FILOSOFISK TIDSKRIFT

4/95

från filosofidagarna 1995

med bidrag av:

PAUL NEEDHAM	MARTIN GUSTAFSSON THOMAS WETTERSTRÖM
LARS BERGSTRÖM KRISTER BYKVIST	ERIK CARLSON INGMAR PERSSON
JAN ÖSTERBERG INGAR BRINCK	PETER PAGIN JOHN CANTWELL
JAN ODELSTAD BERTIL ROLF	WLODEK RABINOWICZ RYSIEK SLIWINSKI

Förord

Syftet med detta specialnummer av Filosofisk tidskrift är att manifestera den nationella konferens i filosofi och vetenskapsteori som ägde rum i Umeå 7-9 juni 1995 och som gick under namnet *Filosofidagarna*. Arrangör var det nybildade *Svenska filosofisällskapet* i samarbete med institutionen för filosofi och vetenskapsteori vid Umeå universitet. Konferensen samlade drygt hundra deltagare varav ett sjuttiotal föredragshållare. Vi hoppas att uppsatserna i denna volym – samtliga baserade på föredrag från filosofidagarna – skall ge en uppfattning om den vitalitet, bredd och höga kvalitet som genomgående präglade inläggen vid filosofidagarna i Umeå.

Förhoppningsvis kommer Filosofidagarna att bli ett stående inslag i svenskt filosofiskt liv. Nästa Filosofidagar planeras till Lund 1997.

Sten Lindström
Bertil Strömberg

INNEHÅLL

Om Ivar Segelberg, medvetandet och världen 5
THOMAS WETTERSTRÖM

Duhems quineska realism 26
PAUL NEEDHAM

Några kommentarer till Wittgensteins "Om visshet" 41
MARTIN GUSTAFSSON

Utilitarismens fel och förtjänster 49
LARS BERGSTRÖM

Är de flesta utilitarister deontologer? 62
KRISTER BYKVIST

Några teorier om egenvärde 75
ERIK CARLSON

Lika för olika 86
INGMAR PERSSON

Hur många tolkningar har "Macbeth"? 97
JAN ÖSTERBERG

Offentlighet och obestämdhet hos språklig mening 111
PETER PAGIN

Indexikal kunskap och social mening: om ordet 'jag' 128
INGAR BRINCK

Pågående handlingar 137
JOHN CANTWELL

Om trovärdighet 151
BERTIL ROLF

Förväntad nytta i ljuset av alternativa numeriska representationer 160
JAN ODELSTAD

Om Seidenfelds kritik av sofistikerade brott mot oberoendeaxiomet 169
WLODEK RABINOWICZ

Baklänges induktion i spelteori 186
RYSIEK SLIWINSKI

Notis 196

Filosofisk tidskrift nr 4 1995 årg 16. Filosofisk tidskrift utges av Stiftelsen Filosofisk tidskrift och Bokförlaget Thales.

Redaktör och ansvarig utgivare: Lars Bergström
Redaktionskommitté: Aant Elzinga (Göteborg) Göran Hermerén (Lund), Göran Lantz (Uppsala), Sven-Eric Liedman (Göteborg), Åke Löfgren (Karlstad), Giuliano Pontara (Stockholm), Dag Prawitz (Stockholm) och Dag Westerståhl (Stockholm)
All redaktionell post sänds till redaktören, Lars Bergström, Reimersholmsgatan 39, 117 40 Stockholm
Produktion och prenumerationer:
Bokförlaget Thales, Djupdalskvior 5, 620 12 Hemse, förlagsredaktör Ulf Jacobsen, tel 0498-48 48 93
Prenumeration: Tidskriften kommer ut med 4 nummer/år, pris 120 kronor, pg 50 79 91-8 Filosofisk tidskrift
Lösnummer: 35 kronor. Nr 4 1995 75 kronor.
Satt av Ord och Grafik, Stockholm
Tryckt hos Lagerblads tryckeri AB, Karlshamn, 1996
ISSN 0348-7482

THOMAS WETTERSTRÖM

Om Ivar Segelberg, medvetandet och världen

Ett av två syften med denna uppsats är att stimulera intresset för några, som jag tycker, beaktansvärda inslag i Ivar Segelbergs relativt förbisedda filosofi, främst i hans teori om upplevelsers natur. Det andra syftet är att med utgångspunkt från diverse kritiska synpunkter på den teorin formulera några fenomenologiska och metafysiska idéer ur egen fatabur. Framhållas bör dock att frågan om *medvetandets enhet* inte kommer att diskuteras, detta trots att lösningen av just det problemet är den övergripande målsättningen för Segelbergs medvetandefilosofiska huvudarbete (*MJ*).

En annan upplysning som redan här bör ges är att jag genomgående använder orden *'ontologi'* och *'metafysik'* efter högt föredöme: på Anders Wedbergs vis. "Med en ontologisk teori menar jag", säger Wedberg, "en teori om de allmännaste kategorier, vari "allt" [...] kan indelas. [...] Med en metafysisk teori menar jag en teori [...] som – om den förknippas med en explicit ontologi – talar om vad som skall stoppas in i de skilda kategoriernas fållor" (1966, s 47 n).

1. Kvaliteter och moment

Endast i ringa mån kan Segelbergs medvetandefilosofi formuleras annat än i termer av hans *momentontologi.* För att kort kunna presentera denna, måste jag be läsaren betrakta två konkreta objekt. Låt oss ta två homogent rödfärgade, cirkelrunda ytor som är exakt likadana med avseende på *kvaliteter,* d ä inre egenskaper som noga måste skiljas från *relationsegenskaper,* t ex att vara far, bosatt i en storstad eller medlem av Skattebetalarnas förening. Exempel på kvaliteter är röd, färgad, cirkelformad, rumsligt utsträckt och si-och-så-strukturerad.

De båda ytorna kan naturligen sägas ha *samma* kvaliteter, nämligen cirkularitet, utsträckning och (en lägsta determinat av) rödhet. Var och

en av dem kan emellertid också ses som en helhet i vilken individuella förekomster eller *moment* av nyssnämnda kvaliteter ingår. Med detta momentontologiska betraktelsesätt kan *den ena ytan* sägas innehålla ett moment *c* av cirkularitet, ett moment *u* av utsträckning och ett moment *r* av rödhet. Vidare kan momentet *c* sägas vara exakt likt (men numeriskt different från) ett kvalitetsmoment *c** som ingår i *den andra ytan,* och såväl *u* som *r* kan sägas ha samma relation till momenten *u** resp. *r** vilka tillsammans med av *c** ingår i sistnämnda yta.

'Moment' är Edmund Husserls term för i viss mening abstrakta delar av helheter (se 1913, del III). Husserlska *Momente* är abstrakta = osjälvständiga: de kan inte tänkas existera oberoende av de helheter i vilka de ingår. Det finns emellertid åtminstone en prominent momentontolog som förkastar distinktionen mellan självständiga och osjälvständiga delar. Den momentontologen är ingen mindre än Segelberg. Distinktionen i fråga attackerar han sinnrikt i *BE* (kap 4, sekt 2). Trots att jag för egen del håller på den, måste jag här avstå från varje försök att utlägga eller försvara den. Termen 'moment' reserverar jag emellertid för entiteter som i likhet med *c, u* och *r* uppfyller följande villkor: *om* det över huvud taget finns någon plausibilitet hos tesen att de med avseende på själv- eller osjälvständighet har samma status som uppenbart självständiga delar av världen, t ex hästhuvuden eller de tegelstenar av vilka en tegelmur är byggd, så härleder sig denna plausibilitet från en sofistikerad och för common sense främmande argumentation.

Även förekomster av komplexa kvaliteter kan uppfylla detta villkor: entiteter som i *min* terminologi är moment kan mycket väl vara komplexa. Men eftersom det för segelbergska och husserlska moment verkar gälla att de alltid – eller, vad Husserl beträffar, åtminstone vanligtvis – är enkla, uppstår för denna framställnings vidkommande ett terminologiskt problem som jag löser så: fram till fjärde stycket av sektion 10 är om motsatsen inte tydligt markeras (explicit eller av kontexten) termen 'moment' förbehållen enkla entiteter. Därefter spelar det ingen roll om de entiteter som kallas moment är enkla eller komplexa.

Att umbära *relationsuniversalia* är inte lättare för en momentontolog än för någon annan, men *kvalitetsuniversalia* kan en sådan ontolog om hon så vill skära bort med Ockhams rakkniv; ty kvaliteter kan hon konstruera som klasser av moment mellan vilka *likhets-*

relationer råder (jfr *BE,* kap 2, särskilt s 28 f resp 47 ff; och *MJ,* s 6 f). Lägsta kvalitetsdeterminater blir i så fall segelbergska *kongruensklasser,* d ä klasser av *exakt* likadana kvalitetsmoment, och determinabler blir unioner av kongruensklasser (jfr *BE* och *MJ,* s 28 resp 6 f).

2. Ur Segelbergs allmänna föremålslära

I sin undervisning angrep Segelberg ofta tes 1.1 i Wittgensteins *Tractatus*: att världen är inbegreppet av fakta, icke av ting. Världen är, hävdade han, ett ting i följande bemärkelse: den är en komplex enhet bestående av ett otal med varandra *elementarförbundna* kvalitetsmoment. (Till begreppet elementarförbindelse återkommer jag.)

Med avseende på Segelbergs värld måste emellertid en distinktion dras mellan den *konkreta* och den *totala* världen (termerna är mina); och det är endast den förra som är en komplex enhet av elementarförbundna moment. I Segelbergs terminologi är såväl moment som enheter av nyssnämnt slag *föremål av första ordningen.* Den konkreta världen är alltså själv är ett föremål av första ordningen, och samma sak gäller varje innehåll i den. Beträffande den totala världen gäller att den utöver föremål av första ordningen bl a innehåller *fakta.* Ett faktum är ett *föremål av andra eller ännu högre ordning,* och utmärkande för varje entitet av sistnämnt slag är att den innehåller åtminstone en *relation.* Relationer i sin tur är inte föremål av vare sig första eller högre ordning; de är, menar Segelberg (och jag protesterar inte), för varje rimlig ontologi oumbärliga universalia sui generis. Relationer och moment "utgöra i viss mening de element av vilka universum ytterst är uppbyggt" (*BE,* s 137).

För mig gåtfull är emellertid den existens-status som relationer och föremål av högre ordning har enligt Segelberg. *Existerar* dessa entiteter i samma bemärkelse som vanliga ting? Av allt att döma menar han att så är fallet, men för mig är denna ståndpunkt i det närmaste obegriplig. Vi har här att göra med ontologiska spörsmål, och sådana uppfattar jag inte som sannings- utan som lämplighetsfrågor. Uppgiften att utlägga denna tanke kan inte tacklas här. Men tesen att inga relationer ingår i komplexa föremål av första ordningen bör förtydligas, och *då* kommer begreppet elementarförbindelse in i bilden.

3. Om elementarförbindelser, kollektioner och efs-komponenter

Betrakta en av de båda cirkelrunda ytor som figurerade i sektion 1. Låt

oss ge namnet 'Caesar' åt den yta som innehåller cirkularitetsmomentet *c*, utsträckningsmomentet *u* och rödmomentet *r*. Caesar är en komplex enhet i kraft av att elementarförbindelser råder mellan vissa av dess delmoment. Vilka? Uppenbart torde vara att *r* och *u* har en relation till varandra som med fog kan kallas elementarförbindelse; och att samma sak gäller för *c* och *u*. Men med *r* och *c* ligger det annorlunda till, ty *r* saknar utsträckning, och en outsträckt entitet kan inte begripligen sägas vara formad.

Mer allmänt kan sägas att det verkar finnas fundamentala lagar enligt vilka somliga kvaliteter *inte passar ihop*. Sålunda torde det vara *omöjligt* för ett ljud, en smärtförnimmelse eller en smaksensation att vara *bokstavligen* formad eller färgad. Har vi här att göra med syntetiska sanningar a priori? Jag tror det ogärna men är långt ifrån säker. Hur som helst torde mina exempel belägga att det finns lagar som separerar vissa sorters kvalitetsmoment från varandra, och jag tror att jag uttalar mig i överensstämmelse med Segelbergs intentioner när jag nu säger att det är i kraft av sådana lagar som elementarförbindelser råder eller fallerar att råda mellan de kvalitetsmoment som ingår i den konkreta världen.

Givet att '><' och '|' betyder 'är elementarförbundet med' resp 'är icke elementarförbundet med' gäller alltså följande: $u >< c$ & $u >< r$ & $r \mid c$. Men kan man då med plausibilitet förneka att Caesar innehåller relationer? Ja. Det finns en uppenbar bemärkelse in vilken ett relationsuniversale, nämligen *likhet,* ingår i t ex det *faktum* att rödmomentet *r* är likt det i sektion 1 figurerande rödmomentet *r**. Analogt gäller att relationsuniversalia ingår i de *fakta om Caesar* som satsen '$u >< c$ & $u >< r$ & $r \mid$ c' korresponderar till. Två av dessa tre fakta innehåller relationen *elementarförbundenhet* och det tredje innehåller relationen *icke-elementarförbundenhet*, men att Caesar själv skulle *innehålla* relationsuniversalia torde framstå som ett absurt påstående. En helt annan sak är att Caesar med god mening kan sägas *exemplifiera* ifrågavarande relationsuniversalia.

Men vad innehåller då Caesar utöver *c, u* och *r*? Så länge den frågan är obesvarad kan en eller annan läsare fortfarande förväntas betvivla att första ordningens komplex ej innehåller relationer.

Trivialt är att det i Caesar ingår komplexa enheter innehållande *c* och *u* respektive *r* och *u*. Har Caesar härutöver några hittills obeaktade innehåll, och har dessa i så fall momentstatus?

Även bland segelbergska föremål av första ordningen finns det *komplex* som inte är komplexa *enheter*. Förutom komplexa enheter finns det en helt annan sorts komplex, nämligen *kollektioner*. Betrakta följande citat:

> En hög böcker betraktas i allmänhet ur kollektionens synpunkt: bokhögen anses vara densamma även om ordningsföljden mellan böckerna förändras, så att t.ex. den bok, som ligger överst vid ett första tillfälle, ligger näst överst vid ett senare tillfälle. Men man kan också ändra aspekt och säga, att den helhet, som böckerna bilda i det första fallet är en annan helhet än den som böckerna bilda i det senare fallet. Man har då på bokhögen anlagt enhetens aspekt (BE, s 74 f).

Man har dock, framhåller Segelberg (*BE,* s 75), "uttryckt sig inadekvat" om man säger att de båda här figurerande aspekterna "äro skilda synpunkter på *samma* föremål [...]. Bokhögen med abstraktion från ordningsföljden är ett annat föremål än bokhögen *inklusive* ordningsföljden. Bokkollektionen är en abstrakt komponent i den komplexa enhet, som den konkreta bokhögen utgör" (ibid.). Mot denna bakgrund torde det stå klart att kollektioner enligt Segelberg har – eller (oavsett vad han månde ha tänkt och menat) *borde* anses ha – en med kvalitetsmoment analog, ontologisk status: de måste anses vara individuella, genom intellektuella abstraktionsprocedurer konstruerade entiteter.

Är vi nu färdiga med inventeringen av i Caesar ingående komponenter med momentstatus? Nej. Den *kollektion* som bildas av *c, u* och *r* måste man rimligtvis kunna abstrahera från Caesar, och *det som därvid blir kvar* – låt oss kalla det en *efs-komponent* ('ef' för 'elementarförbindelse'; 's' för 'struktur') – torde näppeligen kunna förmenas momentstatus. Såvitt jag kan förstå måste efs-komponenter klassas som *enkla* moment av *kvaliteter* – strukturkvaliteter av visst slag. Låt *'K(x, y, ...)'* och *'x* och *y* och ... bildar en kollektion' vara synonyma uttryck. Givet denna stipulation kan en av de säregenheter som efs-komponenter ger prov på karaktäriseras så: om *efs* är den i Caesar ingående efs-komponenten, så förmedlas exakt samma information av *'K(c, u, r, efs)'* och *'u >< c & u >< r & r | c'*. I mitt fortsatta ordande om i olika sorters komplexa enheter ingående moment förtiger jag för korthets skull alltid enheternas efs-komponenter.

4. Segelbergs metafysik i huvuddrag

Hur ser Segelbergs *metafysik* ut? Vilka huvudtyper av kvalitetsmoment ingår i hans värld? Följande diagram ger ett översiktligt svar.

Icke mentala moment	Mentala moment		
	Hyletiska	Intentionala	M-moment, d ä moment av kvaliteten medvetenhet

Ingenting tyder på att Segelberg var idealist. Tvärtom. Visserligen har han mig veterligt aldrig rest några anspråk på att kunna *vederlägga* allt vad idealism heter, men det är ändå helt klart att det i hans värld finns icke-mentala föremål. Icke-mentalt är såvitt jag förstår ett segelbergskt *moment* om och endast om det ingår i en komplex enhet (inte nödvändigtvis, menar *jag,* en självständig dylik) som, till skillnad från exempelvis en vaken och verksam person, är *helt och hållet* icke-mental. Ett för Segelberg väsentligt faktum om varje icke-mentalt moment torde vara att det, till skillnad från ett intentionalt eller hyletiskt dito, inte *kan* vara elementarförbundet med ett m-moment.

Till m-momenten återkommer jag, och samma sak gäller andra mentala moment. En diskussion om såväl m-moment som hyletiska dito förutsätter att fenomenet intentionalitet blir hjälpligt klarlagt, och en sådan klarläggning kräver en egen sektion.

