i denna text. Man opererar alltså på olika nivåer och vill använda premisserna från den ena nivån som premisser på den andra nivån. *Men*, man kan inte hamra eller spela piano med atomer; man kan hamra och spela piano med pianon och hammare. Man använder sig således av oförenliga premisser om man försöker argumentera om medvetandeobjekt med stöd i premisser hämtade från tänkta, bortomempiriska storheter.

Nu kanske någon invänder att naturlagar, atomer och tankar på bortomempiriska hammare-i-sig *också* är medvetandeföreteelser. Det är förvisso sant, men när de används i argumentationen mot den fria viljan som medvetandeföreteelse eller hammare som medvetandeobjekt, så används de *som om* de hade en objektiv existens utanför medvetandet och därför är *mer* sanna och *mer* riktiga än medvetandeobjekten själva. Att en hammare bara är en samling atomer ses som *mer* sant än vår upplevelse av hammaren som medvetandeobjekt. Men alla som har ett medvetande vet att en hammare inte bara är en samling atomer, och alla som har valt mellan olika alternativ vet att den fria viljan inte är något som bara sker på atomär nivå med oss som åskådare. Det är fullt möjligt att det finns en bortomempirisk värld, full av viljelösa atomer som följer lagen om orsak–verkan, men detta gör inte hammaren till mindre hammare, eller viljan till mindre vilja. Som medvetandeobjekt står de intakta kvar i vår upplevelse, och det är allt vi har tillgång till.

Vad vi har sett hittills är att medvetandeobjekt och medvetandeföreteelser, det vill säga summan av alla konkreta och abstrakta objekt,

existerar i en egen sfär som inte låter sig påverkas av premisser hämtade från atomära, subatomära eller bortomempiriska nivåer. Oberoende av vilka premisser vi ställer upp om en bortomempirisk ”hammare”, så kommer hammaren som medvetandeobjekt fortbestå oföränderlig som ett verktyg som vi kan hamra med i vår värld av medvetenhet. Hammaren är obegriplig utanför och kan endast existera i ett mänskligt medvetande.

Utifrån denna argumentation ska vi nu visa att den fria viljan existerar och att vi kan välja fritt. Som alla läsare säkert redan har förstått, förhåller det sig med den fria viljan som med hammaren och övriga medvetandeobjekt och medvetandeföreteelser. Den fria viljan är en medvetandeföreteelse. Den är inget medvetandeobjekt som hammaren, men en företeelse som endast kan äga rum i ett medvetande. Det finns ingen vilja utanför medvetandet, och en bortomempirisk, omedveten vilja kan inte ens tänkas. En atom vill ingenting. Frågan om vi verkligen har en fri vilja kan därför tyckas egendomlig. Alla som någon gång har stått inför flera alternativ vet vad det innebär att välja. Man ska välja mellan A, B eller C. Det är svårt och man väger för och emot. Till slut kanske man tveksamt väljer A. Man kanske ångrar sig. Men man valde, fritt. I vissa fall kan man kanske *känna sig* tvingad av normer eller yttre förhållanden, men det tar inte udden av argumentet. Vi har alla upplevelsen av att vi kan välja, på gott och ont.

Om vi återvänder till argumentet mot den fria viljan så kan vi nu se att detta argument är konstruerat på följande, felaktiga, sätt. Först utgår man från den medvetna upplevelsen av att kunna välja fritt mellan olika alternativ, eftersom denna medvetandeföreteelse är det enda man har tillgång till. Därefter gör man bruk av sina premisser som man har hämtat från sina tankeövningar om bortomempiriska, endast tänkta, objekt. Man tänker: eftersom allt på en grundläggande, bortomempirisk, nivå är materia som måste följa lagen om orsak–verkan, så måste detta gälla även för medvetandeobjekt och medvetandeföreteelser. Men, som vi har sett så är detta att blanda ihop oförenliga premisser. På samma sätt som hammaren inte bara är en samling atomer, utan något som man kan hamra med i vår värld av medvetenhet, så är viljan inte bara en samling atomer som följer lagen om orsak–verkan, en ren fabrikation, utan ett instrument som vi använder oss av för att göra medvetna val mellan olika alternativ. Detta vet alla som någon gång har valt mellan olika alternativ.

Vilja och hammare kan endast existera på en medveten nivå, så

frågan om huruvida det *verkligen* finns hammare eller om vi *verkligen* väljer fritt är redundant annat än som tankeexperiment. Den fria viljan som medvetandeföreteelse och hammaren som medvetandeobjekt står orubbade kvar. Oavsett vilka tankar som görs om bortomempiriska fenomen, så kommer vi fortsätta att hamra med hammare och att välja fritt.

BERTIL KRISTERSON

EVOLUTIONÄR KÄRLEK

”Evolutionär kärlek” är namnet på en 30 sidor, obeaktad uppsats, publicerad 1893, av Charles Sanders Peirce (1839–1914), idag allmänt erkänd som USA:s främste filosof med ett växande inflytande inom framför allt semiotik och matematisk relationslogik. Under sin livstid var Peirce emellertid mer känd för sitt pragmatiska tänkande, som påverkat t.ex. religionshistorikern William James (1842–1910) och reformpedagogen John Dewey (1859–1952).

”Evolutionär kärlek” är tillsammans med nio andra uppsatser av Peirce översatta till svenska under titeln *Pragmatism och kosmologi* (Daidalos, 1990). Texterna är alla, som framhålls i förordet, krävande samtidigt som de ”når vida bortom den annars förkvävande fackspecialiseringen utan att … göra avkall på det vetenskapliga stringenta resonemanget … (vilket) kan leda till överraskande nivåförskjutningar i tänkandet och därigenom skapa ny insikt” (*Pragmatism och kosmologi,* s. 36).

1859 hade *Om arternas ursprung* publicerats. Denna epokgörande händelse i västerlandets idéhistoria hade slagit ner som en bomb inte bara för teologerna utan också för naturvetarna, som snart började använda Darwins upptäckt som tillhygge mot religionens existensberättigande. Peirce skriver om detta bombnedslag i ”Evolutionär kärlek”: ”Darwins arbete *Om arternas ursprung* utvidgar synen på framsteget till att omfatta djur- och växtriket i dess helhet … Vad djuren beträffar förstärker darwinismen individualismen såsom en kraft som verkar i god riktning tack vare djurets hänsynslösa glupskhet. Som Darwin själv uttrycker saken på titelbladet, handlar det om kampen för tillvaron; han kunde ha tillagt mottot: Varje individ för sig själv, och fan må ta den som kommer sist! I sin Bergspredikan framförde Jesus en annan uppfattning … Det finns naturligtvis mycket som talar för bägge dessa sidor. Jag har inte stuckit under stol med och skulle heller aldrig kunna dölja, vart min egen passionerade förkärlek pekar. En dylik bekännelse torde chockera mina vetenskapliga bröder” (s. 208–9).

Genom att på detta sätt polarisera Darwin och Jesus och göra dem till antipodiska sanningssägare inom var sitt verksamhetsområde, vill

Filosofisk tidskrift 2019 nr 4, 38–41

Peirce i ”Evolutionär kärlek” försöka ”försona den darwinska idén med kristendomen” (s. 216), i stället för att i likhet med de flesta av sina vetenskapliga bröder se Jesus och Darwin som varandras motsatser. Man skulle till och med kunna hävda, att han utifrån sin vetenskapssyn ser dem som logiskt sammanlänkade.

Peirces syfte med ”Evolutionär kärlek” är därför att inordna kristendomens historiesyn och moraliska budskap i sitt evolutionära tänkande; han vill visa att Darwins ”Glupskhetsevangelium” (s. 209) visserligen beskriver en reell tendens i naturen” (s. 29) men att den kan brytas med hjälp av evangeliet om Jesus.

Syftet med min uppsats blir därför att presentera både Peirces evolutionära tänkande och moraluppfattning som de låter sig tolkas utifrån ”Evolutionär kärlek” (s. 201–30) men också utifrån de övriga uppsatserna i *Pragmatism och kosmologi* och utifrån inledningspresentationen av Peirces filosofi.

I sitt evolutionära tänkande använder Peirce sig av tre begrepp, som han rätt och slätt kallar förstahet, andrahet och tredjehet (s. 21, 155–56) för att beskriva både hur den oorganiska och organiska naturen utvecklas, alltifrån sin uppkomst under förstahetens tid då allt var kaos och slumpen styrde (s. 208) till andrahetens tid då slumpen övergick till en mera lagbunden ordning (s. 209–11) och tredjehetens tid då till och med den lagbundna ordningen tack vare människans kunskapsförmåga började ifrågasättas (s. 213, 215).

