Människans mörka materia: Om implicit bias och moraliskt ansvar[1]

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).[2] 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.

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 istället 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 de 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,

[1] Berndt Rasmussen, Katharina & Åsa Burman (kommande), *Filosofisk tidskrift*.
[2] För en närmare beskrivning av IAT, se avsnitt två nedan.

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.[3] 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.[4] 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.[5] 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 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.[6]

---

[3] 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 (Saul 2013b), (Di Bella m.fl. 2016).

[4] 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/).

[5] Se t.ex. Greenwald m.fl. 1998; Krumpal 2013.

[6] Jfr. Holroyd & Sweetman 2016, s. 81, Saul 2013b, s. 40.

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.[7]

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 bias", "implicita fördomar", "implicita skevheter”, och "implicit social kognition"[8] 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.[9] 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

[7] Se Rubin 1983.
[8] Jfr. Greenwald & Banaji 1995.
[9] Gendler 2008a, Gendler 2008b.

attityder.[10] 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.[11] Mandelbaum kritiserar också begreppet ”alief” för att vara en icke-enhetlig ”sammelsurium”-kategori.[12] Tim Bayne och Anandi Hattiangadi hävdar att begreppet ”alief” står inför ett 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.[13] 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.[14]

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.[15] 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”.[16] Sarah-Jane Leslie menar istället 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”.[17]

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 kvinnor och män är lika lämpade för att utbilda sig inom och jobba med naturvetenskap.[18]

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

[10] Jfr. Brownstein 2017, paragraf 2.1.
[11] Se t.ex. Brinol m.fl. 2009.
[12] Mandelbaum 2016.
[13] Bayne & Hattiangadi 2013.
[14] Glüer & Wikforss 2013.
[15] Mandelbaum 2016.
[16] Haslanger 2013, s. 11, vår övers.
[17] Leslie 2017.
[18] Schwitzgebel 2013.

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.[19] 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,[20] kallar någon med arabisktklingande namn till anställningsintervju,[21] uppför sig hövligt mot kvinnor respektive mörkhyade på arbetsplatsen,[22] 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.[23]

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.[24] Andra studier indikerar att IAT-resultaten speglar kulturella stereotyper och ojämlikhet som föreligger i det omgivande samhället.[25] 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:

---

[19] 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).

[20] Greenwald m.fl. 2009. Deras resultat motsägs dock i viss mån av Ditonto m.fl.2013.

[21] Rooth 2010.

[22] Cortina m.fl. 2013, Dovidio m.fl. 2002.

[23] Greenwald m.fl. 2009.

[24] 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.

[25] Se t.ex. Karpinski & Hilton 2001, Arkes & Tetlock 2004.

> 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.[26]

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 mellan 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.[27] 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).[28] 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.[29] Detta indikerar att det föreligger en skevhet redan i vår *varseblivning* – oavsett vilken typ av implicit skevhet som antas ligga 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."[30]

---

[26] Gendler 2008b, s. 577, vår övers.
[27] Mekawi & Bresin 2015.
[28] Unkelbach m.fl. 2008.
[29] Payne 2001, jfr. Saul 2013a.
[30] 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).

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.[31]

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.[32]

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.[33]

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 inre mörka materia som det som IAT mäter. Testerna kanske spårar helt skilda mentala fenomen?[34] 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 bias-tester,[35] och även för testerna där självsäkra kvinnors lämplighet för chefspositioner skulle utvärderas.[36] 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.[37]

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

[31] 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 svenskklinagnde namn följer mönstret vanlig könsdiskriminering (där män kallas oftare på intervju än kvinnor).
[32] Rudman & Glick 1999.
[33] Jfr. Holroyd 2012, s. 276.
[34] Jfr. Holroyd & Sweetman 2016.
[35] Se t.ex. Correll m.fl. 2002, Banks m.fl. 2006.
[36] Rudman & Glick 2001.
[37] Rooth 2010.

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. Statistisk 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; dessutom jobbar kvinnor och män inom en till stora delar könssegregerad arbetsmarknad, och så vidare.[38] 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.[39]

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,[40] respektive anser att jämställdhet är ett politiskt viktigt mål.[41] 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 dynamiska 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.[42] 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

[38] SCB 2016.
[39] Socialstyrelsen 2010. För undersökningar kring etnisk diskriminering på den svenska arbets- resp. bostadsmarknaden, se Bursell 2014 resp. Ahmed & Hammarstedt 2008.
[40] 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.
[41] SIFO 2014. Jfr. även World Values Survey Association 2014.
[42] Schelling 1971. För en enkel webbaserad simulering av Schellings modeller, se www.ncase.me/polygons/.