5. Om intentionalitet

Anta att en stenåldersmänniska med hjälp av en s k tidsmaskin förflyttades in i ett nutida vardagsrum. Om det funnes krukväxter i rummet, skulle hon antagligen identifiera dessa som *växter,* och samma sak torde mutatis mutandis gälla för djur och människor. Men i övrigt skulle hennes upplevelser skilja sig radikalt från våra. Till skillnad från oss skulle hon inte se bokhyllor, gardiner, lampor, soffor etc. För *dessa* ting skulle hon sakna *begrepp* (inte bara språkliga uttryck); endast hennes mest allmänna begrepp – ting, form, o d – skulle ha tillämpning. *Förnimmelser utan begrepp är blinda,* säger Kant (1781, s 51; 1787, s 75). Sanningen bakom denna slogan torde bl a yttra sig i att stenåldersmänniskans varseblivningar av vardagsrummets inventarier skulle ha ett magert *meningsinnehåll*: om det varken funnes

växter, djur eller människor i rummet och om fönstergardinerna vore fördragna, så skulle *alla* hennes visuella varseblivningar sannolikt bli *Vad är det där för något?*-upplevelser. Någonting som ens avlägset liknade våra *soff-, lamp-, gardin-* eller *bokhylle*varseblivningar skulle hon i varje fall inte kunna få.

Dessa olikheter mellan stenåldersmänniskans och våra varseblivningar är samt och synnerligen skillnader med avseende på *intentionalitet.* Långt ifrån att vara en för *avsikter* specifik egenskap är intentionalitet en egenskap som tillkommer alla upplevelser, varseblivningar såväl som tankar och känslor. Visserligen finns det icke-intentionala *komponenter* i upplevelser, men dessa är inte själva kvalificerade för benämningen 'upplevelse' (mer om detta nedan).

Traditionellt karaktäriseras en upplevelses intentionalitet som dess *riktadhet* mot ett objekt ett s k *intentionalt objekt* som mycket väl kan vara non-existent. Men exemplet med stenåldersmänniskan torde visa att en upplevelses intentionalitet med fog kan sägas vara dess *mening* eller *innebörd.* Att om en upplevelse säga att si-och-så är dess intentionala objekt är *ett* sätt att ange en del av dess mening. Stundom fungerar det väl; ibland är det mindre lyckat.

6. Segelbergs upplevelseteori i huvuddrag

Att intentionalitet tillkommer *varje* entitet som förtjänar att kallas upplevelse är en med Segelbergs upplevelseteori oförenlig tes (mer om detta nedan). I övrigt torde det i föregående sektion sagda vara acceptabelt för honom. Att med sidhänvisningar belägga allt som i det följande sägs om hans medvetandefilosofi är svårt, inte minst. p.g.a. att jag i viss mån stöder mig på minnen av samtal med honom. Den som vill kontrollera min Segelberg-exeges torde i första hand konsultera *MJ,* s 29 ff och 71 ff.

Som framgått finns det enligt Segelberg – eller, försiktigare sagt, *kan* det enligt honom finnas – upplevelser som är renons på intentionalitet. Men den komplikationen skjuter jag på framtiden genom att tills vidare begränsa diskussionen till *tankar* och (sinnliga) *varseblivningar.*

Varje segelbergsk varseblivning innehåller ett m-moment och en intention. Den sistnämnda är en identifierande eller meningsgivande komponent som består av intentionala moment. En segelbergsk tanke – en oåskådlig dylik – består av ett m-moment och *idel* intentionala dito. (Personligen tror jag inte att detta kan gälla för någon tanke, hur

oåskådlig den än må vara – jfr Wetterström 1977 s 47 f och 57 – men den komplikationen behöver inte diskuteras här.)

Utöver intentionala moment innehåller varseblivningar kvalitetsmoment på vilka jag i likhet med Segelberg och utan Husserl-exegetiska ambitioner applicerar den husserlska termen *'hyletisk'*. Låt oss jämföra en oåskådlig tjur-tanke – varför inte en klar och med avseende på medvetenhet relativt intensiv förståelse av yttrandet "Det går en tjur på ängen"? – med en av uppmärksamhet präglad, *visuell* varseblivning av det på ängen betande djuret. Just den olikhet mellan dessa båda upplevelser som konstitueras av att den ena är en tanke och den andra en visuell varseblivning exemplifierar en skillnad med avseende på hyletiska moment. Samma sak gäller just den olikhet mellan en visuell varseblivning och en auditiv dito som består i att den ena är visuell och den andra auditiv; för olikheten mellan en luktvarseblivning och en upplevelse som inte är en luktupplevelse, och analogt för en taktil eller intrasomatisk varseblivning. Vidare – och nu upphäver jag mitt tillsvidare-beslut att begränsa framställningen till upplevelser som är varseblivningar eller tankar – gäller det ovan sagda mutatis mutandis för en smärt-, kväljnings-, andnöds-, skräck-, ilske-, sorg- eller glädjeupplevelse. En upplevelse av sistnämnt slag skiljer sig alltså med avseende på hyletiska moment från varje upplevelse som inte är en glädjeupplevelse.

Segelbergs teori om upplevelsers natur kan nu sammanfattas så: *Varje upplevelse består av ett m-moment* m *och ett antal intentionala* eller *hyletiska moment för vilka gäller att* vart och ett av dem *är elementarförbundet med* m. *Med undantag för* m *är alla dessa moment* upplevelseinnehåll, *och samma sak gäller varje komplex enhet som samtliga eller somliga av dem bildar.*

Värt att notera är att vi i Segelbergs terminologi inte kan uppleva någonting annat än upplevelseinnehåll. Om jag t ex visuellt varseblir en kråka, så upplever jag inte kråkan. Det jag *upplever* är en kråk-intention och en på visst sätt formad, färgad och mönstrad synbild vilken består av hyletiska moment och *illustrerar* kråk-intentionen. *Kråkan,* som jag alltså inte upplever trots att jag ser den, har jag *"intention på"*.

Det kan enligt Segelberg mycket väl vara fallet att varje moment förutom *m* i en upplevelse är hyletiskt, och om så är fallet, så har vi förstås att göra med en *upplevelse som är renons på intentionalitet.*

Idén att det skulle kunna finnas upplevelser av detta slag utvärderas i sektion 9.

7. En antifenomenalistisk teori om hyletiska moment

Föreställ dig att du har visuell varseblivning av en rödfärgad yta och begrunda denna fråga: *Är ett moment av* den rödhet *som du upplever i varseblivningen ett hyletiskt moment i denna och därmed* i bokstavlig *bemärkelse en del av den?*

Segelberg ger ett jakande svar. Färg-, ljud-, gestalt-, smak-, olustmoment o d är, menar han, i bokstavlig bemärkelse innehåll i upplevelser. Stipulativt inför jag termen *'sensationsmoment'* för att beteckna i varseblivningar upplevda moment av färg, ljud, smak, lukt, gestalt, smärta, äckel, fruktan, vrede, glädje o d, och ordet *'sensation'* applicerar jag på komplexa enheter som består av sensationsmoment och som med lika god plausibilitet som t ex synbilder och fenomenella ljud kan anses vara självständiga delar av de helheter i vilka de ingår.

Segelberg kan nu sägas identifiera sensationsmoment med hyletiska dito. Den identifikationen kan man acceptera endast om man ger ett jakande svar på frågan om sensationsmoment i bokstavlig bemärkelse är delar av upplevelser. I tidigare skrifter (1974, 1977) har jag tagit detta svar för givet, men numera vill jag svara nej (därtill i avsevärd mån förmådd av Dick Haglund – inte minst av hans argumentation i 1984 – men han får naturligtvis inte hållas ansvarig för något som sägs i det följande). Detta mitt ställningstagande leder mig fram till en teori om hyletiska moment som i en uppenbar bemärkelse är antifenomenalistisk. Självsvåldigt reserverar jag i det följande termen *'antifenomenalism'* för denna teori. Grovt kan den formuleras så:

Hyletiska moment skiljer sig radikalt från sensationsmoment. De är teoretiska entiteter. *En* av deras uppgifter i en upplevelseteori är att förklara den stora skillnad som sinnliga varseblivningar och oåskådliga tankar alltid ger prov på. För den som inte tillämpar en *instrumentalistisk* teori på dem utan delar min *realistiska* syn på dem är de också en sorts *ting i sig*. Trots att de är komponenter i upplevelser, finns det ingen bemärkelse i vilken de själva kan upplevas. Att de genom sin närvaro i upplevelser påverkar dessas beskaffenheter är en helt annan sak.

Vad talar till förmån för denna teori? Om man identifierar sensationsmoment med hyletiska dito, så förskriver man sig åt tesen att *sen-*

sationer är ting i sig för vilka gäller att upplevelser av dem uppenbarar den slutgiltiga, metafysiska sanningen om deras natur. Men finns det egentligen några goda skäl för tron att synbilder och fenomenella ljud m fl s k sinnesdata i metafysiskt och kunskapsteoretiskt hänseende så radikalt skulle skilja sig från *externa fenomen,* d ä den *vardagligt upplevda yttervärldens* fenomen – tingen omkring oss sådana vi upplever dem? Inte såvitt jag kan finna.

8. F-konstituerade föremål

Mer tilltalande synes mig då en hypotes enligt vilken sensationer och externa fenomen i lika mån är konstituerade i en bemärkelse som jag fortsättningsvis täcker med termen *'f-konstituerad'*.

Ett f-konstituerat föremål existerar inte i den metafysiska världen, d ä den värld som en sann, fullständig och väl bestyrkt metafysisk teori (om den per impossibile kunde åstadkommas) skulle ge en beskrivning av. F-konstituerade föremål tillhör således inte den del av universum som, om den finns, är den av alla upplevelser oberoende yttervärlden. (Personligen betvivlar jag lika lite som Segelberg dess existens, men jag kan inte *vederlägga* allt vad idealism heter.) Ej heller hör f-konstituerade föremål hemma i någon av upplevelser bestående del av universum. Om de är externa fenomen som – i likhet med t ex stolar, bord och tjurar och till skillnad från t ex pegaser, dryader och troll – korrekt kan sägas existera, så finns de i den vardagligt upplevda yttervärlden. Ett *x* är f-konstituerat som ett si-och-så-beskaffat, externt fenomen om och endast om följande villkor är uppfyllt:

(A) Det finns (i den metafysiska världen) en upplevelse *u* i vilken *x* i kraft av *u*:s intentionalitet *framträder som en si-och-så-beskaffad, extern företeelse,* t ex som ett granitblock, ett sileshår, en kråka, en man, ett hus, en spårvagn, en svensk hundralapp, en domare under ämbetsutövning, ett köpekontrakt, en militär order, ett påstående eller en äkta fråga.

Och *x finns* qua si-och-så-beskaffad, extern företeelse om och endast om detta gäller:

(B) *u* är inte en hallucination, illusion e d.

Analysen av villkor (B) aktualiserar problem, men dessa kan inte behandlas här, och de är dessutom ovidkommande i en diskussion om

sensationer. Ty om sensationer är f-konstituerade entiteter, så är det uppenbart poänglöst att för deras existens ställa upp ett villkor analogt med (B); *x* kan med rätta sägas existera som si-och-så-beskaffad sensation om och endast om ett villkor analogt med (A) är uppfyllt. Detta kan synas harmoniera med kunskapsteoretiskt motiverade fenomenalisters idé att vissa entiteter, vanligen kallade sinnesdata, har en epistemologiskt privilegierad status. Så är dock inte fallet. Väsentligt för fenomenalister av här aktuellt slag är ju att sinnesdata tillhör det jag kallar den metafysiska världen, men om antifenomenalismen är sann, så hör de inte alls hemma där. De hör hemma i en *f-konstituerad* värld – inte i den vardagligt upplevda *yttervärlden,* men väl i vardagsvärlden.

9. Är medvetenhet en kvalitet?

Om antifenomenalismen är sann, så är det helt klart att endast entiteter med intentionalitet förtjänar titeln *'upplevelse'.* Ställer sig saken på annat sätt om antifenomenalismen är falsk?

Vi kan tryggt ignorera varje jakande svar som är fotat på en fenomenalistisk teori vars företrädare, likt Bertrand Russell (1921 och 1926 s 141 ff respektive 83), förkastar distinktionen mellan *sensationer* och *upplevelser av sensationer.* Doktriner av detta slag implicerar att vi är mentala blindstyren: att våra "upplevelser" *samt och synnerligen* består av kantska förnimmelser utan begrepp vilka ju är "blinda"; och inte nog med detta: våra "blinda" förnimmelser är enligt ifrågavarande teorier också renons på *medvetenhet.* Aningen mer tänkvärd är Segelbergs idé om icke-intentionala upplevelser vilken återgavs i sista stycket av sektion 6.

Segelbergs version av tesen att medvetenhet är en kvalitet – låt oss kalla den *MK-S* – är viktig för hans hantering av frågan om *medvetandets enhet* och aktualiserar därför diverse med det problemet förknippade spörsmål. En del av dessa har jag diskuterat annorstädes (1974); här lämnas de åt sitt öde. *Ett* resultat av 1974-diskussionen bör dock nämnas. Segelberg verkar förutsätta att medvetenhet är en kvalitet för vilken gäller att alla moment av den (ev frånsett skillnader med avseende på intensitet) är exakt likadana. Den förutsättningen är inte hållbar: *om* medvetenhet är en kvalitet, så är den, i likhet med t ex rödhet, en determinabel med många determinater (se a a, s 222). Men om ett *flertal* medvetenhetsmoment ingår i en upplevelse, så kan man

med fog misstänka att de i själva verket är *intentionala* moment. *Ett* skäl för den misstanken är att idén om sensationer om vilka ingen är medveten, även om den i och för sig kan te sig egendomlig, *intuitivt* är avsevärt mindre absurd än motsvarande idé om *tankar*: om det funnes en tanke om vilken ingen vore medveten – vore den då inte själv *med nödvändighet* ett medvetande i miniatyr? (fr a a, s 223). Att begreppen medvetenhet och intentionalitet är intimt förknippade verkar alltså plausibelt. Men tesen att det skulle finnas intentionalitetslösa upplevelser à la Segelberg finns det, såvitt jag kan finna, inga som helst skäl att fästa tilltro till. Tvärt om: ifall m-moment inte är intentionala dito, varför är vi då inte mentala blindstyren – varelser som är renons på *information* om sina sensationers beskaffenheter, trots att de (vad *det* nu månde innebära) är medvetna om sensationerna? Härtill kommer att upplevelser kan vara renons på intentionalitet endast om antifenomenalismen är falsk.

Skall vi då konkludera att ingen kvalitet förtjänar titeln 'medvetenhet'? Nej.

I polemik mot *MK-S* framhåller Haglund (1984a, s 93) att vi, om vi betraktar de sätt på vilka orden 'medveten' och 'medvetenhet' "vanligen" används, så finner vi att vi har att göra med två "huvudtyper av fall". När vi säger om ett *x* att *x* är medvetet menar vi i det ena fallet att *x* är en person som är medveten om något; och i det andra fallet menar vi att *x* är föremål för en persons medvetenhet; och i "båda dessa huvudfall tycks det [...] vara så, att om ordet "medveten" står för ett attribut, så är detta en relationsegenskap och inte en egentlig egenskap".

Dessa haglundska iakttagelser torde framstå som helt okontroversiella, men följande två kommentarer till dem är på sin plats:

(1) De diskrediterar på intet sätt tesen att *varje upplevelse har en kvalitet som med fog kan kallas medvetenhet.* I fall där vi om en person säger att hon är medveten om *x förutsätter* vi normalt att *x* finns. *Om* vi gör denna förutsättning – ja, *då* är det uppenbart att vad vi avser att säga är att hon har en viss relation till *x*. Men hur ställer sig saken i fall där vi inte gör denna förutsättning?

(2) Haglund menar, såvitt jag kan förstå och i så fall med rätta, att den som är medveten om något har intention på detta något. Hans iakttagelser styrker därför tesen *att begreppen medvetenhet och intentionalitet är intimt förknippade.*

Mot denna bakgrund torde det te sig naturligt att jag gärna vill slå ett slag för följande tes:

MK. Varje upplevelse är intentional, och en kvalitet värd benämningen 'medvetenhet' tillkommer en entitet om och endast om den är intentional.

Till förmån för *MK* kan följande anföras.

Medveten om något kan jag inte vara annat än i viss grad. *En* sorts gradskillnad har med mina upplevelsers *klarhet* att göra. Surret från humlor som jag hör då jag dåsar på en sommaräng kan jag mer eller mindre klart identifiera som humlesurr eller som någonting helt annat (eller vara mer eller mindre klart medveten om att jag ingen aning har om vad för slags ljud det kan vara). En *p*-tanke, där *p* är en proposition, kan jag tänka mer eller mindre klart. Osv.

Det finns emellertid också *en annan* sorts gradskillnad: medveten om något kan jag inte vara annat än med en viss grad av *intensitet.* Höggradig klarhet är visserligen normalt korrelerad med höggradig intensitet. Om jag t ex klart identifierar humlesurret som humlesurr, så *uppmärksammar* jag det, och ett uppmärksammande av *x* involverar en medvetenhet om *x* som är intensivare än den medvetenhet om *x* som föreligger då *x* ceteris paribus varseblivs utan att bli uppmärksammat. Men intensitet och klarhet är icke desto mindre skilda ting: om en varseblivning i vilken ett ljud uppmärksammas som humlesurr kan vi meningsfullt säga att den *klart* är en humlesurrsvarseblivning, men är den *intensivt* en sådan? Näppeligen. En (relativt) intensiv upplevelse är den, och den medvetenhet med avseende på vilken den är intensiv är en medvetenhet om humlesurr, men *qua humlesurrsupplevelse* är den inte intensiv utan klar.

Variationer i medvetenhetsintensitet får alltså inte förväxlas med variationer i klarhet hos upplevelser. Ej heller får de förväxlas med variationer i intensitet hos kvaliteter som tillkommer upplevda sensationer eller externa objekt. Även när jag är medveten om något som inte *kan* graderas med avseende på intensitet, t ex *strukturen* hos ett föremål, är min medvetenhet mer eller mindre intensiv. Vidare är intensiteten hos en upplevd färg *uppenbart* någonting annat än intensiteten hos min medvetenhet om färgen, och samma sak gäller mutatis mutandis för lukt- och smaksensationer, för taktila dito och för fenomenella ljud. *Att en viss från allt vad sensationsintensitet heter skild grad av medvetenhetsintensitet tillkommer varje upplevelse* synes mig vara en plausibel tes.