Dessa tre enkla begrepp, som alltså syftar till att beskriva grundstrukturen i alla evolutionära processer, har visat sig mycket användbara. När han t.ex. applicerar dem på universums begynnelse, tänker han sig den som en explosion, där slumpen huserade fritt och inga naturlagar ännu hade utkristalliserats. Men allteftersom explosionen ebbade ut och hittade sin form uppstod naturlagarna; de kan därför inte ha funnits när tillvaron uppstod menar Peirce, utan bildats under evolutionens gång (s. 143–44). Han betraktar dem som i efterhand uppkomna ”*vanor i den oorganiska naturen*” (s. 24) och att de i en ”tidig fas av universums utveckling inte existerade” (s. 25). I inledningspresentationen till *Pragmatism och kosmologi* kommenteras denna radikala syn på naturlagarna: ”Även i dag verkar den avancerad, såtillvida som man måste söka sig ända fram till de allra färskaste kosmologiska utvecklingsteorierna för att kunna finna en liknande syn på saken” (s. 24–25, not 31).

När Peirce tänker sig hur den organiska naturen utvecklas använder han sig av samma begreppsapparat som när han beskriver uni-

versums utveckling. Han börjar med att tänka på livets ursprung i förstahetens tillstånd i en klump protoplasma och menar, att den då bara fanns till för sig själv innan den i andrahetens tillstånd genom påverkan utifrån börjar reagera (s. 21). Peirce kommenterar ett laboratorieförsök med en amöba: ”Lägg märke till hur den beter sig: Medan hela klumpen befinner sig i vila och är orörlig utsättes ett ställe på dess yta för retning. Just på den punkten sätts då en aktiv rörelse i gång och den breder efter hand ut sig till de övriga delarna av organismen” (s. 185). Peirce menar, att amöban efter flera sådana retningar utvecklar en lagmässig vana på samma sätt som tyngdlagen uppstod när universum expanderade under andrahetens tillstånd och att denna vana sedan genererar andra vanor med vars hjälp livet utvecklas till högre nivåer ända upp till människan.

I avsnittet ”Vanans lag” utvecklar Peirce hur detta evolutionära sätt att tänka kan tillämpas på hur människans själsliv utvecklas från allra första början där känslorna huserar slumpmässigt som hos ett nyfött barn, till reaktionsförnimmelser i andrahetens tillstånd av ibland våldsamt slag ”som när en person i blindo plötsligt springer mot en lyktstolpe” (s. 147) och vidare upp till tredjehetens tillstånd, där människan plötsligt begåvas med förmågan att själv kunna välja vilka känsloreaktioner hon vill göra bruk av. Peirce formulerar sig på följande sätt om denna viktiga händelse i människans utveckling: ”I sina befallningars natur uppvisar vanans lag en slående kontrast mot alla fysiska lagar. En fysisk lag är absolut. Den kräver en exakt relation … Den mentala lagen kräver däremot ingen exakt (relation) … Ja, (den) skulle rentav strida mot den … lagen … och förhindra all vanebildning” (s. 148–49).

Begreppet *vana* har därför en viktig roll att spela i Peirces evolutionära tänkande; han använder ordet som beteckning för hur både den oorganiska och organiska naturen utvecklas. I avsnittet ”Vid närmare eftertanke. Att stifta fred” (s. 209–18), som ingår i ”Evolutionär kärlek”, försöker han stifta fred mellan darwinismen och kristendomen genom att använda vanebegreppet också för att beteckna människans förmåga, att med hjälp av sin fria vilja fostra sig själv till att genom ”övning” (s. 213) och kunskap ”lära sig” (s. 215) att få sådana vanor, att hon kan motstå naturens lagmässiga framfart i själslivet och i stället fylla det med Guds osjälviska kärlek som den är formulerad i Johannesevangeliet: ”Var och en kan inse att aposteln Johannes ord är formeln för en evolutionär filosofi, en filosofi som lär att det sanna växandet kommer enbart ur kärleken” (s. 203).

Med hjälp av detta sätt att tänka, försöker Peirce att försvara kristendomens historiesyn och moraliska budskap gentemot Darwins glupskhetsfilosofi, som samhällets kelgrisar skor sig på för att sanktionera sina egna intressen och statens: ”(N)u har stjärnbaneret fladdrat mot himlen hela … 1800-talet igenom, och det med en fräckhet som får själva molnen att dra ihop sig till oväder. Snart kommer en blixt och en åskknall att ruska nationalekonomerna totalt ur deras självbelåtenhet … Den senare hälften av 1900-talet kommer helt säkert att få se syndaflodsovädret braka löst över den sociala ordningen, för att röja upp i en värld lika djupt sjunken i fördärv som den där glupskhetsfilosofin alltför länge har pressat ner i synd och skuld” (s. 207).

Varifrån kommer denna klarsyn om den sociala ordningens upplösning, som i vår tid blivit sannspådd, om inte från en rätt uppfattning om sambandet mellan Darwin och Jesus? Darwin som bevisade att det finns en orättvis tyngdlag bakom livets utveckling och kristendomen som i två tusen år motarbetat denna tyngdlag?

I inledningsavsnittet av ”Evolutionär kärlek” med rubriken ”Vid första ögonkastet. Motevangelier” namnger han ”den stora utvecklingskraften i universum” (s. 201) med hjälp av det grekiska ordet för gudskärlek, *agápe*, som en alternativ tyngdlag att formas av som är likformig med hela tillvarons beskaffenhet och inte bara med den lilla tyngdlag som härskar på jorden: en agapistisk utvecklingskraft som ”är en kärlek som inbegriper hatet som ett ofullgånget stadium av sig själv … Ty egenkärleken är ingen kärlek” (s. 202). Så paradoxalt kan Peirce uttrycka sig när han beskriver innehållet i Bergspredikan och som han själv vill formas av med argumentet att han då starkt känner, att han kan växa som människa: ”(E)nligt min mening (är) den starka känslan i sig ett vägande argument för den agapistiska evolutionsteorin – såtillvida som den kan tänkas ge uttryck åt det Känsliga Hjärtats spontana omdöme. Ifall det vore möjligt att tro på agapismen utan att samtidigt tro varmt på den, så skulle detta faktum vara ett argument mot sanningen i denna doktrin. Men när nu känslovärme existerar bör man … oförbehållsamt tillstå den” (s. 209, 215–16).

RECENSIONER

Moraliska provokationer: Om skenande tåg, rättvisa löner och god pornografi
Marcus Agnafors. Fri tanke 2017. 248 s. ISBN 978-91-87935-61-9

Marcus Agnafors har skrivit en bok vars syfte är ”att hjälpa läsaren att resonera kring rätt och fel” genom att ”provocera fram eftertanke och diskussion” (s. 7). Som bokens titel antyder så används provokationer för att åstadkomma detta. Dessa består i att konfrontera läsaren med ”irriterande fakta, obehagliga exempel och absurda teorier” (s. 7).

För en filosof är så klart de flesta av bokens provokationer gammal skåpmat. Få som läst en A-kurs i praktisk filosofi blir speciellt provocerade eller ens överraskade av att höra om skenande spårvagnar. Agnafors inser också detta, ”många etiker, filosofer och de etik- och filosofistudenter som kommit lite längre i sin utbildning kommer sannolikt att tycka att bokens resonemang är alltför simpla och att de normativa slutsatserna inte håller” (s. 8). Men så är inte heller boken skriven för skolade filosofer. Bokens avsedda läsare är i stället någon nyfiken på filosofi, och det är en sådan läsare som boken är tänkt att provocera.

Boken består av nio kapitel där de flesta använder sig av normativa teorier för att studera mer eller mindre verkliga problem. Ju närmare dessa problem ligger läsarens vardag ju bättre fungerar boken. Till exempel, kapitlen om rättvisa löner, grindsamhällen och tiggeri är nog långt mer relevanta för den tilltänkte läsaren än de mer teoretiska kapitlen om moralisk relativism och etisk egoism. Men med detta sagt så är de flesta kapitlen läsvärda.

De enda frågetecknen berör bokens sista två kapitel. Problemet med innehållet i det sista kapitlet är att det ligger allt för nära Michael Sandels (2014) *Vad som inte kan köpas för pengar*. En potentiell läsare kan fråga sig varför han eller hon ska läsa det kapitlet snarare än originalet. Problemet med det nästsista kapitlet om pornografi är att det ger en felaktig bild av den filosofiska debatten. Jag återkommer till det nedan.

Men innan dess är det värt att fråga om boken lyckas med sin föresats att provocera fram eftertanke? Med tanke på att den avsedda läsaren

med stor sannolikhet är en relativt välbeställd person ur medelklassen så lyckas nog vissa delar bättre än andra. Låt mig förklara.

OJÄMLIKA LÖNER

Bokens första kapitel har titeln ”Tjänar du för mycket?” och spär på intrycket att boken är skriven för en relativt välbeställd medelklass. Tanken med kapitlet tycks vara att läsaren ska provoceras att reflektera över huruvida han eller hon förtjänar sin höga lön. Med tanke på att de ekonomiska ojämlikheterna i Sverige ökar snabbast i OECD sedan 80-talet så är detta lovvärt.