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.[43]

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.[44] 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.[45]

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

[43] Martell m.fl. 1996. Jfr. även Martell m.fl. 2012. För en enkel webbaserad simulering av deras modell, se www.doesgenderbiasmatter.com.
[44] Nosek m.fl. 2009.
[45] Hehman m.fl. 2018.

> 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.[46]

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.[47] 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 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”.[48] 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.

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,

[46] Saul 2013b, s. 55, vår övers.
[47] Holroyd 2012.
[48] Ibid., s. 294, vår övers.

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.[49] 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.[50]

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. Istället 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 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 mikro-beteenden: 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

[49] Se t.ex. Devine m.fl. 2002.
[50] Holroyd 2012, s. 282.

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 urskiljningslöst), 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.[51]

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 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

[51] Saul 2012, s. 260–261, vår övers.

gruppmedlemmar i exempelvis en anställningskommitté eller som en grupp, exempelvis som en myndighets personalavdelning.[52]

Referenser

Ahmadi, Fereshteh, Irving Palm & Nader Ahmadi. 2016. ”Mångfaldsbarometern”. Gävle: Högskolan i Gävle. https://www.hig.se/Ext/Sv/Press/2016-10-19-Mangfaldsbarometern-2016---Forandringar-i-opinionen.html.

Ahmed, Ali M. & 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 & 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. & 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 & 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 & Anandi Hattiangadi. 2013. “Belief and Its Bedfellows”. I: *New Essays on Belief*, London: Palgrave Macmillan, . https://doi.org/10.1057/9781137026521_7.

Brinol, Pablo, Richard Petty & Michael McCaslin (2009). “Changing Attitudes on Implicit versus Explicit Measures: What is the Difference?” I: R. Petty, R., Fazio & P. Brinol (red.) *Attitudes: Insights from the New Implicit Measures*. New York: Psychology Press.

Di Bella, Laura, Eleanor Miles & Jennifer Saul. 2016. “Philosophers Explicitly Associate Philosophy with Maleness”. I: Michael Brownstein & Jennifer Saul (red.) *Implicit Bias and Philosophy, Volume 1*. Oxford: Oxford University Press , s. 283–308. https://doi.org/10.1093/acprof:oso/9780198713241.003.0012.

Brownstein, Michael. 2017. “Implicit Bias”. I: Edward N. Zalta (red.) *The Stanford Encyclopedia of Philosophy*. 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 & 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.

[52] Författarna vill tacka Moa Bursell, Johan Ekelund, Martin L. Jönsson, Filip Olsson, Jennifer Saul, samt deltagarna i forskarseminariet på Institutet för framtidsstudier (okt. 2018) för värdefulla kommentarer. Artikeln skrevs inom forskningsprojektet ” Individuellt och kollektivt ansvar för diskriminering p.g.a. implicita fördomar”, finansierat av Jane och Dan Olssons Stiftelse för Vetenskapliga Ändamål (Dnr. 2016-23).

Cortina, Lilia M., Dana Kabat-Farr, Emily A. Leskinen, Marisela Huerta & 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. 835–848.

Ditonto, Tessa M., Richard R. Lau & 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 & 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 & 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 & Åsa Wikforss. 2013. "Aiming at Truth: On The Role of Belief". *Teorema: International Journal of Philosophy* 32 (3), s. 137–162.

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

Greenwald, A., D. McGhee & J. Schwartz. 1998. "Measuring individual differences in implicit cognition: The Implicit Association Test". *Journal of Personality and Social Psychology* 74, s. 1464-1480.

Greenwald, Anthony G., T. Andrew Poehlman, Eric Luis Uhlmann & 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 & 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, & 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 & Joseph Sweetman. 2016. "The Heterogeneity of Implicit Bias". In *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, & James L. Hilton. 2001. "Attitudes and the Implicit Association Test". *Journal of Personality and Social Psychology, 81*(5), s. 774-788.

Krumpal, Ivar. 2013. "Determinants of social desirability bias in sensitive surveys: a literature review", *Quality & Quantity* 47 (4), s. 2025–2047.

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 & 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 & Cynthia G. Emrich. 1996. "Male-Female Differences: A Computer Simulation". *American Psychologist* 51 (2), s. 157–58.

Mekawi, Yara & 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 & Irving Palm. 2014. "Mångfaldsbarometern". Gävle: Högskolan i Gävle. https://www.hig.se/Ext/Sv/Arkiv/Tidigare-nyheter-2014/2014-11-28-Mangfaldsbarometern-2014-visar-att-fler-ar-positiva-till-mangfalden---men-det-finns-orosmoln.html.

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 & 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. & 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. & 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?* 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/levnadsforhallanden/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/10.1080/0022250X.1971.9989794.

Schwitzgebel, Eric. 2013. “A Dispositional Approach to Attitudes: Thinking Outside of the Belief Box”. I: *New Essays on Belief*. London: Palgrave Macmillan,s. 75–99. 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 & 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.