"Men smärtupplevelser då?", spörjer kanske någon. Såvitt jag finna är frågan problematisk för min tes endast om alla smärtor förutsätts vara *svåra.* När det gäller lindriga smärtupplevelser synes det mig nästan lika uppenbart som när det t ex gäller färgupplevelser att medvetenhetsintensitet och sensationsdito är skilda ting, och därför är det enligt min mening föga troligt att distinktionen mellan medventenhets- och sensationsintensitet inte ens *i princip* skulle vara tillämplig på svåra smärtor. Distinktionen är *uppenbart* tillämplig på *flertalet* sorters upplevelser. Att det i ett relativt fåtal typer av fall, t ex när det gäller glädje-, olust- eller *svåra* smärtupplevelser, *i praktiken* är vanskligt eller rent av omöjligt att skilja medvetenhetsintensitet från sensationsdito är näppeligen en tungt vägande invändning.

MK torde nu framstå som plausibel. Om varje upplevelse i högre eller lägre grad ger prov på någonting som förtjänar benämningen 'medvetenhetsintensitet', så kan ingen upplevelse frånkännas en kvalitet värd namnet 'medvetenhet'.

10. Medvetenhet och intentionalitet (I)

Men *hur* är medvetenhet och intentionalitet relaterade till varandra? För den frågans hantering är det lämpligt att införa en terminologi för en distinktion som redan figurerat på ett väsentligt sätt.

Trots att de är exakt likadana i övrigt, kan somliga kvalitetsmoment uppvisa olikheter med avseende på *intensitet* (jfr *ZP,* kap 4). För kvaliteter vars lägsta determinater kan identifieras med kongruensklasser av sådana moment inför jag härmed termen *'i-kvalitet',* och *'x-kvalitet'* får bli beteckningen på kvaliteter som inte är i-dito. Begreppet i-kvalitet exemplifieras bl a av färger och av ljud-, lukt-, smak- och smärtkvaliteter. Bland x-kvaliteterna finner vi kvadratiskhet, elliptiskhet m fl formkvaliteter samt utsträckning och strukturkvaliteter av varierande slag.

Att medvetenhetskvaliteten tillhör i-kategorin är ju uppenbart. Hur förhåller det sig med intentionalitet?

Att såväl *aktmateria* som *aktkvalitet* alltid är involverad i en upplevelses intentionalitet är en husserlsk tes (se 1913, § 20). Segelberg godtar den inte (mer om detta nedan), men för mig är den evident. Innan jag vidareutvecklar den är dock några terminologiska anmärkningar på sin plats.

'Aktmateria' och 'aktkvalitet' kan brukas som benämningar på

såväl moment som kvaliteter. Båda användningarna av de två termerna finns hos Husserl, men den förra dominerar. I det följande är ordet 'aktmateria' förbehållet moment; *'aktmaterialitet'* är min paraplyterm för kvaliteter som upplevelser har i kraft av sina aktmaterier. Ordet 'aktkvalitet' reserverar jag naturligt nog – för kvaliteter; moment av aktkvaliteter betecknar jag alltid som moment.

Varje upplevelses intentionalitet tillkommer den i kraft av att den innehåller en aktmateria och ett aktkvalitetsmoment. Genom att ange en varseblivnings s k intentionala objekt fångar man i bästa fall bara dess aktmaterialitet. En fullständig karaktäristik av meningen hos en varseblivning måste också ange aktkvaliteten hos den sistnämnda. Den måste t ex besvara frågan om varseblivningen är en s k *sättande* akt eller ej, dvs om varseblivningen involverar *tro* på det upplevda objektets existens eller ej.

Att en fullständig karaktäristik av en upplevelses intentionalitet måste ange en aktkvalitet är ännu mer uppenbart när det gäller upplevelser som utan avvikelse från normalt, svenskt språkbruk kan kallas *tankar*. För flertalet sådana gäller att man hjälpligt kan fånga deras aktmaterialiteter genom att ange de *propositioner* som tänks i dem – genom att karaktärisera dem som *Att guld är en metall*-tankar, *Att finkel är mer hälsovådligt än etylalkohol*-tankar, *Att tjurar ibland är folkilskna*-tankar etc. Men propositioner kan tänkas på skilda sätt. Om *p = Att Adam slutar röka från och med nyårsdagen 1996* och om Adams hustru under hösten 1995 är engagerad i en kampanj för att få Adam att fimpa för gott från och med nyår, så kan vi lätt föreställa oss att hon vid skilda tillfällen före nyåret tänker *p*-tankar med åtminstone fyra klart skilda aktkvaliteter: *tro, undran, önskan* och *vilja*.

Att en upplevelses aktmaterialitet är en *x-kvalitet* och därmed inte en determinat av medvetenhet torde stå klart från den ovan förda diskussionen om klarhet och upplevelseintensitet. Men de fyra aktkvaliteter som nyss nämnts är inte x-kvaliteter. Såväl tro som undran, önskan och vilja kan ju variera med avseende på intensitet. Kunde då inte dessa aktkvaliteter klassas som determinater till determinabeln medvetenhet? Kunde inte moment av dem sägas *vara* medvetenhetsmoment?

Inte gärna. Om man i en tanke med viss grad av tro omfattar en proposition, är då *trons styrka* proportionell mot *intensiteten hos den medvetenhet* om propositionen som tanken involverar? Inte såvitt jag kan finna, och samma sak gäller mutatis mutandis för en undring, en

önskning och en vilja. Att moment av tro, undran etc. skulle kunna identifieras med moment av medvetenhet torde därför vara en förfelad idé.

En komplikation i sammanhanget är emellertid att aktkvaliteter kan vara såväl neutrala som icke-neutrala. Innan jag tar upp den vill jag dock gärna försöka skingra diverse dubier som en eller annan läsare kan antas hysa beträffande distinktionen mellan aktkvalitetsmoment och den sorts sensationsmoment som är typiska för *känslor.*

11. Exkurs om distinktionen aktkvalitet/känsla

Enligt min mening är åtminstone flertalet upplevelser som förtjänar titeln *'känsla'* varseblivningar eller tankar i vilka det ingår hyletiska moment som gör dem *emotionellt färgade.* Här begränsar jag mig till känslor som är emotionellt färgade *tankar.*

För somliga aktkvaliteter, t ex tro och undran, gäller att tankar som har dem utan tvekan kan vara *emotionellt ofärgade.* Andra aktkvaliteter ter sig kanske mer tvivelaktiga detta hänseende. Kan t ex en *önskning* vara emotionellt ofärgad? Eller en av stark övertygelse präglad *evaluativ* eller *normativ* tanke? Till dessa spörsmål återkommer jag inom kort.

Låt oss först koncentrera oss på tankar vars aktkvalitet är den sorts övertygelse med vilken propositioner omfattas med tro. En tro på sanningen hos en proposition *p* är ofta emotionellt ofärgad, men den kan också vara emotionellt färgad på mångahanda sätt: du kan t ex vredgat, sorgset eller euforiskt tro att *p,* och din tro att *p* kan också involvera leda, rädsla, skam etc.

Nu kommer vi till frågan om *önskningar* kan vara emotionellt neutrala. Om emotionellt neutrala önskningar framstår som dubiösa entiteter, så torde detta bero på lingvistiska intuitioner enligt vilka orden 'känsla' och 'emotion' är applicerbara på *varje* önskning. I detta sammanhang är det dock lämpligt att sätta dessa intuitioner ur spel. För tankar som i min terminologi har *samma aktkvalitet* gäller att de *radikalt* kan skilja sig från varandra i hänseenden som enligt normalt språkbruk är *tveklöst* emotionella eller känslomässiga. Något motsvarande gäller *inte* för tankar som i min terminologi har *samma emotionella färgning.* Beträffande önskningar kan vi nu konstatera att de kan vara emotionellt färgade på radikalt skilda sätt: man kan önska något *glädjefullt* eller *sorgset* eller *vredgat* etc. Vidare kan man önska något

utan att vara vare sig *glad* eller *sorgsen* eller *vred* e d; och de önskningar som man i så fall hyser torde med fog kunna sägas vara emotionellt neutrala.

Att det i föregående stycke sagda mutatis mutandis gäller för tankar i vilka normativa eller evaluativa teser omfattas med övertygelse torde vara uppenbart.

12. Medvetenhet och intentionalitet (II)

De fyra aktkvaliteter som hittills figurerat är alla *icke-neutrala*: i en tänjd men ändå rimlig bemärkelse är *ställningstaganden* involverade i moment av dem; dessa moment är *icke* "ohne jedwede 'Stellungnahme'" (Husserl 1913, s 457). *Neutrala* aktkvaliteter exemplifieras av *rena förståelser,* d ä den sorts upplevelser i vilka yttranden *enbart* blir förstådda. Man kan adekvat förstå ett påstående utan att själv tro (eller betvivla) det sagda; man kan adekvat förstå en äkta fråga utan att själv undra osv. Rena förståelser har neutrala aktkvaliteter av skilda slag beroende på om de förstådda yttrandena är påståenden, äkta frågor, optativer eller uppmaningar. Låt oss säga att aktkvaliteten hos en ren förståelse av ett påstående är *P-neutral* och att de aktkvaliteter som exemplifieras av rena förståelser av äkta frågor, optativa yttranden och uppmaningar är *?-*, *O-* resp. *!*-neutrala.

Hur skall vi nu ställa oss till frågan om neutrala aktkvaliteter kan klassas som determinater av medvetenhet? Det torde ligga i öppen dag att ett ogarderat och jakande svar kan försvaras endast om såväl *P-* som *?-*, *O-* och *!*-neutralitet är en *i-kvalitet.* Är detta villkor satisfierat? Näppeligen. Det är *icke* fallet att en tanke mer eller mindre *intensivt* är en ren och adekvat förståelse av ett påstående, en äkta fråga, ett optativt yttrande eller en uppmaning. Mer eller mindre intensiv med avseende på medvetenhet är den, men qua *P-*, *?-*, *O-* eller *!*-neutral tanke är den mer eller mindre *klar,* inte mer eller mindre intensiv.

P-neutralitet är alltså en *x-kvalitet*; samma sak gäller för såväl *?-* som *O-* och *!*-neutralitet, och att det skulle förhålla sig på annat sätt med någon annan neutral aktkvalitet finns det, såvitt jag kan se, ingen som helst anledning att tro.

I sektion 9 kom jag ju fram till att en kvalitet värd benämningen 'medvetenhet' tillkommer en entitet om och endast om den är intentional. Men om en upplevelses intentionalitet består av dess aktmaterialitet och aktkvalitet, så är medvetenhet och intentionalitet klart skilda

ting. Förutsatt att upplevelser med *icke-neutrala* aktkvaliteter lämnas ur räkningen torde dock det ovan sagda ge visst stöd för följande hypotes:

Det är inte fallet att medvetenhet, som Segelberg tycks mena, är en kvalitet för vilken gäller att varje moment av den (frånsett dess grad av intensitet) är exakt likt varje annat medvetenhetsmoment. I likhet med t ex rödhet är medvetenhet en determinabel med många determinater, och sin intentionalitet har en upplevelse i kraft av att ha en strukturkvalitet som i sin tur *emergerar från* ett *renodlat medvetenhetskomplex,* d ä ett komplex som är sådant, att varje i-kvalitetsmoment i det är ett medvetenhetsmoment. Den bemärkelse i vilken strukturkvaliteten emergerar från medvetenhetskomplexet torde vara analog med den i vilken t ex en melodikvalitet emergerar från en tonsekvens. I upplevelser ingående medvetenhetsmoment av skilda slag bildar strukturer som manifesterar sig i att *innebörder* tillkommer upplevelser – innebörder som är klara om medvetenhetsmomenten är intensiva och oklara om intensiteten hos tillräckligt många involverade moment av medvetenhet är låg.

Denna hypotes om relationerna mellan medvetenhet och intentionalitet döper jag härmed till *den enkla emergenshypotesen.* Hurpass väl står den sig om upplevelser med icke-neutrala aktkvaliteter tas med i beräkningen?

Det finns en uppenbar bemärkelse i vilken en *motsvarighetsrelation* råder mellan moment av *P*-neutralitet och moment av *tro.* Lika uppenbart är det att moment av såväl *?*- som *O*- och *!*-neutralitet är analogt relaterade till moment av *undran, önskan* resp. *vilja.* Hur förhåller det sig med de egenskaper – låt oss kalla dem *motsvarighetskvaliteter* – med avseende på vilka *P*-, *?*, *O*- och *!*-neutrala tankar *liknar* tankar som orden 'tro', 'undran', 'önskan' resp 'vilja' är applicerbara på? Är moment av motsvarighetskvaliteter alltid eller i viss utsträckning *exakt* likadana som neutrala aktkvalitetsmoment? Kanske, och i så fall kan alla eller somliga motsvarighetskvaliteter sägas *vara* neutrala aktkvaliteter. (Ingen terminologisk stipulation i det föregående hindrar att ett neutralt aktkvalitetsmoment säges *ingå* i ett icke-neutralt dito.) Härmed må det emellertid förhålla sig hur som helst. Det i sammanhanget viktiga är att även motsvarighetskvaliteterna tillhör x-kategorin; att de verkligen hör hemma där synes det inte finnas några skäl att betvivla. Ett icke-neutralt aktkvalitetsmoment kan sägas innehålla ett

moment av en motsvarighetskvalitet och ett *avneutraliserande moment,* och de skillnader med avseende på intensitet som icke-neutrala aktkvalitetsmoment av i övrigt samma slag kan uppvisa torde bero på deras avneutraliserande moment.

Men nu inställer sig frågan vad dessa avneutraliserande moment har för status. Emergerar *de* från renodlade medvetenhetskomplex? Mot ett jakande svar talar iakttagelsen att intensiteter hos moment av tro, önskan, vilja etc. inte generellt är proportionella mot intensiteter hos medvetenhetsmoment. Det kan mycket väl tänkas att *icke-neutraliteten* hos en upplevelse med icke-neutral aktkvalitet tillkommer upplevelsen i fråga på grund av att *hyletiska* moment ingår. Segelberg hävdar att "det kan [...] vara lämpligt att vidga omfånget för begreppet hyletiskt innehåll, så att aktkvaliteterna inkluderas i denna klass" (*MJ,* s 74). Att denna idé i sin segelbergska version kunnat attrahera någon förbluffar mig; att hänföra *allt* vad aktkvalitetsmoment heter (t ex moment av *P*- eller *?*-neutralitet) till kategorin sensationsmoment är *uppenbart* bort i tok. Om hypotesen revideras till att enbart gälla avneutraliserande moment, så ter den sig mer rimlig – såpass plausibel att jag från och med nu låter den vara en av premisserna för mitt resonemang.

Detta beslut genererar emellertid ett dilemma: antingen måste begreppet intentionalitet revideras så, att avneutraliserande moment klassas som icke-intentionala eller också måste intentionaliteten hos en upplevelse med icke-neutral aktkvalitet antas emergera från komplex i vilka såväl medvetenhetsmoment som hyletiska dito ingår. Det förra alternativet må synas motbjudande, men om antifenomenalismen vore falsk, borde vi nog ändå föredra det och därmed acceptera en modifierad variant av Segelbergs åsikt om aktkvaliteter. Saken kommer emellertid i ett annat läge om antifenomenalismen approximerar sanningen. Givet antifenomenalismen är *alla* sensationsmoment – om vi även i varseblivningar upplevda moment av färg, form, ljud, smak, lukt etc – i viss mening intentionala. Att det t ex för en visuell varseblivning av en svart tjur gäller att den är en upplevelse av något *svart* beror inte på att den bokstavligen innehåller ett moment av svärta; egenskapen svart och egenskapen att vara tjur är i lika mån förbehållna föremål som är f-konstituerade. Om vi – vilket jag för egen del beslutar – förutsätter antifenomenalismen, så bör vi rimligtvis välja dilemmats andra horn och anamma hypotesen att vissa determinater av

intentionalitet emergerar från komplex i vilka såväl medvetenhetsmoment som hyletiska dito ingår. Därmed faller den enkla emergenshypotesen. I dess ställe måste vi, om vi inte radikalt vill överge ovan skisserade tankebanor, gå in för en *komplex emergenshypotes.* Men hur ser det intentionalitetsbegrepp ut som vi då erhåller?

En distinktion måste dras mellan *f-konstituerad* och – låt oss säga – *metafysisk* intentionalitet. Intentionalitet = meningsfullhet är f-konstituerad. Det finns, såvitt jag kan finna, ingen anledning att tro att man någonsin genom att specificera en upplevelses *mening* anger den egentliga beskaffenheten hos något av de kvalitetsmoment som bokstavligen är ett innehåll i upplevelsen. Men att upplevelsen har en viss mening måste rimligtvis antas bero på att vissa icke-hyletiska kvalitetsmoment bokstavligen ingår i den. Den egentliga beskaffenheten hos dessa moment avslöjar sig inte annat än i begränsad mån för oss i våra upplevelser: med de medvetenhetsmoment som ingår i våra upplevelser tycks vi ha direktkontakt i så måtto som att deras intensitetsgrader men inte deras beskaffenheter i övrigt avslöjar sig. Vi har här att göra med moment av metafysisk intentionalitet.

Med avseende på metafysisk intentionalitet kan en distinktion dras mellan *emergent* och – låt oss säga – *basal* intentionalitet. Till den senare kategorin hör moment av kvaliteter som uppfyller följande villkor:

(1) Det är icke fallet att de emergerar från komplex av andra kvalitetsmoment.

(2) De tillhör en klass *K* för vilken detta gäller: om en entitet *x* fallerar att innehålla ett moment tillhörande *K,* så är *x* renons på f-konstituerad intentionalitet.

Även för en anhängare av den komplexa emergenshypotesen torde det te sig naturligt att identifiera basal intentionalitet med unionen av alla kongruensklasser av medvetenhet, men hans teori implicerar att moment av emergent intentionalitet kan vara mer eller mindre *hyletiskt influerade.* Ett moment *mi* av emergent intentionalitet är hyletiskt oinfluerat om och endast om det emergerar från ett renodlat medvetenhetskomplex. Om det komplex från vilket *mi* emergerar i någon mån innehåller hyletiska moment, så är *mi* i motsvarande mån hyletiskt influerat, och samma sak gäller den f-konstituerade intentionalitet som tillkommer en upplevelse i vilken *mi* ingår.

Not:
Till Mats Furberg vill jag rikta ett varmt tack för värdefulla synpunkter och råd.