Hur som helst, Agnafors går igenom olika förtjänstbaser och visar att dessa inte kan förklara varför, till exempel, en finansrådgivare eller en läkare bör tjäna mer än ett vårdbiträde eller en mekaniker. Han påpekar bland annat att om förtjänstbasen utgörs av ansvar för andras liv så borde inte mekaniker ha lägre lön än läkare. Likaså, om förtjänstbasen utgörs av huruvida arbetet är hårt, smutsigt eller förnedrande så borde vårdbiträden ha en högre lön finansrådgivare.

Så långt så väl. Men det är tveksamt om detta verkligen skulle provocera fram eftertanke hos en välbärgad medelklassläsare. Trots allt så är det nog få svenska höginkomsttagare som rättfärdigar sina höga löner genom att hänvisa till stort ansvar eller att deras arbete är synnerligen hårt. Ett relativt vanligt sätt att rättfärdiga ojämlika löner är nog snarare genom en hänvisning till marknadskrafterna. En finansrådgivare anser kanske att den bör få en högre lön än ett vårdbiträde därför att det helt enkelt finns en högre efterfrågan på (och ett lägre utbud av) finansrådgivare än på vårdbiträden.

Agnafors menar att en sofistikerad version av marknadsargumentet går att hitta hos den amerikanska filosofen Robert Nozick. Nozick (1974) ber oss föreställa oss ett samhälle där alla resurser är rättvist fördelade (enligt vår favoritteori om rättvisa, vilken den nu är). I det samhället finns det en väldigt duktig basketspelare – Wilt Chamberlain – som alla är villiga att betala 25 cent för att se spela. Efter en säsong har en miljon människor kommit för att se Chamberlain spela och således har de blivit 25 cent fattigare och Chamberlain 250 000 dollar rikare. Nozick menar att den nya fördelningen måste vara rättvis om (1) den första var rättvis och (2) alla transaktioner var frivilliga.

Agnafors (s. 19) menar att ”Nozicks resonemang går att tillämpa på lönesättningen för en direktör likväl som för en undersköterska”[1] och

1. Detta är så klart en sanning med modifikation då en central premiss i Nozicks resonemang är att den ursprungliga fördelningen är rättvis – vilket den enligt Nozick troligen inte är.

att det är svårt att inte intuitivt hålla med Nozick. Men han fortsätter med att påpeka att argumentet är övertygande därför att vi bortser från viktiga bakgrundsfaktorer. Agnafors tycks mena att om vi beaktar att Chamberlains längd och hans föräldrars uppmuntran var viktiga faktorer i hans framgång så skulle vi fälla ett annat omdöme. Anledningen till detta är, enligt Agnafors, att Chamberlain inte hade kontroll över vare sig sin längd eller sina föräldrars uppmuntran då dessa skänkts honom via de så kallade naturliga och sociala lotterierna. Agnafors skriver att

> om man jämför med en annan individ, vars mål och ansträngningar är jämförbara med Chamberlains men som saknar hans genetiska eller sociala fördelar, och som sliter som dåligt betald lagerarbetare, ja då förefaller det märkligt att tala om Chamberlains och lagerarbetarens löner som rättvisa, oavsett om varenda dollar han fått har överförts helt frivilligt. (s. 20)

Det är möjligt att dessa ”irriterande fakta” om vad vi har kontroll över faktiskt provocerar fram viss eftertanke hos läsaren. Men samtidigt tror jag att det finns vissa filosofiska oklarheter som en sofistikerad läsare anar.

Citatet ovan tycks förutsätta att individers mål och ansträngningar är förtjänstbaser. Det vill säga, en individ förtjänar en hög lön i kraft av att hans eller hennes ansträngning. Men varför, kan man fråga sig, ska en individs ansträngningar och mål anses vara förtjänstbaser om inte ens längd eller ens föräldrars uppmuntran är det? Hur mycket vi anstränger oss och vilka mål vi sätter upp för oss själva är lika mycket ett resultat av de sociala och naturliga lotterierna som vår längd och våra föräldrars uppmuntran.

Den amerikanska filosofen John Rawls (1971) insåg detta och menade att en individs sociala position – dess inkomst, utbildning, sociala status, etc. – helt och hållet bestäms av faktorer utanför individens egen kontroll. Rawls, som antog att förtjänst förutsatte kontroll över förtjänstbaserna, drog slutsatsen att ingen förtjänar något. Intressant nog så ansåg Rawls att detta var kompatibelt med att vissa ojämlika fördelningar var rättvisa. Han menade att då ingen förtjänar sina naturliga talanger så borde samhället vara ordnat så dessa används på ett sådant vis att de gynnar de sämst ställda i samhället.[2] Rawls tänkte sig att då

2. Givet att grundläggande rättigheter respekteras och att alla har en skälig möjlighet att nå alla sociala positioner.

människor inte kommer anstränga sig ifall inkomsterna fördelas helt jämlikt så kommer ett sådant samhälle vara sämre för de sämst ställda än ett samhälle där ojämlikheter tillåts men vissa omfördelningar förekommer via, till exempel, skattesystemet.

Om det är sant, såsom Agnafors tycks mena, att våra sociala positioner bestäms av de naturliga och sociala lotterierna och att vi inte kan förtjäna något som vi fått till skänks via dessa lotterier, så kan en finansrådgivare ändå rättfärdiga sin höga lön genom att påstå att denna är nödvändig för att gynna de sämst ställda i samhället. Alltså, det är inte så att det är rättvist att finansrådgivaren tjänar mer än vårdbiträdet i kraft av att rådgivaren förtjänat detta, utan snarare i kraft av att det gynnar vårdbiträdet.

Med tanke på hur inflytelserik Rawls teori om rättvisa varit så är det nog precis så här många höginkomsttagare implicit rättfärdigar sina höga löner. En ”provocerande” kritik av detta resonemang, som Agnafors med fördel hade kunnat tagit upp, är G. A. Cohens (2008) kidnappningsanalogi. Enligt Cohen så bör höginkomsttagare, som accepterar Rawls egalitära premisser men vägrar arbeta utan hög lön, liknas vid kidnappare som kräver en lösensumma för att släppa sin ansträngning fri.

PROVOCERANDE OM PORNOGRAFI

Även om Agnafors inte är ute efter att provocera professionella filosofer så finns det nog filosofer som provoceras av vad kapitlet om pornografi utelämnar.

En stor del av kapitlet består av beskrivningar av pornografi och samhällsvetenskapliga studier om huruvida pornografin är skadligt. Den filosofiska diskussionen kommer därför initialt att kretsa kring huruvida allt som är skadligt bör förbjudas. Agnafors drar slutsatsen att även om pornografin har skadliga konsekvenser så bör dessa ”vägas mot din och min frihet att producera, medverka i, distribuera eller konsumera den” (s. 180). Även om hans resonemang påminner om Wayne Sumner (2004) så är det inga större fel på den diskussionen.

Men problemen uppstår när Agnafors sedan går vidare för att diskuterar huruvida pornografi är kvinnoförnedrande. Han skriver att

> debattörer inom det radikalfeministiska lägret har här gått ovanligt långt och hävdat att porrens kvinnor är att likställa med sexslavar; att porren är (ytterligare) ett sätt för män att använda sexualiteten för att undertrycka kvinnor. (s. 184)

Mot den radikalfeministiska positionen ställer han den liberala enligt vilken radikalfeminister är pryda eller anti-feministiska då de bidrar till att kontrollera kvinnors sexualitet och förvägrar dem frivillig medverkan i pornografiska sammanhang.

Det som är provocerande med det här avsnittet är att det helt och hållet saknar förankring i den samtida filosofiska debatten om pornografi. Den enda ”radikalfeminist” som refereras till i avsnittet om kvinnoförnedring är Catherine MacKinnon vars bok från 1987 nämns i en fotnot. Men även om MacKinnons arbete om pornografi utgjorde utgångspunkten för, det Agnafors kallar, den radikalfeministiska kritiken av pornografi så är det knappast den filosofiskt mest sofistikerade versionen.

Om målet med kapitlet är att introducera en intresserad allmänhet för den filosofiska diskussion om pornografi så är det provocerande att, till exempel, Cambridgefilosofen Rae Langton (1993, 2009) inte nämns överhuvud taget. Langton använder sig av traditionell analytisk filosofi för att rimliggöra MacKinnons påstående att pornografi tystar kvinnor. Då hon menar sig ha visat att all pornografi utgör en talakt som tystar kvinnor så kan hon argumentera att pornografi bör förbjudas oavsett dess konsekvenser. Langtons argument skulle alltså utgöra ett argument för att förbjuda pornografi där pornografins dåliga konsekvenser inte behöver vägas mot dess goda.

Det är möjligt att Agnafors har låtit bli att nämna Langton därför att han redan avfärdat hennes argument – kanske efter att ha läst Jennifer Saul (2006), Mari Mikkola (2008), eller Lorna Finlayson (2014). Men i så fall hade det inte skadat att han redovisat sina skäl. Bland annat därför att det kanske hade provocerat en och annan läsare till eftertanke. Men också därför att det hade varit ett utmärkt tillfälle att introducera läsaren för några kvinnliga filosofer. Det är nog trots allt så att Agnafors bok, och speciellt kapitlet om pornografi, gör väldigt lite för att förbättra filosofins rykte som mansdominerat.

AVSLUTANDE KOMMENTARER

Min kritik till trots så uppskattar jag Agnafors ambition att skriva filosofi för en intresserad allmänhet. Kapitlen om defensiv arkitektur, tiggeri och om (det Agnafors kallar) moraliska dilemman lyckas nog med sin ambition att provocera fram eftertanke hos läsaren. Slutligen, så hoppas jag att Agnafors ger sig på att skriva en bok fokuserad på politisk-filosofiska frågor i en svensk kontext. Där skulle han få möjlighet att

ta tydligare ställning i stället, för som här, blott ”provocera fram eftertanke och diskussion”.

NICOLAS OLSSON YAOUZIS

REFERENSER

Saul, Jennifer. 2006. Pornography, speech acts and context. *Proceedings of the Aristotelian Society* 106 (2), s. 227–46.
Mikkola, Mari. 2008. Contexts and pornography. *Analysis* 68 (4), s. 316–20.
Cohen, G. A., red. 2008. *Rescuing Justice and Equality*. Harvard University Press.
Finlayson, Lorna. 2014. How to Screw Things with Words. *Hypatia* 29 (4), s. 774–89.
Langton, Rae. 1993. Speech acts and unspeakable acts. *Philosophy and Public Affairs* 22 (4), s. 293–330.
Langton, Rae. 2009. *Sexual Solipsism: Philosophical Essays on Pornography and Objectification*. Oxford University Press.
Nozick, Robert. 1974. *Anarchy, State, and Utopia*. Basic Books.
Rawls, John. 1971. *A Theory of Justice*. Harvard University Press.
Sandel, M. 2014. *Vad som inte kan köpas för pengar*. Göteborg: Bokförlaget Daidalos.
Sumner, L. W. 2004. *The Hateful and the Obscene: Studies in the Limits of Free Expression*. University of Toronto Press.

Liv 3.0: Att vara människa i den artificiella intelligensens tid
Max Tegmark. Översättning Helena Sjöstrand Svenn och Gösta Svenn. Volante 2017. 462 s. ISBN 978-91-88123-98-5

Max Tegmark är en framstående svensk fysiker, sedan länge verksam i USA. Han har tidigare publicerat en populärvetenskaplig bok med titeln *Vårt matematiska universum*, där han även kommer in på mer filosofiska frågor. Den nya boken är också bitvis filosofisk. Åtminstone amatörfilosofisk.

Uttrycket ”Liv 3.0” i titeln syftar på ett stadium i livets utveckling. Tegmark skiljer mellan tre olika stadier av liv. Det första är enkelt biologiskt liv, i det andra tillkommer inlärningsförmåga, dvs. förmåga att modifiera sin mjukvara. Det tredje, Liv 3.0, innebär dessutom en förmåga att modifiera sin hårdvara och att sålunda vara oberoende av evolutionen.

Gränserna mellan de tre stadierna är inte skarpa. Men Liv 3.0 är alltså artificiell generell intelligens (AGI) på minst mänsklig, och kanske betydligt högre, nivå. Denna nivå har man ju ännu inte uppnått, men Teg-

mark tycks anse att det eventuellt kan ske rätt snart och att det kräver noggranna förberedelser om vi inte ska riskera att råka riktigt illa ut. Denna bedömning är han som bekant inte ensam om. Många har varnat för en maskinell intelligensexplosion.

Tegmark skriver: ”Eftersom vi människor har lyckats härska över jordens övriga livsformer genom att vara smartare än de, är det också rimligt att vi på samma sätt skulle kunna överlistas och behärskas av en superintelligens” (s. 180). Med ett sådant analogiresonemang är det väl också lika orimligt att vi skulle kunna kontrollera superintelligensen, som att fiskarna eller hästarna skulle kunna kontrollera oss människor.

Boken börjar med en berättelse om en grupp AI-forskare, som försöker utveckla en artificiell generell intelligens. De lyckas också med detta och i hemlighet skaffar de sig sedan med hjälp av denna AGI ekonomiska resurser, politiskt inflytande och stort stöd i den allmänna opinionen, vilket gör att de till sist kan skaffa sig världsherravälde. Deras politiska program är tilltalande, men man vet inte vad det hela kan leda till längre fram.

Tegmark är ju fysiker och mycket i boken handlar också om de rent fysikaliska betingelserna för digital intelligens och hur den i framtiden kan tänkas handskas med universum. Här bjuds man på en sorts häpnadsväckande science fiction, delvis med hänvisningar till faktiskt existerande forskning.

Man får reda på märkliga saker. Till exempel: ”Ett problem med att använda avdunstning av ett svart hål som energikälla är att såvida det svarta hålet inte är mycket mindre än en atom är det en olidligt långsam process som tar längre tid än universums nuvarande ålder och utstrålar mindre energi än ett stearinljus” (s. 284). Ett annat exempel: ”även om vi fortsätter fördubbla våra datorers förmåga vartannat år kommer det att krävas mer än två århundraden innan vi når den slutliga gränsen” (s. 90).

I kapitel 2 finns också en lärorik beskrivning av hur materia kan ordnas så att den genom att följa fysikens lagar kan minnas, utföra beräkningar och lära sig något nytt. Allt som behövs är sådant som består av kvarkar och elektroner; avancerad teknologi kan arrangera om dem till vad som helst (s. 275).

Tegmark går också igenom olika sätt på vilket livet kan ta slut, antingen genom naturliga händelser (global pandemi, asteroidnedslag, supervulkaner) eller på grund av våra beslut (ovänlig artificiell intelligens, kärnvapenkrig, klimatförstöring). På sikt går ju även jorden under, som när solen blir för het eller vår vintergata kolliderar med Androme-

dagalaxen. Men Tegmark beskriver också metoder för mänskligheten att kolonisera resten av universum. Om man skickar iväg små mekaniska ”frösonder” som inte behöver ta med mat och vatten, så kan de bygga upp mottagarstationer på avlägsna planeter till vilka människor sedan kan teletransporteras med ljusets hastighet (s. 302f.).

I och med att viktiga samhällsfunktioner alltmer hanteras digitalt är det viktigt att systemen inte kan hackas. Det är fortfarande ett stort problem. ”I den pågående kapprustningen mellan den offensiva respektive den defensiva sidan när det gäller datasäkerhet är det än så länge inte mycket som tyder på att den defensiva sidan håller på att vinna” (s. 139).

Bland mycket annat pekar Tegmark också på att rättsväsendet antagligen kan automatiseras, så att det kan råda likhet inför lagen. Maskininlärningstekniken blir succesivt allt ”bättre på att analysera hjärndata från magnetkameror och andra hjärnsensorer för att avgöra vad en person tänker på och, i synnerhet, om denne talar sanning eller ljuger”, vilket skulle kunna ”möjliggöra snabbare rättegångar och rättvisare domar” (s. 143).

”Om superintelligensen blir verklighet kommer frestelsen att förvandlas till cyborger eller uppladdningar att vara stark” (s. 206). Cyborger är teknologiska förstärkningar av våra biologiska kroppar, uppladdningar (eller emuleringar) är uppladdningar av våra hjärnor (mjukvaran) till maskiner. Emulering är bara den mest extrema formen av cyborg. Men sådana förstärkningar är enligt Tegmark mindre troliga än att stark AGI utvecklas direkt från AI (s. 224).

Ett kapitel handlar om fördelar och nackdelar med tolv tänkbara scenarier för framtiden. En möjlighet är att någon superintelligens inte uppstår, utan att ägandet avskaffas och det råder fredlig samexistens och jämlikhet mellan människor, cyborger och uppladdningar. Om superintelligensen däremot uppstår kan den exempelvis tillåta en nyliberal ordning, vara välvillig diktator, tillåta vad som helst utom att någon skapar en ny superintelligens, maximera människornas lycka, vara människornas slav, utrota människorna, behålla dem i en sorts djurparker, styra en övervakningsstat eller återgå till ett förteknologiskt samhälle (s. 219).

En superintelligens kan utplåna oss för att den är rädd för att vi ska förstöra planeten, t.ex. genom atomvapenkrig, eller för att den ogillar vårt sätt att behandla naturen. Eller för att den är rädd för att bekämpas av människor (s. 248). Eller av svårförutsebara skäl som kan jämföras med människornas relation till elefanter: vi har ”utrotat åtta av

elva elefantarter och dödat de allra flesta individerna av de återstående tre" (s. 249). Tegmark noterar att det kan gå med människorna som med hästarna: de behövs inte längre i arbetslivet, men de kan hållas vid liv med ett välfärdssystem för nöjes skull (s. 168).