Litteratur

Följande förkortningar används: *'GUF'* = 'Göteborgs universitet, Institutionen för filosofi'; *'FH'* = 'Malmgren, H (ed): (1974): *Filosofiskt herbarium tillägnat Ivar Segelberg på hans 60-årsdag. GUF,* Göteborg'; och *'SKM'* = *'Språk, kunskap, medvetande. Festskrift tillägnad Ivar Segelberg på hans 70-årsdag, GUF,* Göteborg 1984'.)

HAGLUND, D (1984): *Phenomenological Studies I. On the Nature of Hyletic Data* (stencil). Filosofiska meddelanden, Röda serien nr 24, *GUF*, Göteborg.

(1984a): 'Ivar Segelberg och frågan om medvetandets enhet.' I *SKM.*

HUSSERL, E (1913): *Logische Untersuchungen* II/1, 2a uppl Halle a d S (Nyutg från Max Niemeyer Verlag, Tübingen 1968).

KANT, I (1781): *Kritik der reinen Vernunft.* 2a uppl. 1787.

RUSSELL, B (1921): *The Analysis of Mind.* London.

(1926): *Our Knowledge of the External World.* Omarb. utg., London.

SEGELBERG, I (1945): *Zenons paradoxer. En fenomenologisk studie* (ZP) Stockholm.

(1947): *Begreppet Egenskap. Några Synpunkter* (BE). Stockholm.

(1953): *Studier över medvetandet och jagidén* (MJ). Stockholm.

WEDBERG, A (1966): *Filosofins historia. Från Bolzano till Wittgenstein.* Stockholm.

WETTERSTRÖM, T (1974): 'Consciousness from a Quality Instance Point of View'. I *FH.*

(1977): *Intention and Communication. An Essay in the Phenomenology of Language.* Lund.

PAUL NEEDHAM

Duhems quineska realism

1. Inledning

Med få undantag (såsom Maiocchi 1990) brukar Duhem beskrivas som antirealist eller instrumentalist – en beskrivning som motiveras dels genom hänvisning till hans välkända angrepp på atomistiska teorier, dels genom hans påstådda konventionalistiska uppfattning. Popper, som inte skiljer mellan Duhems och Poincarés uppfattningar (1972, s 78, fn 1), har varit en inflytelserik förespråkare för den senare tolkningen. Denna tolkning återfinns inte bara hos dem som hör till Poppers skola; även idéhistoriker som yttrar sig om Duhems hållning beskriver honom ofta som konventionalist (t ex Dolby 1984, ss 388, 389). Nyligen har en författare gått så långt att hon (Christie 1994) citerar Duhems formulering av den konventionalistiska tesen – vilken Duhem sedan invänder mot utifrån sin allmänna syn – men tolkar den som en formulering av hans egen uppfattning! Alltså, Duhem argumenterar mot konventionalismen och det argument han ger illustrerar väl hans hållning. Detta argument tas upp i avsnitt 3 som en central tråd i den tolkning av Duhems realism som här skisseras.

Jag tänker inte fördjupa mig i alla förekommande versioner av realism och antirealism för att sedan utreda Duhems förhållande till dessa, utan håller mig till några från litteraturen välkända uttolkningar av Duhem som antirealist. En sådan ryms i en inställning till den s k Duhem-Quine tesen (jfr också Ariew 1984) som Vuillemin uttrycker i följande termer:

> Mot Duhems neutrala naturvetenskap renad från all metafysik ställde Quine en realistisk naturvetenskap som förvandlas till en metafysik i kraft av sina ontologiska åtaganden (*commitments*). (Vuillemin 1986, s 609)

Den uppfattning av Quine som här antyds ger ett mått på realism med vilket Duhems hållning kan jämföras. Det kan tilläggas att Quine själv tycks acceptera Vuillemins Duhem-tolkning, och i sitt svar till Vuillemin i Schilpp-volymen förklarar han att

> ... det finns en skillnad mellan Duhems fiktionalistiska inställning gentemot fysik och min realistiska inställning. Den beror på min naturalism, som inte erkänner någon högre sanning än den vi söker i vårt sammansatta naturvetenskapliga världssystem. (Quine 1986, s 619)

Senare i samma svar skiljer Quine mellan det som är tillräckligt för att en sats i fysiken är sann och det som räcker för att han skall 'kanske felaktigt, erkänna ett fysikaliskt påstående som sant' (s 622), där han betonar sin immanenta sanningsuppfattning, som (s 316) sägs vara inbyggd i hans naturalism.

Det förefaller mig, å andra sidan, att Duhems hållning har mycket gemensamt med Quines naturalism. Naturligtvis finns den skillnaden att Duhem bygger på sina väldiga kunskaper om såväl naturvetenskapen som dess historia, och att han utvecklar sin filosofiska hållning, dels sida vid sida med sitt eget program att ena fysik och kemi inom en generaliserad termodynamik, dels i ljuset av insikter hos Aristoteles som bibehållits av skolastikerna och som Duhem menar överlevt 1600-talets s k vetenskapliga revolution. Kort sagt, hans vetenskapliga gärning ger oss ett typfall av Quines dröm om en kontinuitet mellan vetenskap och filosofi om en man 'som bedriver filosofi ... som en del av sitt världssystem, i kontinuitet med vetenskapen i övrigt' (Quine 1986, s 430), vägledd av det som enligt hans vetenskapliga omdöme kunde försanthållas. Quine själv däremot tycks inte ha någon detaljerad kunskap om något naturvetenskapligt ämne, utan argumenterar utifrån ett filosofiskt perspektiv där naturvetenskaplig kunskap antas vara central. Naturvetenskapen, eller fysiken, är för Quine den disciplin som har till uppgift 'att säga vilken ... minimal katalog av tillstånd som vore tillräcklig för att berättiga oss att säga att det inte finns någon förändring utan en förändring i läge eller tillstånd' (1986, s 430), men Quine gör inte anspråk på att kunna uttala sig om vad det är den egentligen säger. Det kan dock noteras att Quine inte bygger in atomism av något slag i en karakterisering av realismen.

Fastän angrepp mot atomismen ofta betraktas som ett kännetecken

på en antirealistisk hållning, kommer inte Duhems argument mot atomismen att utförligt behandlas här. Diskussionen fokuseras i stället kring det han anför mot konventionalismen, vilket med fördel kan ses i ljuset av Quines immanenta sanningsuppfattning som fordrar att vi försanthåller de påståenden som åberopas när vi tar till oss erfarenhet. Men Duhems argument mot konventionalismen berör förhållandet mellan matematiska och fysikaliska påståenden, och därmed uppmärksammas en fråga som Quine inte behandlar särskilt ingående. I motsats till vad vissa andra filosofer anser (t ex Papineau 1993), medför inte naturalismen för Quine att det som sägs om s k abstrakta entiteter – grovt uttryckt, entiteter som inte berörs av fysikens lagar – måste betraktas som en fiktion. Quine är ingen anhängare av fiktionalism inom matematiken, för han anser att matematiska påståenden ingår i den systematiska redogörelse för fysiken som åberopas i vår förståelse av vår erfarenhet. Men när han talar om vad han kallar för fysikaliska påståenden, så syftar han antagligen inte på rent matematiska påståenden, som t ex hur en viss differentialekvation kan lösas. Duhem kan knappast heller tolkas som matematisk fiktionalist, men han menade att fysikaliska påståenden kan skiljas från rent matematiska påståenden genom det sätt på vilket man begränsar sig till vad man är beredd att hålla för sant. Explicit uttalad hänsyn till felmarginaler indikerar vilket fysikaliskt påstående ett matematiskt påstående kan användas för att uttrycka.

För att tydliggöra hur frågan uppkommer om vilka de satser som Quine kallar fysikaliska påståenden egentligen är och vari Duhems insats består, börjar jag med en presentation av vissa aspekter av Quines realistiska syn i samband med hans utveckling av den s k Duhem-Quine tesen. Vad Duhem har att säga om vad det är som försanthålles försöker jag förklara i avsnitt 3.

2. Quines realism

Om direkt observation inte kan avgöra om en hypotes är sann eller falsk, måste den testas genom sina konsekvenser. Från denna utgångspunkt utvecklade Duhem, som bekant, sitt argument mot möjligheten av ett avgörande experiment. Den centrala frågan är hur man redogör för en implikation som innebär att en sats är en konsekvens av en hypotes. Den innebörd man lägger i en hypotes bestäms av de konsekvenser man tänker sig hypotesen medför. I allmänhet följer inte en

sats enbart från en hypotes utan hjälp av andra hypoteser. Det kan hända att detta inte framgår av vardaglig praxis, utan först när man försöker grundligt och rigoröst redogöra för alla stegen i resonemanget, då 'dolda' eller ditsatta hypoteser (ditsatta för att täppa en lucka) dyker upp. Likaså kan det förefalla oss att en gången tids redogörelse är ofullständig, som när Hilbert fyllde ut Euklides' axiomatisering av geometri. Vilka slags tilläggshypoteser Duhem betonar kommer att diskuteras i nästa avsnitt.

Quine håller med om att observationer inte kan implicera hypoteser; han hänvisade till Duhem i 1953 års omtryckning av 'Two Dogmas', när likheterna med Duhems tankegång påpekades för honom. Men Quine pekade i första hand på den allmänna omständigheten att sidoförhållanden åberopas snarare än på de detaljer som Duhem redovisade. Quine uppmärksammade det holistiska drag denna omständighet medför, samt att de tilläggsantaganden som behövs för en fullständig redogörelse för en typisk slutsatsdragning från en hypotes typiskt omfattar antaganden traditionellt betraktade som oantastliga – s k *a priori* sanningar – omfattande t o m de logiska grundprinciperna. Under senare år har Quine modifierat och begränsat sin holism, infört idealen av observationssatser vilka holofrastiskt betraktade har en egen innebörd, och medgivit att en given hypotes inte involverar alla andra försanthållna satser utan berör en begränsad omgivning. Men hans syn på ontologi förblir oförändrad. Existensantaganden spelar en roll då redogörelsen av logiska länkar mellan satser kräver en regimentering som överstiger den satslogiska nivån. Satser antas då innehålla gemensamma predikat som är sanna för entiteter av olika slag. Att acceptera en slutsatsdragning innebär då att allt som har lett till slutsatsen accepteras, vilket innebär vad han kallar för *ontological commitment* gentemot de entiteter till vilka referens görs för att upprätthålla godtagna logiska länkar. Denna doktrin står i motsats till 'filosofiskt hyckleri, vilket ej vill kännas vid en ontologi men ändå utnyttjar dess fördelar, [och] gynnas av vardagligsspråkets nyckfullhet' (Quine 1960, s 242) – en doktrin Putnam har beskrivit som 'ett grundkrav för intellektuell ärlighet'.

Men ärlig eller inte så har denna hållning gett upphov till en spänning i samband med svårigheterna att ge en bokstavlig fysikalisk tolkning av teorier formulerade i matematiska termer. När Koyré (1965) talade om Newtons syntes, menade han att Newton lyckades

med konststycket att visa hur matematiken kan användas för att beskriva naturen. Man kunde då upplösa den traditionella åtskillnaden mellan matematiska och sanna hypoteser vilken Bellarmine åberopade när han tillrättavisade Galileo som höll den Kopernikanska hypotesen för sann. Dock dröjde det inte länge innan problem uppdagades. Berkeleys kritik av Newtons infinitesimalkalkyl var inte begränsad till föga smickrande beskrivningar av de nyss införda entiteterna som 'andar av bortgångna kvantiteter'. Han pekade på skriande motsägelser i Newtons argumentationssätt. Och de tvister hans kritik gav upphov till ledde till ett klarläggande av kalkylens centrala idéer, då Bolzano, Cauchy, Weierstrass och Riemann under 1800-talet utvecklade acceptabla definitioner av derivator och integraler, baserade på en noggrann axiomatisering av de reella talen. Det framgick bl a att tätheten hos de rationella talen – att mellan varje två ligger ett tredje – inte räcker, utan ännu flera tal måste packas in i ordningen för att gränsvärdet av varje (konvergerande) Cauchy-sekvens ska kunna existera. Men skillnader mellan sådana oändligheter som krävs för att begrepp i infinitesimalkalkylen skall vara väldefinierade, är svåra att uppfatta i fysikaliska termer. Ska de matematiska distinktioner som krävs för matematisk kontinuitet verkligen tas på allvar som distinktioner som återspeglar motsvarande distinktioner i naturen? Den traditionella åtskillnaden spökar fortfarande i bakgrunden. Men Quinesk ärlighet föreskriver att man inte kan blankt förneka principer man använder hela tiden.

Ett sätt att reagera på vad som närmast tycks vara ett dilemma är att försöka visa att man egentligen inte behöver använda matematiska teorier om kontinuitet och de entiteter som därmed krävs. Hartry Field har i sin bok *Science Without Numbers* utvecklat en sådan uppfattning, som han ömsom kallar för nominalism, ömsom matematisk fiktionalism. Field gör anspråk på att presentera 'en alternativ formulering av naturvetenskapen som inte förutsätter någon del av matematiken som refererar till eller kvantifierar över abstrakta entiteter' (s 2). Han tillägger att vad som menas med 'abstrakt' 'kanske inte är fullständigt klart, men ... [omfattar] sådana förmenta entiteter som tal, funktioner och mängder' (s 1). Trots sina fyndiga teoribildningar är dock Fields program allmänt betraktat som misslyckat p g a argument som Shapiro (1983) i ljuset av Gödels ofullständighetssats gett mot Fields konservatism – dvs dennes önskan att bibehålla en fysikalisk teori med samma kraft som den ursprungliga. Det tycks som om den Quineska

poängen måste vara den att matematiken är oundviklig, eftersom den inte kan urskiljas från en rent fysikalisk del av teorin. Om man vill upprätthålla de logiska samband som innebörden av olika fysikaliska påståenden tycks bero på, så föreligger det samtidigt ett åtagande (*commitment*) gentemot mycket annat som inte lämpar sig för fysikalisk tolkning. I avsaknad av en alternativ formulering som inte representerar naturen med dessa matematiska inslag måste vi erkänna att vi försanthåller en teori som faktiskt representerar naturen på detta sätt.

Field tycks ha förstått Quines dilemma som en utmaning att visa att man kan undvara abstrakta entiteter. Men problemet med en fysikalisk tolkning av kontinuitet behöver inte uppfattas på detta sätt (även om den traditionella distinktionen bygger på en förutsättning att det föreligger en skarp åtskillnad mellan abstrakta och s k konkreta entiteter). I synnerhet kan inte Quine själv ha förstått sina krav på intellektuell ärlighet på detta sätt. För Quine avgränsas ingen speciell ontologisk kategori med beskrivningen 'abstrakt'. Det finns endast ett existensbegrepp, såsom Russell vidhöll gentemot Meinong. Denna hållning innebär att man inte utan vidare prövning kan acceptera huruvida det som tycks ha den grammatiska formen av ett substantiv verkligen ska behandlas som ett substantiv. Ontologin framgår av regimentering, inte av ytlig grammatisk struktur.

Det skulle kunna tänkas att en nominalist i Fields anda trots detta logiska argument, anser att hans elimineringsstrategi till viss del kan motiveras med utgångspunkt i en klar åtskillnad mellan föremål som hör till den fysikaliska ontologin och de – de s k abstrakta föremålen – som inte gör det. Möjligen kan det tänkas att materia såsom den uppfattades före Newtons tid lämpade sig för kontraster mellan kropp och själ, materia och tal, osv, vilka begrepp kanhända var informativa då. Men nu när man är van vid att tala om accelerationsmotstånd, attraktiv förmåga, magnetiskt fält, elektronmoln, osv, kan de knappast uppfattas som upplysande. Uppenbarligen är inte Quine påverkad av sådana motiveringar och den nominalistiska tesen kan lämnas åt sidan.

Men frågan om fysikalisk tolkning är inte därmed avgjord och bör skiljas från frågan om matematikens abstrakta karaktär, vad nu det kan innebära. Att det ändå finns ett problem med fysikalisk tolkning framgår av exempel som det följande. I sin utveckling av postulaten för termodynamik säger Callen (1985) att det finns en funktion S, kallad entropin, som är en funktion av de s k extensiva variablerna, kallade

energin, volymen, och moltalen hos ett kemiskt sammansatt system:

$S = S(U, V, N_1, \ldots, N_r)$.

Callen tillskriver funktionen de matematiska egenskaperna kontinuitet och differentierbarhet i sitt Postulat III. Tidigare i boken ger han en förklaring av hur de oberoende extensiva variablerna skall tolkas. I denna diskussion förekommer följande passage:

> ... ett enkelt system har en bestämd *kemisk sammansättning* som beskrivs med en lämplig uppsättning parametrar. En rimlig uppsättning sammansättningsparametrar är antalet molekyler i var och en av de kemiskt rena komponenterna utav vilka systemet är en blandning. Alternativt, för att erhålla tal av en mer hanterlig storlek, inför vi *moltalen,* definierade som det faktiska antalet molekyler av varje typ dividerad med Avogadros tal. (Callen 1985, s 9)

Enligt denna förklaring av hur moltalen skall uppfattas varierar variablerna N_i över heltal eller rationella tal, dvs bråk av heltal. Men detta är inkonsistent med det som sägs om differentierbarheten av entropifunktionen. De partiella derivatorna $\partial S/\partial N_i$ införs faktiskt explicit (s 41) tillsammans med andra s k entropiska intensiva parametrar i termer av vilka den s k entropiska fundamentala relationen är definierad. Detta förutsätter uppenbarligen att variablerna N_i varierar över de reella talen.

Matematiska påståenden säger något om tal (eller andra s k matematiska entiteter). Vad som är oklart är hur påståenden om tal kan användas för att uttrycka fysikaliska påståenden. Ett bättre sätt att uttrycka detta är nog att säga att det inte är klart vilka dessa fysikaliska påståenden är. Det är därför det finns ett problem med den fysikaliska tolkningen av matematiskt formulerade påståenden som är avsedda att handla om något annat än tal. Duhem tar upp just detta problem, även om han inte skulle ha illustrerat problematiken med ovannämnda exemplet. Vi får ett grepp om vilket fysikaliskt påstående som avses med en matematisk formulering genom att betrakta vilka slutsatser som uppfattas som rimliga.