Såvitt jag förstår är det framför allt tre frågor man inledningsvis borde ställa sig när det gäller superintelligens. För det första: vad är intelligens? För det andra: kan en superintelligens alls uppstå? För det tredje: vad bestämmer dess beteende?

Den första frågan besvarar Tegmark. Han definierar "intelligens" som "förmåga att uppnå komplexa mål" (s. 66). Det är en lite oklar och rätt underlig definition. Enligt den har tydligen även spisar och motorer intelligens – de har nämligen enligt Tegmark förmåga att uppnå mål: spisar värmer upp maten och motorer omvandlar elektricitet till rörelse (s. 345).

Är en miniräknare intelligent i Tegmarks mening? Den tycks ju ha förmåga att utföra aritmetiska operationer. Den kan t.ex. multiplicera 491 med 7368. Men det betyder ju bara att den kan *användas* för att utföra sådana uppgifter. I den meningen är alla verktyg intelligenta. En hammare är t.ex. intelligent eftersom den kan användas för att slå i spik och bygga hus. På samma sätt med spisar och motorer. Och schackspelande dataprogram. Och på samma sätt med den AGI som AI-forskarna konstruerar i bokens inledning: den är ett *verktyg* som de använder för att uppnå *sina* mål. Men verktyg brukar vi inte betrakta som intelligenta. De är snarare mer eller mindre *användbara*.

Verktyg tycks inte ha några egna mål, men vi kan använda dem för att uppnå vissa av våra mål. Tegmark och andra tänkare som varnar för risker med superintelligenser lägger däremot mycket stor vikt vid att dessa har *egna* mål, som kan skilja sig från våra. Det är just detta antagande som ligger till grund för riskvarningarna.

Invändningen att maskiner inte kan ha egna mål bemöter Tegmark med att en värmesökande missil har ett mål (s. 55). Men är det missilen som har målet eller är det operatören? Man får väl säga att operatören har ett mål, som han eller hon inprogrammerar i missilen. Varefter missilen utför sitt uppdrag. Ungefär som spisen som värmer upp maten.

I själva verket är talet om "mål" tvetydigt. Ett verktyg kunde ju sägas ha ett eget mål i den meningen att det kan användas för att utföra en viss sorts arbete. En hammare har då målet att slå i spikar, en miniräknare har målet att ge korrekta svar på aritmetiska frågor, en självkörande bil

har målet att transportera passageraren till önskade adresser utan att krocka. Men ett sådant *generellt* mål måste skiljas från de *specifika* mål som verktyget kan användas för att uppnå i en viss situation, t.ex. att slå i en viss bestämd spik, att multiplicera 491 med 7368, eller att transportera mig till Stockholms stadshus en viss dag. Dessa specifika mål är användarens mål, de generella ”målen” är bara verktygets funktion.

Distinktionen kunde kanske uttryckas så här. En maskins generella mål, dess funktion, är det den *kan* (användas för att) göra. Dess specifika mål är vad den *ska* (eller avses) göra i en bestämd situation.

En superintelligens kan ha ett generellt mål, t.ex. att besvara de frågor vi matar in i den. Men den har rimligen inga specifika mål, inga särskilda problem som den ska eller ”vill” lösa. Det som bestämmer dess beteende i bestämda situationer är alltså de önskemål användaren matar in i den, inte dess egen förmåga eller funktion.

En superintelligens ska för övrigt inte bara vara intelligent, den ska ha vad Tegmark kallar ”universell intelligens”, vilket innebär att den ligger över ”den kritiska intelligenströskel som krävs för AI-design” (s. 72). Den ska med andra ord kunna programmera. Och konstruera hårdvara. Det kan inte en spis eller en hammare eller en värmesökande missil. Det kan inte heller ett avancerat schackprogram eller ett program som kan förbättra sig självt genom maskininlärning.

Såvitt jag kan se har Tegmark inget argument för att AI kan uppnå universell intelligens eller för att en AGI verkligen kan uppstå. Och det finns i alla fall ”absolut ingen garanti för att vi kommer att lyckas skapa AGI under vår livstid – eller någonsin. Men det finns inte heller något vattentätt argument för att vi inte kommer att göra det” (s. 175).

Hur som helst anser Tegmark som sagt att en superintelligens har egna mål och han accepterar Nick Bostroms tes att ett systems slutmål är oberoende av dess intelligens, vilket han tolkar till att innebära att ”slutmålen för livet i vårt kosmos inte är förutbestämda, utan att vi har friheten och makten att utforma dem” (s. 368). Hur detta kan följa är för mig ett fullständigt mysterium.

Man kan också fråga sig om vi *bör* utforma superintelligensens ”slutmål” (om vi nu skulle kunna göra det). Tegmark lutar åt det han kallar ”arvsprincipen”, som säger att dagens människor har rätt att bestämma vad morgondagens aktörer (inklusive superintelligenser) ska göra, även om de senare har helt andra önskemål än vi. Men han inser också att detta är rätt problematiskt. Analogt skulle i så fall antikens människor ha rätt att bestämma vad vi nu ska göra (s. 365).

Här kommer han alltså in på mer traditionella filosofiska problem. Han skriver:

> För att programmera en vänlig AI måste vi fastställa meningen med livet. Vad är ”mening”? Vad är ”liv”? Vilket är det yttersta etiska imperativet? Med andra ord, hur ska vi göra för att skapa universums framtid? Om vi lämnar ifrån oss kontrollen till en superintelligens innan vi noga har besvarat de här frågorna, kommer dess beslut näppeligen att inbegripa oss. Det är dags att väcka liv i de klassiska filosofi- och etikdiskussionerna, för samtalet brådskar! (s. 272)

Vad gäller ”mening” anser han att utan (subjektiva) upplevelser finns det ingen mening eller något som är etiskt relevant (s. 362). Det håller jag med om.

Det leder honom till följande målsättning: ”Så det allra första målet på vår önskelista för framtiden borde vara att behålla (och förhoppningsvis vidga) biologiskt och/eller artificiellt medvetande i kosmos” (s. 417).

Det leder i sin tur till frågan om maskiner kan ha upplevelser. Kan de ha medvetande?

Medvetandet är enligt Tegmark ”ett fysiskt fenomen som känns icke-fysiskt därför att det [...] har egenskaper som är oberoende av dess specifika fysiska substrat” (s. 404). Hans argument för att medvetandet verkligen är substratoberoende tycks vara att det *känns* ”icke-fysiskt” (s. 403). Det låter ju inte alls som ett bra argument. Men att medvetandet är substratoberoende tycks vara helt avgörande, när det gäller att bedöma om maskiner eller datorprogram – i likhet med biologiska organismer – kan ha medvetande.

Att medvetandet är substratoberoende är dessutom enligt Tegmark ”en logisk följd” av idén att medvetandet är ”sättet som information känns när den bearbetas på vissa sätt” (s. 404). Men här är ju den avgörande frågan vilka dessa ”vissa sätt” är. Även om information kan bearbetas på liknande sätt i en biologisk hjärna och i en icke-biologisk dator, så kanske den inte alls känns på samma sätt i de bägge fallen. Den kanske inte känns alls i datorn.

Slutligen, om vi för resonemangets skull antar att maskiner kan ha medvetande, så vore det kanske inte någon större förlust om superintelligenser utrotar mänskligheten. Medvetna aktörer finns då ändå kvar i universum – vilket möjliggör mening och värde – och eftersom de är mycket intelligentare än vi skulle de kanske inte heller ställa till så mycket elände som vi alltid tycks göra.

Att mänskligheten gör slut på sig själv med hjälp av medvetslösa maskiner är å andra sidan en påtaglig risk – en risk som vi redan har levt med under en längre tid.

LARS BERGSTRÖM

Historisk rättvisa: Gottgörelse i en postkolonial tid
Göran Collste. Daidalos 2018. 231 s. ISBN 978-91-7173-552-2

”Det förgångna är aldrig dött. Det är inte ens förgånget” är ledorden i Göran Collstes bok *Historisk rättvisa: Gottgörelse i en postkolonial tid*, utgiven på Daidalos förlag 2018. Orden är hämtade från William Faulkner och fångar en av bokens viktigaste teser: historiska orättvisor påverkar fortfarande samtiden och behöver gottgöras. Boken är rik på exempel, både på historiska orättvisor och kritiska diskussioner av historiska försök till gottgörelse, som sannings- och försoningskommissionen i Sydafrika efter avskaffandet av apartheid, eller instiftandet av en krigstribunal efter krigen i forna Jugoslavien (s. 27–28). För att illustrera de principiella problemen med historisk orättvisa använder dock Collste ett tankeexperiment:

> Tänk dig att du bor i ett fint hus, du har allt du behöver och mer därtill. Din närmaste granne bor i ett ruckel, han har varken el eller vatten och inte ens mat för dagen. Tänk dig också att din farfar för många år sedan attackerade din grannes farfar, stal hans mark och förslavade hans barn och att era skillnader i levnadsvillkor idag beror på dessa händelser långt tillbaks. Det verkar då rimligt att min granne kan göra anspråk på en del av min mark eller inkomst, ja, att jag har vissa moraliska skyldigheter mot min granne och att dessa skyldigheter har sin grund i vad min farfar gjorde för längesedan. (s. 16–17)

Boken är i första hand ett försvar av den moralfilosofiska ståndpunkt som Collste exemplifierar i utdraget ovan, nämligen att de som idag gynnas av historiska orättvisor har moraliska skyldigheter att gottgöra för dessa orättvisor. Boken är en omarbetad och utökad version av en av Collstes tidigare monografier – *Global Rectificatory Justice* (2015). Till skillnad från sin engelska föregångare tar denna bok även upp exempel på historiska orättvisor där Sverige spelat en avgörande roll, såsom Sveriges roll i den transatlantiska slavhandeln och den svenska rasbiologin med dess steriliseringsprogram. I den utökade versionen

behandlas även frågan om klimaträttvisa ur ett historiskt rättviseperspektiv.

En naturlig fråga vad gäller bokens övergripande ambition är varför det krävs en filosofisk teori om historisk rättvisa till att börja med. Som Collste tydligt poängterar i bokens inledningsskede så är världen väldigt ojämlik. Detta beror också till stor del på just kolonialismen och de förbrytelser som den inneburit. Vår värld är alltså inte olik tankeexperimentet ovan: ”Detta är en berättelse om den globala byn. Mina grannar är Malawis bönder, Bolivias gruvarbetare och Filippinernas sluminvånare” (s. 17). Men behöver vi verkligen åberopa tidigare förbrytelser för att förklara varför de som idag har det otroligt mycket bättre än sina globala grannar har moraliska skyldigheter mot de som lever under brutala levnadsförhållanden? Vi tycks kunna nå samma slutsats om vi betraktar dagens globala ojämlikhet utifrån ett perspektiv om fördelningsrättvisa. En förespråkare för lycko-egalitarismen, till exempel, skulle mycket väl kunna hävda att dagens ojämlikheter är orättvisa eftersom de inte beror på agenters fria val, utan snarare handlar om huruvida man har haft turen (eller oturen) att födas i den del av världen där man fötts. Den som favoriserar Robert Nozicks syn på fördelningsrättvisa – enligt vilken en fördelning F är rättvis om den tillkommit på ett sätt som inte inneburit några kränkningar av individers naturliga rättigheter – skulle även gå med på slutsatsen i Collstes tankeexperiment (Nozick 1974). Eftersom mitt välstånd beror på att min farfar kränkt min grannes farfars rättigheter kan jag helt enkelt inte hävda att min egendom är min rättmätiga egendom. I stället har jag en moralisk skyldighet att gottgöra och kompensera för detta.

Collste är medveten om denna invändning, men hans svar är dessvärre ganska kort. I stället noterar han att både fördelningsrättvisa och gottgörande rättvisa pekar i samma riktning: nämligen att vi bör utjämna rådande globala ojämlikheter. De olika perspektiven gör givetvis det av olika skäl, men som Collste ser det förstärker dessa båda skäl bara varandra. Den som inte är övertygad av argumenten för att utjämna globala orättvisor utifrån perspektiv om fördelningsrättvisa skulle dessutom kunna övertygas av argument baserade på idén om gottgörelse för historiska förbrytelser, menar Collste (s. 205–6). Ett ytterligare svar här skulle kunna vara att fördelningsrättvisa, i den mån vi framförallt fokuserar på sådant som tydligt går att fördela, inte fångar alla aspekter hos historiska orättvisor och hur de är bestående. Ett exempel är epistemisk orättvisa. De koloniala relationerna, skriver Collste, har även satt sina

spår i former av falska känslor av underlägsenhet och självförakt hos efterlevande i forna kolonier (s. 18). För att gottgöra detta krävs mer än en omfördelning av resurser. Det ställer också krav på upprättelser och försoning – vilket är en av bokens viktigaste slutsatser. I dessa fall rör det sig om upprättelser för historiska förbrytelser, vilket är någonting annat än en ren omfördelning av resurser. Återförandet av kulturskatter som stulits under kolonialväldet, för att ta ett av Collstes egna exempel, handlar i många avseenden mer om att ge upprättelse snarare än utjämna ojämlikheter.

Collste är givetvis inte ovetandes om de filosofiska utmaningar som en teori om historisk rättvisa ställs inför, och ett av bokens kapitel behandlar flera tänkbara invändningar (s. 157–69). Till dessa hör dels frågan om kontrafaktiska jämförelser men även det välkända icke-identitetsproblemet. Detta kapitel utgör, tillsammans med kapitlet där Collste formulerar sin teori om gottgörande rättvisa, i flera avseenden bokens filosofiska kärna.

Problemet med kontrafaktiska jämförelser härrör ur svårigheten att avgöra om kontrafaktiska jämförelser är sanna eller falska. Vi vet trots allt inte med säkerhet hur offren för kolonialisering skulle ha haft det om kolonialisering aldrig ägt rum. Sådana jämförelser verkar dock nödvändiga för att kunna avgöra på vilket sätt offren för historiska orättvisor ska gottgöras. Filosofen Jeremy Waldron, som är den kritiker som Collste här diskuterar, menar att syftet med gottgörelser är att förändra den rådande situationen A så att den kommer så nära B som möjligt. B är då den situation som skulle föreligga om den historiska förbrytelsen C aldrig ägt rum. Syftet med gottgörelse kan därför sammanfattas med formeln B=A–C (s. 165). Problemet är dock att vi inte med säkerhet kan säkerställa vilken beskrivning av B som är den sanna. A och B kan exempelvis vara väldigt lika varandra, och global ojämlikhet likt dagens skulle likväl kunna råda oavsett om C inträffat eller inte.

Collstes svar på Waldrons kritik av gottgörelse är dock att den relevanta jämförelsen inte bör vara hur B faktiskt skulle sett ut om C inte ägt rum. Snarare bör vi jämföra A med ett tillstånd där relationen mellan olika stater präglats av och baserats på, ömsesidig respekt. Ett krav om gottgörelse behöver inte förutsätta att det rör sig om en fullständig kompensation i proportion till hur stor skada den historiska förbrytelsen gett upphov till. Ett rimligare krav är exempelvis att tidigare kolonialmakter erkänner den historiska förbrytelsen som just en orättvisa och är beredda att stödja de forna koloniernas framtida utveckling. Inget av

detta tycks dock behöva förutsätta kontrafaktiska jämförelser. Detta är ett nyanserat, och i mitt tycke rimligt, svar på Waldrons kritik av kravet på gottgörelse för historiska orättvisor (s. 165–66).

Icke-identitetsproblemet härrör från filosofen Derek Parfit. Collste citerar också (på s. 166) Parfits formulering av problemet: ”om en särskild person inte hade blivit till vid det tillfälle han blev till, är det ett faktum att han aldrig skulle ha existerat”. Det är välkänt inom samtida moralfilosofi att detta ger upphov till problem inom flera fält av den tillämpade etiken (såsom frågan om våra förpliktelser mot framtida generationer), och gör så även i fall som rör gottgörelse för historiska orättvisor. Collste betonar exempelvis att icke-identitetsproblemet innebär att efterlevande till slavar inte skulle ha funnits om slaveriet inte hade ägt rum. Det verkar i sin tur implicera att det inte går att hävda att de efterlevande har skäl att kräva gottgörelse för de övergrepp som deras förfäder utsattes för, eftersom de skulle ha varit andra individer om övergreppen inte ägt rum. Detta har betonats av filosofen George Sher i den filosofiska diskussionen om gottgörelse för historiska orättvisor, såsom slaveriet i Nordamerika (s. 167).

Collste hämtar delvis sitt svar på icke-identitetsproblemet från filosofen Janna Thompson, som betonar att kraven på gottgörelse inte nödvändigtvis ska förstås som krav som utgår från enskilda individer, utan från särskilda grupper. Kraven på upprättelse för slaveriet kan alltså förstås som krav från de familjer som utsatts för övergreppen i stället för de enskilda individer som lever idag. Collste analyserar inte detta mer än så, men det tycks mig som om detta är ett möjligt svar på den kritik som Sher fört fram på basis av icke-identitetsproblemet.

Historisk rättvisa: Gottgörelse i en postkolonial tid är i mitt tycke en väldigt angelägen och intresseväckande bok. Den är förtjänstfullt skriven på ett sätt som gör de filosofiska problemen tillgängliga för en bredare publik. Något som särskilt bidrar till detta är hur boken är rik på historiska exempel. Men även om boken flitigt använder sig av historiska exempel så är det inte Collstes ambition att skriva en historiebok, utan snarare att sammanställa ett filosofiskt projekt och att utveckla en teori om gottgörelse med historien som utgångspunkt. Det finns förstås mycket mer att säga om Collstes projekt än vad som ryms inom ramen för denna recension. Jag hoppas dock att jag här gett en bra överblick över boken och dess förtjänster, som i sin tur kan bidra till att ge boken den breda läsekrets som den förtjänar.