3. Duhems holism

Duhem utvecklar sitt argument mot möjligheten av ett, bokstavligt uppfattat, avgörande experiment i samband med en diskussion av systematiska och andra felkällor. Hur man redogör för felkällor för att

korrigera observationer och garderar sig genom att inskränka det som hävdas som resultat av ett experiment är något som ingår i grundutbildningen i naturvetenskapliga ämnen. En vetenskapsman är medveten om dessa korrigeringar och garderingar, även om de ofta tas för givet, inte alltid görs explicit i läroböcker och således inte uppenbarar sig för den oerfarna läsaren.

Insamlingen och bedömningen av felkällor beskriver Duhem som en procedur som *ger en bättre bild* av det som verkligen inträffar under ett experiment. Han sammanfattar en lång diskussion med många exempel genom att beskriva ett experiment i fysik som

> den noggranna observation av en samling fenomen åtföljd av en *tolkning* av dessa fenomen; denna tolkning ersätter det konkret givna, genom observation faktiskt insamlade, med abstrakta och symboliska representationer som svarar mot dem i kraft av de teorier som är godtagna av observatören. (1914, ss 221-2)

Att säga precis vad man menar innebär att man explicit specificerar inom vilka gränser man är beredd att stå för de resultat man hävdar. 'Precision innebär approximation', för att uttrycka det i ett slagord. Ju mindre klart gränserna specificeras, desto vagare är påståendet. Motiveringen för Duhems holistiska tolkning av lagar och vetenskapliga påståenden är att den är inbakad i vad det innebär att säga att de är precisa. En genomgång av tilläggshypoteser som talar om inom vilka gränser man kan dra acceptabla slutsatser förtydligar innebörden av en hypotes.

Det är då klart att de tilläggshypoteser som här avses är hypoteser som försanthålles. Blotta den logiska konsekvensrelationen mellan en hypotes med godtyckliga satser och en slutsats ger ingen upplysning om felkällor. Detta görs bara av faktiskt dragna slutsatser, eller slutsatser som en vetenskapsman kunde dra om han skulle betrakta relevanta omständigheter tillsammans med hypotesen. (Jfr distinktionen mellan logiska regler och inferens i Harman 1973.) Att detta är vad Duhem avsåg framgår med all önskvärd tydlighet:

> ... det han [Regnault] förmedlade till oss från detta experiment är inte en beskrivning av observerade fakta utan de abstrakta symboler som *de godtagna teorierna* tillät honom att byta ut mot den konkreta evidens han insamlat. (1914, s 221; min emfas)

> Öppna vilken som helst tidskriftsartikel i experimentalfysik[;] ... slutsatserna ... är abstrakta påståenden som ni inte kan tillskriva någon innebörd om ni inte känner till de fysikaliska teorier som *författaren godtar*. (1914, ss 222-3; min emfas)

> ... de fenomen som fysikern faktiskt observerar är tolkade i enlighet med *godtagna teorier* ... (1914, s 231; min emfas)

> ... för att kunna tolka resultaten av detta experiment [som skulle visa oriktigheten hos ett påstående] och konstatera att det förutsagda fenomenet inte inträffat, begränsar han sig inte enbart till påståendet i fråga; han använder en uppsättning teorier som han *godtar som* obestridliga. (1914, ss 280-1; min emfas)

I sin diskussion av Regnaults försök att förbättra Boyles lag nämner Duhem att Regnault förbiså en systematisk felkälla som andra sedan har rättat till (att kvicksilvrets täthet varierar med höjd). Alltså redogörelsen för felkällorna – dvs för vad som kan vara relevant för att avgränsa innebörden av ett fysikaliskt påstående – kan mycket väl kritiseras senare. Det slutliga 'fullständiga' förtydligandet av ett påstående kanske aldrig formuleras. Detta är knappast en fråga om vaghet. Det finns vid en viss tidpunkt ingen högre preciseringsstandard än de uttalade precisionavgränsningar som upprättas då. Men denna tanke om innebörden av precision kan uppfattas som vaghet, om man bortser ifrån Duhems tes att olika uttryck som förefaller vara inbördes motsägande matematiska påståenden inte uttrycker olika fysikaliska påståenden.

> Ett ökat tryck på 100 atmosfärer ger upphov till en 0,0845 volts ökning i den elektromotoriska kraften hos ett gasbatteri. Det kunde likaväl ha sagts att denna ökning i trycket ger upphov till en 0,0844 volts ökning i den elektromotoriska kraften, eller en 0,0846 volts ökning. Hur kan dessa olika påståenden vara ekvivalenta för fysikern? För matematiken motsäger de varandra; ...
>
> Det fysikern menar när han påstår att dessa tre omdömen är i hans ögon identiska är följande: utifrån ett värde av 0,0845 volt på förminskningen i den elektromotoriska kraften räknar han med hjälp av godtagna teorier ut vilken avvikelse som galvanometervisaren kommer att undergå när han sänder in i instrumentet den ström som batteriet

åstadkommer. ... Om han upprepar samma beräkning utifrån en 0,0844 volts förminskning eller utifrån en på 0,0846 volt, kommer han fram till andra värden på magnetens avvikelse; men dessa tre således uträknade avvikelser kommer att vara för lite skilda för att det skulle vara möjligt att med ögonen åtskilja den ena från den andra. (1914, ss 229-30)

Trots den realistiska tolkning av Duhem som likheten med Quines filosofi gör påfallande, är den antirealistiska tolkningen av Duhem kanske den som man vanligen möter i litteraturen. Som en recensent nyligen uttryckte sig:

för mig, ... är Duhem, trots allt, en uppdaterad version av Bellarmine, eftersom Duhems 'naturliga klassificering' kan betraktas som en utvidgning av den instrumentalistiska konventionalism som återfinns hos Bellarmine. (Murad 1994, s 315)

Här åsyftas följande välkända passage hos Duhem:

... ju mer perfekt den [den fysikaliska teorin] förblir, ju mer vi anar att den logiska ordning i vilken teorin uppställer experimentella lagar återspeglar en ontologisk ordning; desto mer misstänker vi att de relationer som etableras mellan det som gives i observationen svarar mot relationer mellan ting, och desto mer anar vi att teorin tenderar mot en naturlig klassificering (1914, s 35).

Sådana antirealistiska tolkningar tycks ofta vara baserade på inget annat än Duhems ovilja att acceptera atomteorin för materien. Duhems attityd – som utformades i slutet av 1800-talet då atomteorien var i hög grad kontroversiell – kan inte jämföras med attityden hos 'positivisten', som hävdar att atomer inte existerar eftersom de inte syns. Eftersom hans holism ju är en kritik av föreställningen att observation är en enkel sak som direkt testar teoretiska påståenden, måste hans inställning rimligen förväntas bygga på en utvärdering av detaljerna i atomteorierna. Duhem var tveksam till det godtyckliga valet av en bland många till synes jämställda hypoteser, och han kritiserade de atomhypoteser han kände till som hopplöst vaga eller inkoherenta. Liksom när det gäller eterteorierna är de egenskaper som postuleras för att förklara ett visst fenomen oförenliga med egenskaper postulerade för att förklara andra fenomen. Man kan gissa att det aristoteliska

kravet på logisk ordning vägde tyngre för Duhem än den systematiska förklaring av spektra som Bohrs inkonsistenta atomteori erbjöd några år innan Duhem dog. Duhem bestred att man skulle kunna undvika dessa invändningar genom att göra anspråk på en insikt med ett särskilt sanningskriterium och betonade att *'överensstämmelse med experiment är för en fysikalisk teori det enda kriteriet på sanning'* (1914, s 26). Försök att ge 'mekaniska' förklaringar som inte bygger på testbara allmänna lagar från t ex mekaniken – för vilket Newton kritiserade Descartes' 'mekaniska' förklaringar – kallade Duhem för metafysik. Även om benämningen passar Descartes, är den tämligen olycklig med tanke på den roll Aristoteles spelar i Duhems tankebyggnad – inte minst dennes diskussion av materia, som man nog kan säga är paradigmatisk för vad som är metafysik. (En mer fullständig redogörelse för Duhems realistiska hållning än den som ryms här skulle behandla detta tema i detalj.) Men om Vuillemin kan använda ordet 'metafysik' för att beskriva Quines filosofi trots att Quine själv inte använder det, så kunde han tagit under förnyat övervägande lämpligheten av att ordagrant följa Duhems retorik och i samma sats förneka Duhems intresse för metafysik. Nu håller jag mig till den kritiska tråden i Duhems filosofi, och jag skall avslutningsvis ta upp Duhems invändningar mot Poincarés konventionalism.

Den antirealistiska eller instrumentalistiska tolkning som ofta tillskrivs Duhem är ett slags konventionalism trots hans invändningar mot denna doktrin. Vad Duhem närmare bestämt kritiserade var den form av antirealism som Poincaré kallade för konventionalism. I sin kritiska genomgång av denna doktrin åberopar Duhem olika drag i sin syn på vad det är som utgör en fysikalisk tolkning av en sats, dvs av vad det innebär att en sats är ett fysikaliskt påstående.

Duhem uppfattade konventionalismen som uppfattningen att vissa allmänna vetenskapliga påståenden har en privilegierad status i kraft av att de har en immunitet mot prövning genom experiment och observation; de definierar snarare hur vissa viktiga, centrala begrepp i fysikerns ordförråd bör användas. Poincaré föreslog att t ex tröghetsprincipen uppfattas på ett sådant sätt att den inte kan motbevisas av experiment. Det som till synes är avvikelser redogör man för genom att postulera inverkan av gravitations- och friktionskrafter.

Duhem ser här ingen väsensskillnad mellan olika lagar, utan bara en gradskillnad. Faktorer som införs för att förklara avsaknad av verklig

tröghetsrörelse är av precis samma typ som korrigeringar införda för en exakt tillämpning av en lag. Det Duhem kallar den symboliska representationen tillhandahåller en bild av vad som verkligen händer i ett experiment med hjälp av många tilläggsantaganden, vilka mycket väl kan omfatta friktions- och gravitationsfenomen. Om ett experiment inte överensstämmer med förutsägelserna kan vi inte enbart på logiska grunder säga vilka av dessa många antaganden som bör förkastas. Men

> bland de teoretiska element som ingår i denna symbols sammansättning finns det alltid ett visst antal som fysiker under viss epok accepterar utan granskning och som obestridliga (1914, s 321).

Duhem jämför det Poincaré säger om mekaniken med en liknande åsikt om lagen om multipla proportioner. Enkla grundämnen *A, B* och *C* förenas i viktförhållandet *a:b:c* då de bildar en förening *M*, och de bildar en annan förening *M´* då de förenas i ett viktförhållande *xa:yb:zc*, där *x, y* och *z* är heltal. Man kan hävda att denna lag inte är empiriskt testbar p g a att en kemisk analys kommer att visa *M´*:s grundämnessammansättning enbart till en viss grad av approximation. Och det finns alltid heltal *x, y* och *z* sådana att viktförhållandet mellan de tre grundämnen kan representeras som *xa:yb:zc* till vilken som helst grad av approximation som man i förväg specificerar. Alltså, kan lagen om multipla proportioner aldrig visas vara felaktig medelst experiment.

Duhem avvisar ett sådant resonemang som absurt. Det förutsätter att lagen om multipla proportioner är 'blottad på all fysikalisk innebörd' (1914, s 327). Den erhåller fysikalisk innebörd enbart genom ett lämpligt införande av ordet 'approximativt', utan vilket påståendet urartar till en matematisk truism, nämligen att reella tal kan godtyckligt nära approximeras med ett heltalsförhållande. Ordet 'approximativt' signalerar att en enkel jämförelse med observation inte kommer ifråga. En jämförelse av en hypotes med erfarenheten förutsätter andra hypoteser: 'det vore absurt att underkastar lagen om multipla proportioner direkt kontroll' (1914, s 327). På samma sätt visar Poincarés argument enbart att lagar inte testas direkt. Mekanikens lagar är lika sårbara som alla andra lagar (något som Poincaré senare blev berömd för att, tillsammans med Einstein, ha visat). Det finns ingen garanti för att anpassningar av tilläggshypoteser inte kommer att uppfattas som ad hoc och att en ändring i hypoteser med många år på nacken framstår som den bästa lösningen. Logiken kan inte diktera vilken ändring det är som

skall göras. Ej heller kan logiken hjälpa någon enstaka hypotes genom att ge den en status som enbart tillkommer en definition.

Likväl innebär inte Duhems kritik av konventionalismen att han inte ser någon plats för konventioner. I Duhem (1892, 1902) utvecklas en uppfattning om *la formule chimique developée* utifrån hans uppfattning om *la formule chimique brute*. Den senare termen svara mot den engelska termen *compositional formula*, och jag väljer här att använda 'sammansättningsformel' snarare än den på svenska etablerade termen 'formelenhet'. ('Formelenhet', liksom den franska och den engelska termen, 'ange[r] ämnet och dess sammansättning' (Hägg 1973, s 33), men tyvärr har den en definition som förutsätter atomteorin.) Duhem diskuterar i vilken utsträckning sammansättningsformeln bestäms av lagarna om bestämda och multipla proportioner. Vattnets viktförhållande mellan väte : syre är 1 : 8, vilket brukar anges med den vanliga sammansättningsformeln H_2O. Uppenbarligen är förhållandet 2 : 1 i formeln inte entydigt bestämt av viktförhållandet, och Duhem utreder i vilken mån jämförelse med vätes viktförhållande till t ex svavel, klor, fosfor och kol i ämnen vars vanliga sammansättningsformler är H_2S, HCl, PH_3 och CH_4, osv, begränsar möjliga index i sammansättningsformeln för t ex vatten. Han kommer fram till att det finns ett visst spelrum, vilket innebär att H_2O kan betraktas som den korrekta sammansättningsformeln för vatten endast med avseende på vissa konventioner (1892, s 397; 1902, s 73).

Men konventionalismen är en lära som berör lagarnas status, och att konventioner är inbyggda i den nyss illustrerade tillämpningen av lagar innebär inte att lagarna själva är konventioner. Längre fram i Duhem (1902) presenteras praktiskt taget samma formulering av innebörden av lagen om multipla proportioner som senare återgavs i den mer kända Duhem (1914), dock utan att Poincaré nämns. Där införs tanken om innebörden av lagen som svar på frågan 'Vilken är då den exakta innebörden av *den sanning* som vi nyss formulerat?' (1902, s 145; min emfas). Det är denna innebörd som går förlorad när lagar är 'blottad på all fysikalisk innebörd' och reduceras till 'matematiska truismer'.

4. Slutord

Påstådda olikheter mellan Quine och Duhem m a p den s k Duhem-Quine-tesen är överdrivna, och döljer möjligheten till en ömsesidig belysning av deras filosofiska idéer. Visst finns det skillnader. Quines

inställning till traditionella principer i matematik, t ex, är säkert mer radikal än Duhems. Och Duhem uppmärksammar en aspekt av problematiken kring användningen av matematik inom fysiken som inte kommer fram i Quines diskussioner, nämligen att tolkningen hindrar att man drar slutsatser som beror på distinktioner som inte kan berättigas inom de specificerade precisionsgränserna. Duhems tanke att precision säger något om hur olika matematiska satser hänger samman i kraft av att de används med samma fysikaliska innebörd är ett intressant uppslag, även om det är svepande och vagt. Det väcker många frågor om hur lagar och teorier artikuleras och förtydligas. Och sådana frågor bör undersökas. Uppslaget erbjuder ett alternativ till Fields sätt att besvara den fråga han ställer (s vii) om hur matematiken kan tillämpas på den fysikaliska världen, vilket passar väl in i Quines syn på syftet med regimentering.

Not.

Jag vill tacka Jan Österberg för hjälp med språkgranskning.

Litteratur

ARIEW, ROGER (1984), 'The Duhem Thesis', *British Journal for the Philosophy of Science*, 35 (1984), 313-325.

CALLEN, HERBERT B (1985), *Thermodynamics and an Introduction to Thermostatistics*, 2:e uppl, Wiley, New York, 1985.

CHRISTIE, MAUREEN (1994), 'Philosophers versus Chemists Concerning "Laws of Nature"', *Studies in History and Philosophy of Science*, 25 (1994), 613-29.

DOLBY, R G A (1984), 'Thermochemistry versus Thermodynamics: The Nineteenth Century Controversy', *History of Science*, 22 (1984), 375-400.

DUHEM, PIERRE (1892), 'Notation atomique et hypothèses atomistiques', *Revue des questions scientifiques*, 31 (1892), 391-457.
DUHEM, PIERRE (1902), *Le mixte et la combinaison chimique: Essai sur l'évolution d'une idée*, 1902; omtryckt Fayard, Paris, 1985.
DUHEM, PIERRE (1914), *La théorie physique: son object sa structure*, 2:e uppl., 1914 (1:e uppl 1906), omtryckt Vrin, Paris, 1981.
FIELD, HARTRY (1980), *Science Without Numbers*, Blackwell, Oxford, 1980.
HARMAN, GILBERT (1973), *Thought*, Princeton University Press, Princeton, New Jersey, 1973.
HÄGG, GUNNAR (1973), *Allmän och oorganisk kemi*, 6:e uppl, Almqvist & Wiksell, Stockholm, 1973.
KOYRÉ, ALEXANDRE (1965), *Newtonian Studies*, Methuen, London, 1965
MAIOCCHI, ROBERTO (1990), 'Pierre Duhem's *The Aim and Structure of Physical Theory*: A Book Against Conventionalism', *Synthese*, 83 (1990), 385-400.
MURAD, HAZIM (1994), Recension av Pierre Duhem, *German Science* samt R. N. Martin, *Pierre Duhem, Philosophy of Science*, 61 (1994), 313-5.
PAPINEAU, DAVID (1993), *Philosophical Naturalism*, Blackwell, Oxford, 1993.
POPPER, KARL R. (1972), *The Logic of Scientific Discovery*, Hutchinson, London, 1972.
QUINE, W.V. (1953), 'Two Dogmas of Empiricism', *Philosophical Review*, 60 (1951); omtryckt i *From a Logical Point of View*, Harvard University Press, 1953.
QUINE, W.V. (1960), *Word and Object*, M I T Press, Cambridge, Mass, 1960.
QUINE, W.V. (1986), Replies to the critical essays i L E Hahn och P A Schilpp, utg, *The Philosophy of W. V. Quine*, Open Court, Illinois, 1986.
SHAPIRO, STEWART (1983), 'Conservativeness and Incompleteness', *Journal of Philosophy*, 80 (1983), 521-31.
VUILLEMIN, JULES (1986), 'On Duhem's and Quine's Theses', i L E Hahn och P A Schilpp, utg., *The Philosophy of W V Quine*, Open Court, Illinois, 1986.

MARTIN GUSTAFSSON

Några kommentarer till Wittgensteins "Om visshet"

1. Teoretiskt tvivel och traditionell kunskapsteori

Är det principiellt otvivelaktigt att den här kritan inte innehåller en försåtligt dold sprängladdning som strax detonerar? Är det teoretiskt otänkbart att det, här framför mig i golvet, finns en osynlig fallucka som jag kommer att ramla ner i om jag tar ett par steg framåt? Är det absolut visst att en ambulanssjukvårdare, som just anlänt till en stor skadeplats och ser sig omgiven av nedblodade och kvidande människor, inte är utsatt för ett osmakligt *practical joke*, där de till synes skadade bara är uppsminkade skådespelare?

Det finns inget att invända mot påståendet att det i kunskapsteoretiska sammanhang är riktigt att svara 'nej' på de här frågorna. Det är tänkbart och principiellt möjligt att kritan snart exploderar, att det finns en osynlig fallucka här i golvet, och att ambulansföraren är lurad – så används faktiskt uttrycken 'tänkbart' och 'principiellt möjligt' i de flesta kunskapsteoretiska diskussioner, och det är inte sällan detta språkbruk som håller diskussionerna levande.

Karakteristiskt för detta 'principiella' tvivel på att kritan är ofarlig, eller på att golvet här framför mig är solitt, är att det vore orimligt att ta det i beaktande i praktiska sammanhang. Om vi säger att visst kan vi föreställa oss att här finns en fallucka, menar vi inte att jag ska kontrollera saken innan jag ställer mig på platsen; en sådan kontroll skulle vi uppfatta som komisk, ja rentav paranoid. Det principiella tvivlet på sådana saker, är inte menat som en uppmaning att verkligen noggrannare kontrollera hur saker och ting förhåller sig, innan vi agerar.

Att identifiera en nivå av sådant 'teoretiskt' tvivel, förefaller, som sagt, vara väsentligt för en stor del av den traditionella kunskapsteoretiska diskussionen. Ja, många viktiga kunskapsteoretiska uppfattningar eller positioner kan tydligt formuleras bara mot bakgrund av det sätt

att tala om 'tvivelaktighet', 'tänkbarhet', 'ifrågasättande' osv., som präglar det teoretiska tvivlets diskurs. Den omfattande diskussionen kring huruvida den riktiga användningen av 'vet att p' kräver grunder som 'garanterar' utsagan p:s sanning, begriper man bara om man är införstådd med att det här handlar om 'principiell' eller 'teoretisk' garanti. Och fallibilismen, samt invändningarna mot den, går att helt och fullt begripa bara mot bakgrund av talet om 'principiell' felbarhet och ofelbarhet. Detsamma gäller skepticismen, åtminstone så som den oftast uppfattas i kunskapsteoretiska sammanhang. De skeptiska skräckscenarierna – att jag är bedragen av en ond demon, eller att min hjärna manipuleras av en sadistisk vetenskapsman – är inte menade som uppmaningar till att verkligen ompröva vårt förtröstansfulla handlande, utan sägs vara av 'rent teoretiskt' intresse.

2. Det teoretiska tvivlets oklara innebörd

Nu vore det fel att påstå att det sätt att tala om 'teoretiskt tvivel', 'principiellt rum för ifrågasättande', etc., som präglar mycken kunskapsteori, är alldeles klart och tydligt. Det ger förvisso något slags struktur åt den kunskapsteoretiska diskussionen, tillräcklig för att hålla igång den. Och det finns, som sagt, en tämligen bred konsensus om att det faktiskt existerar påståenden vars sanning det vore absurt att inte lita på i praktiken, men som kan 'teoretiskt betvivlas'.

Men att klargöra meningen hos ett uttryck som, låt oss säga, 'teoretiskt otvivelaktig', genom att ge okontroversiella och riktigt upplysande exempel på eller förklaringar av vad det innebär att ett påstående är teoretiskt otvivelaktigt är inte lätt. Ett exempel på svårigheten är Keith Lehrers försök (1990, s 40–8) att klargöra innebörden av 'infallible foundation' –'ofelbar grundval' – genom att säga att en sådan grundläggande övertygelse 'garanterar' sin egen sanning – som om 'garanterar' här vore klarare än 'ofelbar'! Strax därefter tar han upp somliga tänkares uppfattning att en övertygelse som 'garanterar' sin egen sanning ska förstås som 'okorrigerbar'. Och 'okorrigerbar' förklaras slutligen i termer av 'logisk omöjlighet'. Mot bakgrund av de många välkända oklarheterna kring och problematiseringarna av begreppen logisk omöjlighet och logisk nödvändighet, och då inte minst frågetecknen kring logikens 'okorrigerbarhet', är det svårt att se Lehrers resonemang som något annat än ett ansträngt försök att hålla fast vid en sluten och självgående retorisk apparat. Ett ytterligare

tecken på detta, är att hans resonemang bär den filosofiska desperationens kanske främsta kännetecken, nämligen det ständiga *kursiverandet* av de termer han förlitar sig på i sina 'förklaringar': '*garanterar*', '*okorrigerbar*', '*logisk* omöjlighet'.

3. Var Wittgenstein epistemolog?

En del uttolkare av *Om visshet* menar att Wittgensteins diskussion av i synnerhet skepticismen, rör sig inom den traditionella kunskapsteoretiska begreppsram jag just försökt antyda. Gunnar Svensson skriver till exempel såhär (1992, s 16):

> Det är inte bara så att Wittgenstein [...] försöker visa att en skeptisk ståndpunkt är ohållbar, eller att vi i praktiken är tvungna att omfatta vissa åsikter. Inte heller är det så att han nöjer sig med att försöka visa att skeptiska tvivel saknar mening. Han försöker därutöver visa att vi av logiska skäl är tvungna att med objektiv visshet omfatta vissa påståenden. Wittgensteins anteckningar rör alltså inte bara skeptikerns tvivel utan också epistemologens visshet. Det är en insikt om denna visshet, och inte bara en insikt om att skeptikerns tvivel är tomt, som till sist kan tillfredsställa ett epistemologiskt sökande. Det Wittgenstein i detta sammanhang kommer fram till liknar därför i hög grad det som den traditionella epistemologen varit ute efter.

Den metod Wittgenstein använder sig av i denna sin attack på skeptikern består, säger Svensson, i att ge en översiktlig framställning av de epistemologiskt problematiska uttryckens användning; en framställning som visar, eller gör oss uppmärksamma på, de språkregler som faktiskt styr användningen av uttrycken ifråga. Genom en sådan 'grammatisk undersökning' blir vi klara över att skeptikern måste missbruka språket för att överhuvudtaget kunna formulera de satser som är menade att uttrycka skeptiskt tvivel. Det blir också, tänker sig Svensson, klart att somliga påståenden är 'objektivt vissa'. Därmed skulle skeptikern vara desavouerad.

Denna anti-skeptiska strategi har en fundamental svårighet. Och den är: varför skulle jag, om jag anser mig tvivla på sådana saker som att *detta är en hand* eller *jorden existerade för 10 minuter sedan*, inte tvivla på resultatet av varje grammatisk undersökning som någon producerar i avsikt att tillbakavisa min skepticism? Det är en oerhört

besynnerlig tanke att tvivel på att *detta är en hand* skulle på något sätt kunna tillbakavisas genom ett filosofiskt studium av språkanvändningen. För vilket resultat av en grammatisk undersökning kan sägas vara troligare, mer otvivelaktigt, än att *detta är en hand*? Att tvivla på resultatet av en sådan undersökning, kräver inte ens fantasifulla hjärna-i-bunke- eller bedragande-demon-argument. Det räcker med en smula erfarenhet av hur lätt det är att gå vilse och begå misstag i grammatiska undersökningar av det slag Svensson talar om. Tvivlar jag på att *detta är en hand*, är det närmast inkonsekvent att inte tvivla på varje kartläggning av tvivlets och visshetens grammatik. Den anti-skeptiska strategi Svensson tillskriver Wittgenstein, lämnar varje skeptiker av någorlunda kaliber orubbad.

4. En alternativ strategi

Så skeptikern sitter, som vanligt, med trumf på hand, just därför att inget argument som man tycker sig kunna använda mot honom är mindre tvivelaktigt än de saker han redan säger sig tvivla på. Varje försök att tillbakavisa skeptikern genom att *visa* att det han säger sig tvivla på är 'principiellt otvivelaktigt', och att hans tvivel är 'tomt', är fåfängt. Går man med på att föra en diskussion med skeptikern i termer av 'principell otvivelaktighet' och 'teoretisk tänkbarhet', kommer man aldrig tillrätta med hans tvivel.

För att råda bot på den skeptiska sjukan, måste man alltså ställa sig på betydligt större skyddsavstånd från den traditionella kunskapsteorins försåtliga begreppsarsenal än vad Gunnar Svensson menar att Wittgenstein gör. Det innebär att på allvar bearbeta den idé om en sfär av 'teoretiskt', 'principiellt' tvivel, som spökar i såväl skeptikerns som anti-skeptikerns olika sätt att argumentera. Vilket inte betyder att *förneka* möjligheten av teoretiskt, principiellt tvivel; det vore absurt, eftersom sådant tvivel faktiskt förekommer i filosofiska sammanhang. Nej, vad som måste bearbetas är snarare själva *intresset* för detta sätt att tala om tvivel och visshet.

5. Wittgenstein om visshet

Om visshet är en bok som är svår att förstå. Att den innehåller många olika, mer eller mindre motstridiga, tankegångar och uppslag, är naturligt, eftersom manuskriptet består av relativt obearbetade anteckningar. Att påstå sig ha identifierat de mål mot vilka Wittgenstein säger sig

"oavlåtligen ha skjutit" (§387), måste innebära att betrakta en hel del av bokens tankegångar och formuleringar som inkonsekventa snedsteg. Men att hitta entydiga belägg för exakt *vilka* av tankegångarna som utgör sådana snedsteg, låter sig knappast göras.

Om visshet innehåller en hel del undersökningar av hur uttryck som 'vet', 'tvivel' och 'visshet' används i vårt språkbruk. Gunnar Svensson, Marie McGinn och andra uttolkare menar att dessa undersökningar fungerar som anti-skeptiska argument. Jag har försökt visa att de inte kan fungera så; och om Wittgenstein verkligen gav dem denna uppgift, begick han ett misstag.

Jag tror för min del inte att *Om visshet* innehåller någon som helst anti-skeptisk argumentation. En stor del av Wittgensteins arbete där bör snarare karakteriseras som en terapeutisk granskning av det intresse såväl skeptikern som anti-skeptikern fäster vid diskussionen kring den 'principiella' eller 'teoretiska' otvivelaktighetens möjlighet.

Wittgenstein ställer flera gånger frågan varför tvivel som inte har, eller är menat att ha, några praktiska konsekvenser, är något att oroa sig över, eller bry sig om. Angående påståendena att jag har en hjärna, och att bordet finns när ingen ser det, undrar han:

> Men om nu någon betvivlade det, hur skulle hans tvivel visa sig i praktiken? Och kunde vi inte låta honom tvivla i lugn och ro, eftersom det ju inte gör någon skillnad? (§120)

Om de som säger att det inte är alldeles säkert, utan bara högst sannolikt, att vattnet i kitteln som står på elden kommer att koka och inte frysa, frågar han:

> Vilken skillnad gör detta i deras liv? Är det inte bara så att de talar mera om vissa saker än andra människor gör? (§338)

Ibland raljerar han över diskrepansen mellan praktiskt och teoretiskt tvivel, till exempel i den välkända anteckning där han tänker sig sitta framför ett träd tillsammans med en filosof som gång på gång upprepar orden "Jag vet att det där är ett träd". Till en förbipasserande ger Wittgenstein förklaringen till filosofens beteende: "Denna människa är inte galen: Vi filosoferar bara." (§467).

Om man utgår från att Wittgenstein är ute efter att leverera ett anti-skeptiskt argument, är det svårt att betrakta användningen av humor och ironi som mer än en tillfällig, retorisk utsmyckning av hans

tankar. Uppfattar man däremot syftet med undersökningarna på det sätt jag gör, får det raljerande inslaget en mer väsentlig roll: att lära sig skratta åt de filosofiska talesätt som bekymrar en, är ett viktigt steg i behandlingen av ens intresse för dessa talesätt.

Nu är ironiserandet över de kunskapsteoretiska språkspelen inte ett dominerande inslag i *Om visshet*. Betydligt mer energi ägnar Wittgenstein åt att uppmärksamma det tillitsfulla, grundlösa handlandets centrala plats i vårt liv och vårt kunskapssökande. Jag ska avslutningsvis förklara varför jag tror att detta tema ska förstås som ytterligare ett sätt att bearbeta intresset för frågorna kring det 'principiella' tvivlet och den 'teoretiska' ofelbarheten.

Det principiella tvivlet på att marken framför oss är solid, vore inte bara komiskt eller onödigt att ta i beaktande i praktiken. Nej, det vore faktiskt *förödande* om vi, innan varje nytt steg, undersökte platsen där vi planerade att sätta foten, för att se om där finns en dold fallucka. Vi skulle bli fördröjda i allt vi gjorde, inte förmå utföra viktiga arbeten, få problem att försörja oss själva, och så vidare. Att i praktiken hysa ogrundad tillit till att marken framför oss är trygg att gå på, är helt enkelt nödvändigt i en enorm mängd väsentliga sysselsättningar.

Analogt gäller att om ambulanssjukvårdaren, innan han påbörjar sin vårdinsats, verkligen kontrollerar att de kvidande människorna omkring honom inte är uppsminkade skådespelare, skulle det inte bara vara onödigt omständligt; nej, det skulle betyda att sjukvårdaren grovt missköter sitt jobb, och bör omplaceras utan vidare.

Wittgenstein har ett annat belysande exempel (§315). En person har tappat bort något, och vill finna det igen. Han öppnar en låda, och finner inte föremålet där. Han stänger lådan, väntar, och öppnar den sedan på nytt, för att se om föremålet kanske finns där nu. Han fortsätter på det sättet. Vad ska vi säga här? Jo, att personen inte lärt sig söka; ty till sökande hör, att man litar på att ett föremål inte hux flux dyker upp på ett ställe som man nyss genomsökt. Om vi tvivlade här, skulle sökandet aldrig komma igång.

Wittgenstein uppmärksammar en mängd analoga exempel: om vi i praktiken betvivlar naturens lagbundenhet, kommer vi inte att kunna tillägna oss vetenskaplig skolning (§315), om vi tvivlade på att jorden existerade för 100 år sedan, skulle historisk forskning och undervisning köra fast (§316), ständigt praktiskt tvivel på experimentapparaturens tillförlitlighet och existens skulle förhindra allt experi-

menterande (§337), och i matematiken måste jag lita på att siffrorna inte av sig själva byter plats på papperet och på att mitt minne inte spelar mig ständiga spratt (ib.). Ja, Wittgenstein menar att till och med fällandet av omdömen, själva språket, vilar på den ogrundade tilliten till ens egen förmåga att använda språket riktigt i en mängd elementära fall. Han skriver:

> Hur avgör någon, vilken som är hans högra hand och vilken som är hans vänstra? Hur vet jag att mitt omdöme kommer att stämma överens med de andras? Hur vet jag att denna färg är blått? Om jag inte här litar på *mig*, varför skulle jag lita på den andres omdöme? Finns det ett varför? Måste jag inte någonstans börja lita? Dvs. jag måste någonstans börja med icke-tvivlandet; och detta är inte, så att säga, förhastat men förlåtligt, utan det hör till fällandet av omdömen. (§150)

Slutsatsen man kan dra av dessa iakttagelser, är att mycken grundlös tillit är väsentligen förknippad med verksamheter som vi inte kan avstå från, med mindre än att avstå från allt som betyder något för oss. Oavsett om det vi i dessa fall förlitar oss på är principiellt tvivelaktigt eller otvivelaktigt, måste vi fortsätta att i praktiken hysa denna tillit som nu, *och utan att söka grunder för den*, om vi vill ägna oss åt att exempelvis promenera, ta hand om skadade medmänniskor, räkna matematik, betvivla saker och ting, och använda vårt modersmål – kort sagt, ägna oss åt det som utgör våra liv (§344). Att endast lita på det man har goda grunder för, är alltså ett ideal som det vore förödande att ens sträva efter att förverkliga inom något för oss riktigt betydelsefullt område. Det vore inte bara så att säga onödigt försiktigt att alltid i praktiken söka grunder innan vi börjar lita; nej, det skulle rasera hela vår tillvaro.

Men på vilket sätt kan nu detta konstaterande hjälpa till att råda bot på vårt intresse för frågorna kring det principiellt, absolut vissa? Spelar inte Wittgenstein snarast skeptikern i händerna här, genom att tillskriva det grundlösa handlandet sådan betydelse?

Jag tror inte det. Vad Wittgenstein gör, är att visa att vår grundlösa tillit till saker och ting som i många fall svårligen kan karakteriseras som 'i princip otvivelaktiga' genomsyrar vårt tänkande och vårt liv så till den grad *att det framstår som absurt att alls beklaga denna tillit*, eller bekymra sig över den på något sätt, eller vilja förankra den i den

'objektiva vissheten'. Att vilja ha något utöver denna grundlösa tillit, att verkligen *sakna* grunder för den, verkar med ens orimligt. Vi klarar oss inte bara bra med den; vi klara oss överhuvudtaget inte utan den. Wittgenstein ägnar sig alltså inte åt någon Humesk *resignation* inför det faktum att vi i praktiken faktiskt nöjer oss med det ena och det andra. Han *gläder sig* snarare inför detta faktum! Där Hume på ett teoretiskt plan trots allt fortsätter att känna sig *besvärad*, frågar sig Wittgenstein ihärdigt: men varför, ens i teorin, känna sig besvärad? Han liksom vrider perspektivet, och tycks vilja säga: Den grundlösa tilliten står till vår tjänst! Se bara på allt den sätter oss i stånd att göra! Se hur oersättlig den faktiskt är! Så vad har ni skeptiker och anti-skeptiker egentligen att vara så upphetsade över?