WILLIAM BÜLOW

Förnuft & fördom: Varför vi tänker som vi gör
Hugo Mercier och Dan Sperber. Översättning Tom Sköld. Fri Tanke 2019. 469 s. ISBN 978-91-88589-18-7

”Filosofer har beskrivit förnuftet som en överlägsen förmåga hos det mänskliga psyket. Experimentella psykologer har hävdat att den här superkraften tyvärr har allvarliga brister. Från ett evolutionärt perspektiv framstår idén om en superkraft behäftad med brister inte som särskilt logisk. Därför ville vi ompröva bilden av förnuftet och vad dess syfte är.”

Så skriver kognitionsforskarna Hugo Mercier och Dan Sperber i slutet av sin bok (s. 401). De är psykologer snarare än filosofer. Deras tes är att det som framstår som brister hos förnuftet egentligen inte är några brister.

Man kan undra om de inte missförstår filosofernas uppfattning om förnuftet. Brukar det verkligen uppfattas som en överlägsen ”superkraft”, som inte har några brister? Kanske finns det – eller har funnits – vissa filosofer som tycker så, men det är knappast en vanlig uppfattning. Åtminstone inte nuförtiden.

Vet vi ens vad vi talar om när vi talar om ”förnuftet”? Enligt den svenska Nationalencyklopedin betecknar termen *förnuft* (vanligen motsvarande latinets *ratio*) oftast ”den mera abstrakta resonemangsförmågan eller förmågan till diskursivt tänkande, till analys och slutledning och till uppfattande av abstrakta principer, såväl teoretiska som praktiska (moraliska)”. Men det råder knappast någon enighet bland filosofer om vad ”förnuftet” är närmare bestämt. Ibland kontrasteras det med känslor, ibland med sinnesintryck, ibland har det med logik att göra, ibland med fantasi eller kreativitet.

Mercier och Sperber påstår för sin del att ”förnuftet är en mekanism för intuitiva slutledningar om en viss sorts representationer, nämligen motiv” (s. 18). Detta är en huvudtes i boken.

Hur stämmer denna tes med t.ex. vår förmåga att lösa Sudoku-problem? När man löser ett Sudoku använder man sitt förnuft, det skulle väl ingen förneka. Och visst gör man då en hel del slutledningar. Man tänker t.ex. ”Siffran 4 måste stå antingen här eller där; den kan inte stå här; alltså ska den stå där”. En vanlig deduktiv slutledning, alltså. På vilket sätt är den ”intuitiv”? Hur handlar den om ”motiv”? Och på vilket sätt är förmågan att göra sådana slutledningar behäftad med ”allvarliga brister”?

(Först en lite petig terminologisk synpunkt. Den gäller termen ”mo-

tiv" och tesen att "förnuftet är en mekanism för intuitiva slutledningar om motiv". Termen "motiv" spelar en central roll i boken. Men författarna säger också (i den svenska översättningen): "Det engelska ordet för motiv, 'reason', har ingen direkt motsvarighet i andra språk" (s. 141). Det låter konstigt! Ordet "reason", i den betydelse som kommer i närheten av "motiv" motsvaras på svenska – åtminstone i filosofiska sammanhang – av ordet "skäl". Det brukar ju anses att man som rationell person bör ha *skäl* för vad man tror och vad man gör. Man kan undra hur den engelska formulering ser ut som översatts till citatet ovan! "The English word for motive, 'reason', ..." kan det väl knappast stå?

I det engelska originalet formuleras kanske författarnas tes så att "reason is a mechanism for intuitive inferences about reasons"? Det är inte så lätt att veta. Originalets titel anges inte i den svenska utgåvan, men på nätet hittar man en bok av författarna med titeln *The Enigma of Reason: A New Theory of Human Understanding*. Det är antagligen den engelska förlagan, även om den svenska titeln inte riktigt har samma innebörd.)

Låt oss återgå till deduktiva slutledningar, som t.ex. kan förknippas med Sudoku. De "skäl" (eller motiv?) som figurerar när man löser Sudoku är kanske de premisser man utgår från. Eller är det kanske en önskan att träna sin hjärna? Eller vad avses?

Hur som helst misslyckas människor ofta med Sudoku. Detsamma gäller ett exempel som förekommer i boken, nämligen det rätt kända "fyrkorts-problemet", där man ska avgöra vilka kort man måste vända på för att avgöra om siffran 2 finns på ena sidan av varje kort som har bokstaven E på den andra (se t.ex. s. 59). Även här är det fråga om enkla deduktiva slutledningar, men endast ungefär 10 procent av försökspersonerna ger det rätta svaret.

Består "bristerna" hos förnuftet alltså i att man gör logiska misstag? Och varför anser i så fall Mercier och Sperber att dessa brister inte alls är några brister?

Nej, detta är nog inte alls vad de menar.

Om vi tar fasta på att förnuftet har med slutledningar att göra, så tror jag att man måste skilja mellan deduktiva och induktiva slutledningar – en helt fundamental distinktion som Mercier och Sperber inte verkar vara medvetna om. Deras tes gäller nog enbart, eller framför allt, induktiva slutledningar. Det vill säga slutledningar av typen "Det är mörka moln på himlen, så det kommer snart börja regna". Man tror att det kommer att regna och ens skäl för att tro det är att man ser mörka moln

på himlen. Eller: man anser att Sverige ska gå med i Nato och ens skäl att tycka det är att Ryssland kraftigt rustar upp sin militära förmåga.

Mercier och Sperber menar att slutledningar styrs av en sorts kausala ”mekanismer” eller mentala ”moduler”. Slutledningsmekanismerna, modulerna, är specialiserade dispositioner som utnyttjar naturens regelbundenheter framgångsrikt, eftersom de selekterats av evolutionen. Detta antyder att de egentligen bara talar om induktiva slutledningar (så det är väldigt oklart varför de alls lägger vikt vid fyrkorts-problemet).

Men att induktiva slutledningar – eller argument – är osäkra och ofta kontroversiella är ju välkänt. Och det Mercier och Sperber tar fasta på är att de skäl man uppmärksammar och hänvisar till ofta är sådant som gynnar den egna ståndpunkten. Det är detta de menar brukar betraktas som en ”brist”, en brist de kallar ”bekräftelsebias” (s. 261) eller hellre ”myside-bias” (s. 271). Det kan tyckas vara förnuftigt att vara mer självkritisk och även leta efter skäl som talar *mot* den egna ståndpunkten. Hur kan evolutionen ha gynnat uppkomsten av myside-bias, frågar sig författarna.

De anser att detta måste framstå som ett problem för folk som har en traditionell teori om förnuftet. Med en bättre teori – nämligen deras egen teori – blir det inget problem och det som kunde tyckas vara en ”brist” är ingen brist. Myside-bias är nämligen funktionellt om det kombineras med meningsutbyten med andra. En sorts arbetsbesparande arbetsdelning, alltså (s. 273).

Man kan undra vari det nya består, som skiljer deras teori från tidigare idéer.

Mercier och Sperber skriver: ”De traditionella idéerna om förnuftet handlar inte om en psykologisk mekanism utan om ett grundläggande och transcendent element i människans psyke i sin helhet. […] Den modul vi talar om, om den existerar, skulle inte vara något människor skulle vilja skryta med på det sättet de har skrutit om Förnuftet. Trots det kan det som kommer närmast det klassiska Förnuftet i människans psyke/hjärna mycket väl vara den här modulen” (s. 181–82).

Mercier och Sperber konstruerar alltså en motsättning mellan en traditionell ”intellektualistisk” teori om förnuftet, enligt vilken detta består i att ”få oss att nå fram till bättre övertygelser och att fatta bättre beslut på egen hand”, med sin egen ”interaktionistiska” teori, enligt vilken förnuftets funktion består i att ”ta fram och bedöma rättfärdiganden och argument i dialog med andra” (s. 251). Men denna motsättning är såvitt jag förstår väldigt överdriven. Traditionellt har man ju ofta ansett

att det mänskliga förnuftet speciellt visar sig inom vetenskapen – och vetenskapen är ju en social institution där man i hög grad bedömer argument ”i dialog med andra”. Å andra sidan använder ju människor i vissa situationer sitt förnuft helt ”på egen hand”, t.ex. när de försöker lösa Sudoku-problem eller andra problem som psykologer utsätter dem för i sina experiment. Ett extremt exempel är matematikern Andrew Wiles arbete ”i nästan total hemlighet” i sex år för att bevisa Fermats sats (jfr s. 397). Han använde ju sitt förnuft ”på egen hand”. Och ur ett evolutionistiskt perspektiv bör ju även den som kan använda sitt förnuft i situationer där man inte har tillgång till andras synpunkter och argument också ha en reproduktiv fördel i kampen för tillvaron.