Litteratur

LEHRER, KEITH (1990), *Theory of Knowledge*, London: Routledge.
MCGINN, MARIE (1989), *Sense and Certainty*, Oxford: Basil Blackwell.
SVENSSON, GUNNAR (1992), *Wittgenstein om kunskap och visshet*, Stockholm: Thales.
WITTGENSTEIN, LUDWIG (1992), *Om visshet*, Stockholm: Thales, övers Lars Hertzberg.

LARS BERGSTRÖM

Utilitarismens fel och förtjänster

Man kan säga att utilitarismen fått ett icke obetydligt genomslag bland svenska filosofer. Delvis har jag kanske själv bidragit till det. Jag har åtminstone länge tyckt, och tycker fortfarande, att utilitarismen har stora förtjänster. Annars är det väl i Sverige främst Ingemar Hedenius som har gjort propaganda för utilitarismen – och på senare tid Torbjörn Tännsjö.

För drygt tjugo år sedan höll jag ett offentligt föredrag i vilket jag visserligen hävdade att utilitarismen "måste utarbetas närmare" för att undvika allvarlig kritik av olika slag, men att den ändå är det "mest lovande" och "mest acceptabla" av alla de moralsystem som föreslagits under historiens lopp. Idag är jag mer tveksam.

Utilitarismen kan utformas på flera olika sätt. I mitt föredrag hävdade jag att den borde uppfattas som *riktighetskriterium* (inte som beslutsmetod) och som *handlingsutilitarism* (snarare än som regelutilitarism). För övrigt menade jag att den borde innebära att en handlings moraliska status bestäms av dess *totala, faktiska* konsekvenser (snarare än av t ex dess förutsägbara eller sannolika konsekvenser). Så uppfattas den också av t ex G E Moore och i Sverige av Torbjörn Tännsjö. Det innebär alltså att en enskild handling är moraliskt riktig om och endast om den leder till minst lika bra konsekvenser som varje annan handling som agenten kunde ha utfört i stället. Det är denna egenskap som gör riktiga handlingar riktiga. Hur man sedan i praktiken skall bära sig åt för att bestämma sig för vad som är riktigt är en annan sak. Det är inte alls säkert att man enligt utilitarismen skall försöka beräkna de olika alternativens konsekvenser och deras relativa värden. Som redan Henry Sidgwick påpekade är det kanske även ur utilitarismens synpunkt bättre att folk i allmänhet följer vissa enkla moralregler om att hålla löften och tala sanning och så vidare. Sådana

konventionella moralregler kan betraktas som en sorts "tumregler" som man bör lära ut till sina barn och själv efterleva. Därmed blir troligen konsekvenserna bättre än om var och en försöker handla på grundval av noggranna utilitaristiska konsekvensanalyser av varje alternativ.

Tanken är alltså att utilitarismen blir *rimligast* om den tolkas på det sätt jag här har skisserat. Inom denna ram finns sedan en rad andra tolkningsvarianter – se exempelvis Erik Carlsons utmärkta avhandling *Consequentialism Reconsidered* (Kluwer 1995) – som jag dock inte skall trassla in mig i här.

1. Utilitarismens förtjänster

Det enda som ytterst spelar roll är kanske om utilitarismen är *sann*. Detta är dock inte lätt att avgöra. Det syns ju inte direkt på den – lika lite som man direkt kan se på vetenskapliga teorier om de är sanna. Hur skall man då gå tillväga för att värdera utilitarismen som morallära? Vilka krav kan vi ställa på den? Vilka förtjänster har den? Vilka nackdelar? Mycket har sagts om detta. Jag skall här sammanfatta det hela, på mitt eget sätt, i fem fördelar och fem nackdelar. Låt oss börja med fördelarna.

Den är moraliskt tilltalande. I framför allt två avseenden är den en moraliskt attraktiv lära. För det första tar den fasta på hur det faktiskt går för dem som berörs av våra handlingar, hur deras välbefinnande påverkas eller hur deras intressen tillgodoses. Därigenom generaliserar den åsikten att man bör behandla andra som man själv skulle vilja bli behandlad. För det andra är den opartisk. Den avvisar alla former av egoism, inklusive gruppegoism, rasegoism och nationalism, och den insisterar på att alla som berörs skall tillmätas samma betydelse (*each to count for one*), oavsett om de är djur eller människor och oavsett om de lever nu eller tillhör kommande generationer.

Den är hittills optimal. Möjligen kan man säga att den än så länge är bättre än de alternativa moralteorier som har föreslagits. Med detta kan man i så fall mena, dels att den är normativt mer tilltalande (även om smaken här är olika), och dels att den är mer teoretiskt avancerad. En vanlig tanke i vetenskapsteori är ju att man inte bör överge en teori, ens om den har allvarliga problem, innan man har hittat en bättre. Något liknande kunde kanske anföras som stöd för utilitarismen. Framför allt är den konfliktlösande, djärv, och teoretiskt fruktbar. Detta räknar jag som de återstående tre fördelarna.

Den är konfliktlösande. Den säger något om hur man skall lösa moraliska konflikter, dvs situationer där olika moraliska överväganden pekar åt olika håll. Detta är kanske det viktigaste kravet på en moralteori, ty en moralteori skall ju lösa moraliska problem, och moraliska problem uppstår ju framför allt i de fall där olika moraliska överväganden pekar åt olika håll. Andra moralteorier, som t ex Kantianism, rättighetsteori, klassisk dygdeetik och kristen moral, lämnar en här i sticket.

Den är djärv. Man kan nog säga att utilitarismen är en "djärv gissning" (*bold conjecture*) i ungefär den mening som Karl Popper avsåg när han hävdade att bra vetenskapliga teorier skall vara djärva gissningar. Andra moralteorier (som t ex Kantianism, kristen etik, existentialism, etc) är mer intetsägande och oprecisa. Därför är det också svårare att säga vad de påbjuder i enskilda fall, åtminstone om dessa enskilda fall är tänkta exempel av det slag som moralfilosofer brukar diskutera. Utilitarismen tycks ge klarare utslag, och därför är den också lättare att kritisera – vilket Popper, i andra sammanhang, ser som en stor förtjänst.

Den är teoretiskt fruktbar. Utilitarismen är teoretiskt fruktbar i den meningen att den ger upphov till en mängd intressanta problem och till mycket diskussion. Det beror nog bl a på att den bättre än andra teorier lämpar sig för applikation på tänkta exempel, samt på att den – åtminstone vid första anblicken – tycks ge upphov till vissa moraliskt obehagliga konsekvenser som det då för moralteoretikern gäller att bortförklara.

2. *Utilitarismens nackdelar*

Emellertid har utilitarismen även en hel del nackdelar. Dessa påpekas ofta av dess motståndare, men i viss utsträckning erkänns de även av dess anhängare. Jag skall som sagt nämna fem nackdelar.

Den är moraliskt stötande. Den vanligaste typen av invändning mot utilitarismen består av något tankeexperiment som skall visa att den har moraliskt orimliga konsekvenser. Exempelvis att man bör straffa en oskyldig och onyttig person hellre än en som är skyldig men samhällsnyttig, om det går att övertyga allmänheten om att den oskyldige är skyldig. Eller att man med våld bör beröva en frisk person vissa vitala organ, som sedan kan transplanteras in på andra och därmed rädda flera liv, om detta kan göras utan att folks förtroende för sjuk-

vården rubbas. Eller att man bör öka jordens befolkning så länge alla får en dräglig tillvaro hellre än att ha en mindre befolkning där alla har det mycket bättre, men där den totala välfärden är något lägre. Dessutom tycks utilitarismen ofta kunna tillåta sådant som skattefusk, svartjobb, etc. Det är dock något oklart vad dessa exempel visar och hur realistiska de är.

Den är orimligt krävande. Eftersom utilitarismen kräver att jag alltid gör det som har de absolut bästa konsekvenserna, så verkar det rimligt att tro att vi alla alltid handlar moraliskt orätt enligt utilitarismen. Förmodligen kommer vi sällan ens i närheten av det som är rätt. Hellre än att lägga ut sextio kronor på en biobiljett till mig själv bör jag nog enligt utilitarismen ge pengarna till något nödhjälpsprojekt i tredje världen. Det låter i och för sig inte helt fel, men konsekvensen tycks bli att vi aldrig får gå på bio. Det är mindre rimligt. Och samma resonemang tycks kunna visa att vi inte heller får äta oss riktigt mätta eller ge våra barn nya kläder. Detta verkar ganska orimligt.

Den är ofullständig. Utilitarismen förutsätter att det finns objektivt riktiga svar på frågor om hur förbättringar för den ene förhåller sig till försämringar för den andre. Sådana frågor måste ha sanna svar om man på ett meningsfullt sätt skall kunna tala om *sammanlagda* välfärdsvinster eller välfärdsförluster med olika handlingsalternativ. Utilitarismen förutsätter med andra ord *interpersonella välfärdsjämförelser*, och jag tror man kan säga att ingen ännu lyckats visa hur sådana är möjliga. Att vi i praktiken ändå måste göra sådana jämförelser är inget försvar för teorin. Den är ofullständig så länge frågan om interpersonella jämförelser inte är löst på ett teoretiskt tillfredsställande sätt.

Den är grundlös. Det finns såvitt jag kan se *ingen evidens* för utilitarismen, ingenting som talar för att den är sann. Ofta tänker man sig ju att man har evidens för en teori om den utgör den bästa förklaringen av föreliggande fakta, men utilitarismen förklarar egentligen ingenting alls. I en uttunnad mening kan man kanske säga att utilitarismen "förklarar" sina egna konsekvenser – alltså det som kan härledas från utilitarismen i kombination med empiriska tilläggspremisser om enskilda situationer – men dessa konsekvenser ger knappast något stöd för utilitarismen eftersom deras enda grund i sin tur är just utilitarismen. (Att det är rationellt att godta konsekvenserna av vad man godtar är en

annan sak. Detta har ingenting med evidens eller induktiva slutledningar att göra.)

I ett föredrag i Stockholm någon gång i mitten av 80-talet kritiserade Dagfinn Føllesdal mig för denna uppfattning om utilitarismens grundlöshet (som jag tidigare framfört i *Filosofisk tidskrift* nr 1, 1980). Han hänvisade då till John Rawls' idé om ett "reflektivt ekvilibrium" mellan principer och enskilda omdömen, och han hävdade att de senare inte behöver något oberoende stöd för att de skall kunna utgöra ett stöd för de förra. Till en del håller jag med om detta. De enskilda omdömena behöver inget oberoende stöd – men det viktiga är att de inte själva enbart är motiverade av det de skall utgöra evidens för. Poängen är alltså att de inte kan utgöra något stöd för vissa principer, som t ex utilitarismen, om de i sin tur är baserade just på dessa principer (och om vi alltså inte hade accepterat omdömena om vi inte hade härlett dem från principerna).

För övrigt är det ju, som Folke Tersman visat (i *Southern Journal of Philosophy* vol 29, 1991), omöjligt att underbygga utilitarismen i ett reflektivt ekvilibrium, eftersom man inte kan ha tillgång till välgrundade empiriska premisser av det slag som skulle behövas (utom kanske i vissa tankeexperiment – men det faktum att en utilitarist kan tänka ut abstrakta situationer i vilka han skulle hålla fast vid utilitarismen ger knappast något stöd för utilitarismen). Vi kan alltså inte veta vilka konsekvenser utilitarismen har i enskilda fall. Detta leder över till den femte och sista nackdelen med utilitarismen.

Den är oanvändbar. Den är med andra ord *svår eller omöjlig att tillämpa* i praktiken, eftersom vi aldrig kan veta tillräckligt mycket om de totala konsekvenserna av våra handlingar. Även om vi kanske ibland kan har skäl att tro att en handling har bättre konsekvenser än en annan, så har vi troligen aldrig goda skäl att tro att en handling har bättre, eller minst lika bra, konsekvenser som varje möjligt alternativ.

Utilitarister är naturligtvis i allmänhet medvetna om detta problem. Men de brukar slingra sig undan det genom att hävda att ett riktighetskriterium inte behöver kunna tillämpas; det avgörande är om det är sant. Hur man sedan skall bete sig i praktiken för att avgöra hur man bör handla är en annan sak. Det är här man behöver en "beslutsmetod", och en sådan är något annat än ett riktighetskriterium. Denna distinktion går som sagt tillbaka till Sidgwick, och den motsvaras hos exempelvis R M Hare av distinktionen mellan en "intuitiv" och en

"kritisk" nivå i det moraliska tänkandet. Hos Thomas Wetterström (*Basic Ethics*, Oxford 1986) görs på motsvarande sätt en distinktion mellan "basic" och "pragmatic" etik. För egen del är jag dock rätt skeptisk mot dessa distinktioner. (Detta förklarar kanske att jag ibland kritiserats – t ex av Sören Halldén och Torbjörn Tännsjö – för att inte upprätthålla dem.) Låt oss se lite närmare på detta.

3. Utilitarismen i praktiken

Jag har inget speciellt för mig på eftermiddagen och funderar på att gå på bio. Men så slår det mig att ett bättre alternativ är att jag köper den vinnande lotten på penninglotteriet i stället, ty konsekvensen blir då att jag senare *både* kan gå på bio och skicka en miljon till välgörande ändamål. Alltså bör jag göra det.

Mot detta kan man invända att jag *inte kan* köpa den vinnande lotten, och att "bör" implicerar "kan". Alltså kräver inte moralen av mig att jag köper den vinnande lotten.

Men *varför* kan jag inte köpa den vinnande lotten? Den är ju ännu osåld (kan vi tänka oss). Ja, men jag *vet inte hur jag skall bära mig åt* för att köpa just den. Och detta medför att det följande inte gäller: om jag skulle försöka, så skulle jag lyckas. Och därför kan jag inte göra det (i denna hypotetiska mening av "kan", som möjligen är, åtminstone en del av, den relevanta).

Men varför spelar det någon roll? Eftersom vinnarlotten ännu inte är såld, så är det ju ändå *möjligt* att jag köper den. Varför skulle "bör" implicera "kan" i någon starkare mening? (Faktum är att jag i levande livet har stött på två utilitarister som anser att "bör" inte implicerar "kan" i någon starkare mening, och att jag i exemplet alltså bör köpa vinnarlotten. Men jag är övertygad om att de flesta utilitarister, liksom Moore i sin klassiska framställning av detta i *Ethics*, anser att "bör" implicerar mer än bara "möjlig").

Svaret måste väl vara att om "bör" bara implicerar "möjlig", så skulle moralens normer i många fall sakna *praktisk betydelse*, på så vis att de inte skulle ge någon som helst *vägledning* till agenten. Detta torde vara skälet till att man inte kan hävda att jag bör köpa den vinnande lotten, trots att denna handling har de bästa konsekvenserna (av alla möjliga handlingar). Och detta visar i sin tur att även riktighetskriterier – och inte bara beslutsmetoder – måste ha någon sorts vägledande funktion.

En utilitarist skulle emellertid kunna invända mot detta resonemang

att man med hjälp av sin moraliska intuition helt enkelt kan *inse* att "bör" implicerar "kan" i en starkare betydelse än "möjlig", och att detta inte alls beror på att riktighetskriterier (eller moraliska principer) måste ha en vägledande funktion.

Men detta hjälper inte. Ty vi kan då anta att lotteriet i mitt exempel är sådant att jag kan välja lottnummer när jag köper en lott, och att den lott som kommer att vinna har nummer 4711. Utan tvekan *kan* jag nu köpa den vinnande lotten (dvs lott nr 4711), i den meningen att om jag försöker så lyckas jag. Utilitaristen måste väl då anse att jag *bör* göra det (givet att denna handling har de bästa totala konsekvenserna). Men i praktiken är jag ju precis lika hjälplös som om jag skulle plocka en lott på måfå ur en tombola. Ty jag *vet* ju inte, och *kan* inte veta, vilken lott som kommer att vinna. Detta gäller i bägge fallen. Det är därför, ur moralisk synpunkt, *lika orimligt* att hävda att jag (i stället för att gå på bio) bör köpa lott 4711, som att jag bör köpa den vinnande lotten. Den princip som påbjuder detta har ingen vägledande funktion, på grund av att agenten inte kan veta vilken handling den påbjuder i det enskilda fallet.

I detta läge skulle man kanske vilja hävda att "bör" inte bara implicerar "kan", utan dessutom "kan veta att man bör". Men det vore att gå för långt (bland annat kanske för att det skulle förutsätta att moraliska normer kan vara sanna). Däremot skulle jag vilja föreslå att "bör" implicerar "det finns någon egenskap hos handlingen, som gör att den bör utföras, och man kan veta, eller åtminstone ha förnuftiga skäl att tro, att handlingen har denna egenskap". (Sven Danielsson har frågat mig vem denna "man" i min princip är. Den jag främst hade i tankarna var den som kan utföra handlingen, men kanske borde man försvaga principen genom att ersätta "man" med "någon".)

Men om min princip är riktig, så är utilitarismen felaktig! Ty i själva verket vet man väl *aldrig* vilket alternativ som har de bästa totala konsekvenserna, och trots det menar ju Moore och andra utilitarister att man ändå bör utföra just det.

Utilitarismen är alltså inte ett acceptabelt riktighetskriterium. Det är inte heller sant att om man *vet* att en viss handling har bättre konsekvenser än varje alternativ, så bör man göra den – ty dess optimalitet kan bero på att man i framtiden kommer att göra något som man inte borde ha gjort (se min uppsats "Utilitarianism and future mistakes", *Theoria* vol 43, 1977).