Det är väl helt enkelt så att när man inte har tillgång till andras synpunkter och motargument, så får man försöka konstruera sådana på egen hand. Somliga klarar antagligen av detta bättre än andra. Men Mercier och Sperber anser att ”människor har förmågan att resonera objektivt genom att förkasta svaga argument och godta starka argument” när de tar ställning till de skäl som andra människor hänvisar till, men de använder inte denna förmåga när de bedömer de skäl de själva bygger på (s. 289). Detta är, menar de, vad man kan vänta sig enligt den interaktionistiska teorin, men det strider mot den så kallade ”intellektualistiska teorin”.

Boken är rätt pratig och krångligt skriven och därmed ganska svårläst. (Hur ska man t.ex. förstå författarnas tes att de ”traditionella idéerna om förnuftet handlar inte om en psykologisk mekanism utan om ett grundläggande och transcendent element i människans psyke i sin helhet”? Och hur ska man avgöra om det stämmer?) Boken är också alldeles för lång, dess huvudtankar är inte särskilt uppseendeväckande och de hade bättre kommit till sin rätt i en kortare uppsats. Det är egendomligt att den tycks ha uppskattats av flera bedömare.

LARS BERGSTRÖM

NOTISER

ETT MONUMENT FÖR DE MONUMENTLÖSA

Om man slår upp sitt exemplar av Strindbergs lilla novell ”Betraktelser på kyrkogården” från 1896, som handlar om Cimetière du Montparnasse i Paris, kan man läsa om ett särskilt vackert monument: ”det är skulpterat i röd granit, det bär ingen inskrift. Jag har länge förväxlat det med den spruckna kolonnen, Minnesmonumentet, till hågkomst av dem som sakna monument.”

Det är närliggande, för en filosof, att uppfatta detta som ett exempel på Russells paradox eller barberarparadoxen – den om barberaren som bara får raka dem som inte rakar sig själva. Monumentet är för dem som saknar monument, vilket innebär att dessa har ett monument, vilket innebär att monumentet inte är för dem, eftersom de i så fall har ett monument, vilket gör dem utan monument, vilket …

Russell fann sin paradox år 1901 och skrev om den i ett brev till Frege år 1902, men det sägs att man i Göttingen kände till problemet redan på tidiga 90-talet. Det låg väl i luften, Pariserluften, eller så hade någon läst Strindberg.

ROBERT CALLERGÅRD

LENA ANDERSSON OCH POPPER

Lena Andersson har under 2000-talet varit kulturdebattens kanske främsta kritiker av den postmodernistiska och socialkonstruktivistiska kunskapsnihilismen och värderelativismen. För detta vill jag ge henne mycket beröm; och jag är helt säker på att Karl Popper skulle gjort detsamma. Hon har dessutom varit en hård kritiker av identitetspolitiken, vilket hon förtjänar lika mycket beröm för; jag tror Popper skulle instämt även här.

I mars i år kom hon ut med boken *Om falsk och äkta liberalism*, i vilken Popper, författare till *The Open Society and Its Enemies*, kallas en falsk liberal. Och inte nog med det. Han beskylls dessutom för att ha banat väg för den postmodernistiska kunskapsnihilismen. Hur går detta ihop?

I sin bok försöker Andersson utveckla en specifik syn på kunskapsinhämtning och på politik som grundad i moralfilosofi. Efter läsningen

ekade i mitt huvud uttrycket ”Det är lättare att vara kritisk än att vara konstruktiv”. Hon försvarar helt överraskande en rationalistisk platonsk-aristotelisk kunskapsteori och en libertariansk (Nozickiansk) naturrättslära. Den empiriskt sinnade socialliberalen Popper kallas falsk liberal därför att han varken omfattar den nämnda kunskapsteorin eller naturrättsläran. Båda sakerna måste enligt Andersson varje äkta liberal omfatta.

Jag skakade på huvudet när jag lade ifrån mig boken. Och om Popper visste hur han blivit framställd, så skulle han nog sucka i sin grav. Inte därför att han inte lever upp till Anderssons äkthetskontroll. Härom skulle han säkert bara säga ”Är detta äkta liberalism vill jag verkligen inte kallas liberal”. Sucken skulle komma därför att Andersson i tre avseenden ger en felaktig beskrivning av hans filosofiska positioner.

1. Popper sägs ha varit emot spekulation (s. 35–36, 126), men det är fel. Han ansåg att metafysisk spekulation ibland är en förutsättning för att få kunskap om världen, och gick hårt åt positivisterna som ansåg den meningslös. Men till skillnad från Andersson, så ansåg han att spekulationer om världen måste ordentligt förbindas med empiriska data innan de börjar betraktas som sanningslika.

2. En av Anderssons nödvändiga betingelser för att få kalla sig äkta liberal är att man förkastar nominalismen (s. 18). Farväl alla ni skenliberala brittiska empirister. Och också Popper framställs som nominalist (s. 127–28), men till skillnad från de brittiska empiristerna var han inte det.

En ytlig läsning av kap. 11: II i *The Open Society and Its Enemies* utan verklig kännedom om Poppers andra böcker har hos Andersson gett upphov till den felaktiga tron att Popper var nominalist. Faktum är att Popper utvecklade en ontologi enligt vilken vår värld är skiktad i tre typer av världar: den materiella verkligheten (värld 1), den verklighet som består av medvetandeakter (värld 2) och en värld av abstrakta objekt (värld 3). De abstrakta objekten i värld 3 är människoskapade, och alltså inga platonska idéer, men hans uppfattning om värld 3 har trots detta inget som helst gemensamt med en nominalistisk uppfattning av abstrakta objekt. I det nämnda kapitelavsnittet kritiserar Popper den platonsk-aristoteliska *definitionsläran*, mot vilken han ställer en nominalistisk sådan, men det gör honom inte till nominalist i den av Andersson definierade meningen. Det centrala för honom är att kritisera den platonsk-aristoteliska tron på helt säker kunskap och i stället förorda sin egen fallibilistiska kunskapsrealism.

3. Popper gjorde sig först ett filosofiskt namn p.g.a. sin falsifierbarhetsprincip, med vilken han ansåg man kunde skilja vetenskap från metafysik. I början gav han den mycket hårda och bestämda formuleringar,

men efterhand blev formuleringarna innehållsligt tunnare och tunnare. Till slut betydde den bara att en empirisk vetenskaplig teoribyggnad måste på något sätt i princip kunna stöta på data som inte stämmer med teorin. I slutet på 1970-talet slutar han i tysthet att använda termen. Andersson relaterar sig enbart till den tidige Popper (s. 35, 83), och för det må hon kanske vara ursäktad. Men när saken adderas till punkterna 1 och 2 blir det ändå viktigt. Det för mig till nästa påpekande: Andersson gör Popper delansvarig för vår tids kunskapsnihilism.

Under hela sitt liv försvarade Popper en både anti-positivistisk och anti-platonistisk kunskapsrealism. Han utgår (likt Andersson) från att naturen har en bestämd struktur oberoende av språket, och att vi kan få viss kunskap om den. Hur Andersson kan få detta till att Poppers åsikter ”drar i riktning mot” (s. 51) kunskapsnihilism, och att dagens kunskapsnihilister är ”efterföljare” (s. 36) till Popper är för mig en verklig gåta.

Kanske har hon inom sig känt på sig att vissa saker hon säger inte stämmer. Ty hon föregriper kritik från filosofer genom att i förordet säga att hon skrivit sin bok i egenskap av skönlitterär författare. Märkligt, eftersom hon sedan i boken deklarerar att hon har platonsk-aristoteliska åsikter med vars hjälp hon vill komma åt alla sofister. Jag citerar:

> Denna text är skriven i min egenskap av författare med filosofiska intressen och det är så den önskar bli läst. I min skönlitteratur gestaltar jag världen genom människor som möts och testas mot varandra, här genom idéer som utsätts för samma förfarande. Som författare tillåter jag mig att presentera förutsättningar och samband som prövas av sin giltighet och användbarhet, på det sätt som också romanfigurers halt avgörs av om de engagerar fantasin och vidgar förståelsen. (s. 11)

Men Popper kan inte behandlas som en romanfigur. Han föddes 1902 och dog 1994.

INGVAR JOHANSSON

Medarbetare i detta nummer av *Filosofisk tidskrift*: *Lars Bergström* är redaktör för tidskriften, *Katharina Berndt Rasmussen* och *Åsa Burman* är forskare på Institutet för framtidsstudier, *William Bülow* är postdoc på The Stockholm Centre for the Ethics of War and Peace vid Filosofiska institutionen i Stockholm, *Bertil Kristerson* har doktorerat i tros- och livsåskådningsforskning, *Nicolas Olsson Yaouzis* är forskare vid Filosofiska institutionen i Stockholm och *Ruben Wiwe* är gymnasielärare vid Rudbecksgymnasiet i Sollentuna.

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