4. En mer modest utilitaristism

En utilitarist behöver dock inte hänga läpp över detta. Jag föreslår att han eller hon i stället retirerar till att betrakta utilitarismen – eller någon mer tillämpar variant av den – som en beslutsmetod. Eller som en del av en beslutsmetod. Eller som en tumregel. Eller som en moralregel bland andra.

Men hur skall då utilitarismen tolkas för att alltså kunna tillämpas i praktiken? Jag skall inte här gå närmare in på detta. Jag föreställer mig att man till en del helt enkelt kan överta de beslutsregler som brukar figurera inom beslutsteori – "maximin", "minimax risk", "maximering av förväntat värde", osv. Det ur praktisk synpunkt kanske största problemet med dessa regler är att de för sin tillämpning förutsätter att man redan har tillgång till en strukturering av beslutssituationen i form av en matris, där alternativ, möjliga utfall och dessas värden på någon lämplig skala är givna. Hur man fastställer en relevant matris säger däremot beslutsteoretiker inte mycket om. Ur utilitaristisk synpunkt bör naturligtvis värderingen av utfall göras på ett *opartiskt* sätt, utfallen får inte begränsas till något speciellt perspektiv – de bör i princip inkludera *alla relevanta* konsekvenser – och i den mån sannolikheter skall vägas in, så bör de vara *objektiva* (i någon mening) snarare än subjektiva. Detta är en mycket stor beställning. I praktiken får man väl nöja sig med att konstruera någon matris som åtminstone inte är *uppenbart* vilseledande eller partisk. Sedan kan man t ex säga att en handling är moraliskt rätt om den maximerar förväntat värde relativt till en sådan matris – såvida det inte finns andra, starkare skäl för att den är moraliskt orätt.

Kanske kunde man också tänka sig en regel av följande slag: Om de *förutsägbara* konsekvenserna av en handling är bättre än de förutsägbara konsekvenserna av varje *rimligt* alternativ, så bör handlingen utföras – såvida det inte finns andra, starkare skäl som talar mot detta. Detta kan ju sägas vara en utilitaristisk regel, och den kan antagligen ge en viss vägledning i många situationer. Dessutom är det en regel som troligen skulle kunna accepteras av de allra flesta.

Skall man betrakta en sådan regel som ett "riktighetskriterium" eller som en "beslutsmetod" eller som något som varken är det ena eller det andra? Vilket *ord* man här väljer kan knappast spela någon roll. Det viktiga är att man inser att utilitaristiska handlingsregler inte identifierar någon precis, empiriskt avgörbar egenskap som är gemensam

och specifik för alla handlingar som bör utföras. De kräver för sin tillämpning subjektiva bedömningar, t ex vad gäller interpersonella välfärdsjämförelser, relevansen hos givna beslutsmatriser och den relativa resistensen mot icke-utilitaristiska synpunkter. Helst bör de nog betraktas som moraliska tumregler, som kan komma i konflikt med andra moraliska tumregler. För egen del anser jag att de som sådana bör väga tungt.

5. Måste tumregler baseras på ett riktighetskriterium?

Om utilitarismen i sin normativt rimligaste utformning består av en samling tumregler – kanske bland andra – måste det då inte finnas något moraliskt riktighetskriterium i bakgrunden, något som utgör grunden eller motiveringen för dessa tumregler? Hänger de inte annars i luften?

Jag tror inte det. Jag tror snarare att detta är en utilitaristisk fördom – som delvis sammanhänger med svårigheterna att tillämpa det utilitaristiska riktighetskriteriet. Den utilitaristiska uppfattningen om förhållandet mellan riktighetskriteriet och våra mer konventionella moralregler (tumregler) är problematisk. Vilken är egentligen denna uppfattning?

(a) Vi har ju ingen som helst anledning att tro att tumreglerna alltid eller oftast eller ens någonsin pekar ut just de handlingar som det utilitaristiska riktighetskriteriet pekar ut. Tvärtom, att döma av de vanliga normativa invändningarna mot utilitarismen.

(b) Vi har inte heller någon anledning att tro att de mer konventionella tumreglerna utgör ett regelsystem som har högre acceptansvärde än varje alternativt regelsystem. (Ett regelsystems acceptansvärde är värdet av konsekvenserna av att alla accepterar systemet och försöker tillämpa det i sitt handlande.) Om det är svårt att veta att en given handling har bättre konsekvenser än varje alternativ, så är det säkerligen ännu mycket svårare att veta att ett regelsystem har maximalt acceptansvärde. Man kan inte heller förvänta sig att evolutionen har sett till att våra vanliga regler har maximalt acceptansvärde. Evolutionen har nog inte haft tillräckligt lång tid på sig i mänskliga samhällen, och dessa har oftast varit konservativa och föga experimentella.

(c) Inte heller har vi någon anledning att tro att ett system av mer uttalat utilitaristiska tumregler – av det slag som antytts ovan – skulle ha maximalt acceptansvärde. Därtill är de bland annat alltför obestämda.

(d) Även om ett visst regelsystem skulle ha ett maximalt acceptansvärde, så är det ingalunda uppenbart att man därför bör tillämpa det i enskilda fall. Ty konsekvenserna av att göra det kan bero på om även andra tillämpar systemet. Gör de inte det, så blir kanske konsekvenserna bättre om man själv inte heller gör det.

(e) Torbjörn Tännsjö tycks anse att om ett visst regelsystem har maximalt acceptansvärde *för en agent* (dvs om konsekvenserna av att agenten tillämpar dem konsekvent i alla situationer han hamnar blir bättre eller minst lika bra som de konsekvenser som skulle uppstå om han konsekvent tillämpade något annat regelsystem), så är det *rationellt* för denna agent att tillämpa detta regelsystem i enskilda fall, *om* han accepterar det utilitaristiska riktighetskriteriet. Detta är inte särskilt övertygande. Vad som är rationellt för agenten i en enskild situation tycks bero på om han tillämpar regelsystemet i andra fall. Och även om han skulle göra det, så kan det ju mycket väl vara så att han enligt det utilitaristiska riktighetskriteriet inte bör tillämpa systemet i ett enskilt fall. Hur skall han då bete sig? Hur skall man bete sig om ett alternativ är rationellt och ett annat är moraliskt riktigt? Följande är inte något bra svar: man bör bete sig rationellt, om man vill bete sig rationellt, och man bör bete sig moraliskt riktigt, om man vill bete sig moraliskt riktigt. Ett rimligt moralsystem borde ställa större krav än så.

(f) Man kan möjligen ha skäl att tro att många konventionella moraliska regler har ett *visst* positivt acceptansvärde. Ty det förefaller som om de skulle kunna underlätta social samvaro. Och det är väl delvis därför vi lär ut dem till våra barn. Detta är en sorts utilitaristiskt skäl, men det är inte av maximeringstyp.

(g) På liknande sätt tror vi väl att mer direkt utilitaristiska regler också har ett visst positivt acceptansvärde. Det är svårt att säga om det är högre eller lägre än för konventionella regler. Sidgwick trodde väl att det är lägre, medan den mer sentida utilitaristen J J C Smart tycks mena att det är högre. Men här kan vi knappast ha mer än fromma förhoppningar.

(h) Hur som helst tycks det vara så – om man bortser från fall (a) ovan – att det utilitaristiska riktighetskriteriet överhuvud taget inte har någon funktion att fylla när det gäller att motivera en beslutsmetod eller moraliska tumregler. Vad som spelar roll är snarare acceptansvärde eller något liknande. Men en motivering av tumregler i termer

av acceptansvärde blir inte mer vägande om man därutöver lägger till det utilitaristiska riktighetskriteriet. Snarare tvärt om.

Jag anser med andra ord att moraliska beslutsmetoder inte behöver stöd av något riktighetskriterium – åtminstone inte på det sätt som utilitarister tycks tänka sig. Att utilitarister gör ett så stort nummer av distinktionen mellan riktighetskriterium och beslutsmetod beror väl dels på att just deras riktighetskriterium saknar praktisk användning, och dels på att distinktionen tycks kunna användas för att bemöta de vanligaste formerna av moralisk kritik mot utilitarismen. Denna kritik brukar nämligen bemötas genom påpekandet att det utilitaristiska riktighetskriteriet inte skall tillämpas i praktiken, utan att man i stället bör följa mer konventionella moralregler (eller liknande), vilka antas ha högt acceptansvärde. (Här hänvisar man till empiriska hypoteser som är ytterst svårkontrollerade.) Men ett sådant försvar för utilitarismen innebär ju att man i själva verket gör den immun mot moralisk kritik. Detta är knappast mer än ett retoriskt knep.

Jag vill dessutom tillägga, att tron på att det finns ett moraliskt riktighetskriterium – en bestämd egenskap, som är gemensam och specifik för alla handlingar som ur moralisk synpunkt bör utföras – i sig själv mycket dubiös. Såtillvida kan den jämföras med den lika tveksamma (men kanske också lika vanliga) åsikten att det finns ett *sanningskriterium*, en bestämd egenskap som är gemensam och specifik för alla sanna satser eller för alla sanna påståenden. Varför skulle man tro det?

6. Moralens grund

Kontentan av ovanstående är att vi knappast har något behov av ett moraliskt riktighetskriterium som inte med viss lätthet kan tillämpas i praktiken. Vi klarar oss lika bra med enbart tumregler, eller med andra ord med moraliska regler av mer konventionellt slag.

Men på vilken grund vilar då dessa regler? Såvitt jag kan se behöver de ingen grund. Det finns säkerligen *orsaker* till att vi tror på de regler vi tror på. Men det behöver inte finnas några mer bakomliggande *skäl* till att vi tror på dem.

Det finns inga *andra* regler som implicerar att vi bör acceptera just våra regler. Det finns inte heller några *mer generella* regler från vilka våra regler kan härledas. Det kanske inte heller finns några generella regler om hur regelkonflikter bör lösas – bortsett från den självklara

regeln att viktigare regler skall väga tyngre än mindre viktiga.

Och även om det funnes mer fundamentala regler, som utgjorde en grund för våra vanliga moralregler, så skulle det inte göra någon större principiell skillnad. Ty i så fall kunde man på samma sätt fråga sig om *de* kan accepteras, utan att ha en grund i ännu mer fundamentala regler, osv.

Därför kan vi lika gärna godta de *regler som vi faktiskt tror på* (och som vi är benägna att hålla fast vid även om de utsätts för kritik) som moralens grund. Liksom i vetenskap har vi inget annat val än att utgå från det vi faktiskt tror på. Otto Neurath beskrev situationen i en berömd liknelse: vi är som sjömän i en båt på öppna havet; det gäller att hålla sig flytande och om vi behöver reparera båten (dvs modifiera våra åsikter), så måste det göras gradvis, medan det mesta behålls intakt – annars går vi under helt och hållet. Att försöka att hitta en oberoende grund för moralen, från vilken man kan bedöma våra faktiska uppfattningar, är lika orimligt som att konstruera en "första filosofi" som grund för vetenskapen.

Till dessa våra vanliga moralregler hör sådant som: man bör hålla löften och avtal, tala sanning, hjälpa folk i nöd, följa demokratiskt fattade lagar, glädja sina medmänniskor, skydda och uppfostra sina barn, inte skattefuska, inte ta emot socialbidrag som man inte behöver eller som man inte har rätt till, inte sparka folk som ligger försvarslösa på marken, behandla andra som man själv vill bli behandlad, respektera andras autonomi, inte mobba arbetskamrater, inte bära falsk vittnesbörd mot sin nästa, behandla djur humant, osv. Vi uppfattar också olika regler som mer eller mindre viktiga – att hjälpa folk i svår nöd är t ex viktigare än att hålla löften i mer triviala sammanhang. Reglerna kan naturligtvis komma i konflikt, och då får man bedöma deras relativa viktighet.

Man kunde kanske vilja hävda att när olika regler kommer i konflikt, så skall man *enbart* lägga vikt vid utilitaristiska överväganden. Men det vore nog ingen bra regel. Ty utilitaristiska överväganden kan behöva kompletteras med andra regler. Flera av de vanliga motexemplen mot utilitarism visar just detta.

Att "moralens grund" alltså inte är något annat än den moral vi faktiskt godtar innebär dock inte att vi nu har bundit oss för ett nytt, *konventionalistiskt* riktighetskriterium, som innebär att en handling bör utföras om och endast om den är i överensstämmelse med de moral-

regler som accepteras av agenten (eller i agentens samhälle, eller i vårt samhälle). Det inser man om man betänker att denna princip *inte* tillhör det som vi faktiskt tror på.

7. Svåra fall

I vissa fall pekar våra regler som sagt åt olika håll. De kommer i konflikt. Sådana konflikter kan ibland lösas genom att vi tar hänsyn till de olika reglernas relativa viktighet. I vissa fall löser dock olika bedömare konflikten på olika sätt, och ibland kan vi kanske inte hitta några lösningar alls.

Att en uppsättning regler, eller värden, inte är tillräcklig för att garantera ett bestämt resultat i varje enskilt fall är emellertid inte något speciellt för moral. Det gäller även i *estetik* (ett konstverk är bättre i vissa avseenden men sämre i andra), *juridik* (vissa skäl talar för friande dom, andra för fällande, etc), *vetenskap* (en teori är bättre i vissa avseenden men sämre i andra), *texttolkning* (evidens av olika slag kan peka åt olika håll), *socialpolitik* (ett lagförslag är bättre i vissa avseenden men sämre i andra) och för övrigt i en mängd olika fall av *klassifikation* (det kan finnas olöslig oenighet om huruvida något är komiskt, dramatiskt, demokratiskt, rättvist, ett etablerat faktum, sentimentalt, riskabelt, godkänt, välgrundat, kunskap, stort, kallt, tungt, meningslöst, säker slutförvaring, tillförlitligt, ett lyckligt liv, trafikfarligt, etc).

Man kan fråga sig om det i dessa svåra fall (*hard cases*) finns en entydig *sanning*, eller om det snarare är fråga om sanningsvärdesluckor där våra påståenden varken är sanna eller falska. Detta är en filosofiskt spännande fråga, men inget speciellt för moralfilosofi, och ingenting som bör få bli avgörande för frågan om utilitarismen är sann eller falsk. Även om det skulle vara så att det utilitaristiska riktighetskriteriet ger en lösning av alla svåra fall – vilket bland annat förutsätter att problemet om interpersonella jämförelser får en lösning – så kan detta knappast användas som ett argument för att utilitarismen är sann.

KRISTER BYKVIST

Är de flesta utilitarister deontologer?
Om preferentialismens relativisering av värde

1. Introduktion

Det är populärt att hävda att den viktigaste skillnaden mellan deontologer och konsekventialister är att de förra men inte de senare kan ta *agentrelativa* värden eller mål i beaktande. (Se t ex Nagel (1986) kap 8 och 9, Pettit (1993) s xii-xiv och Scheffler (1987).) Till de agentrelativa mål som deontologen vill försvara hör agentrelativa *restriktioner*, dvs vi är förbjudna att utföra vissa typer av handlingar trots att de skulle maximera värdet i en viss situation, och agentrelativa *friheter*, dvs vi är ibland fria, men inte förpliktade, att välja den handling som maximerar det goda. Det är agentens intima relation till sina egna handlingar som är det viktiga här. Eftersom det är omöjligt att, utan att göra våld på begreppet konsekvens, konstruera en agentneutral konsekventialism som inkorporerar dessa deontologiska intuitioner, drar vissa moralfilosofer slutsatsen att vi måste överge den agentneutrala konsekventialismen. Inte nödvändigtvis så att vi ska bli renodlade deontologer. Vi kan hålla fast vid att utfallens värde bestämmer handlingens normativa status. Däremot måste vi överge kravet på agentneutralitet. Om vi inför agentrelativa värderingar av utfallen, så kan vi fånga några av de viktigaste deontologiska intuitionerna.

Vad jag vill visa är att vi har skäl att införa agentrelativa värderingar för att kunna fånga vissa viktiga och vanliga *utilitaristiska* intuitioner. Detta får konsekvenser för hur vi ska se på distinktionen mellan konsekventialism och deontologi. En definition av konsekventialism som utesluter agentrelativitet tycks vara alltför snäv då den inte rymmer några av de vanligaste formerna av utilitarism.

2. *Agentrelativitet*

Ibland sägs det att en teori är agentrelativ om den ger olika mål till olika agenter. En teori som ger olika agenter samma mål sägs då vara agentneutral. Visst ger denna karakterisering ett hum om vad agentrelativitet och agentneutralitet handlar om. Men vi kan inte nöja oss med denna oklara formulering. En något bättre formulering ges i Samuel Schefflers ofta citerade definition av konsekventialism:

> [den specificerar] en princip som rangordnar totala sakförhållanden från det bästa till det sämsta från en opersonlig ståndpunkt. Med andra ord, de rangordningar som genereras av den utsedda principen är inte agentrelativa; *de varierar inte från en person till en annan beroende på vilken bestämd situation man befinner sig i* (Scheffler (1987) s 1, min översättning och kursivering).

Inspirerade av Scheffler skulle vi kunna säga att en teori är *agentrelativ* om och endast om den implicerar att rangordningen av en viss mängd utfall varierar från en agent till en annan. En rangordning av en mängd utfall varierar om något utfall byter plats i rangordningen. En teori är då *agentneutral* om och endast om den implicerar att rangordningen av en mängd utfall måste vara densamma för alla agenter.

Kanske bör vi inte trycka så hårt på att det ska vara en annan *agent* för att vi ska få den relevanta agentrelativiteten. Det är väl relativiseringen till olika *agentpositioner* och inte till olika personer som är det viktiga. En agentposition kan vi då identifiera med en tidpunkt i en värld och en mängd alternativa handlingar. En person kan inte vara olika agenter, men en och samma agent kan vara vid olika agentpositioner.

3. *Preferentialism och agentrelativitet*

Många moderna utilitarister är preferensutilitarister. Deras axiologi, *preferentialismen*, är den teori som säger att värdet av ett utfall uteslutande baseras på preferenser. Om vi kan visa att preferentialismen är agentrelativ, så har vi alltså visat att många moderna utilitarister förespråkar en teori som är agentrelativ. Är nu preferentialismen agentrelativ i den mening som jag skisserat ovan? Något klart besked får vi först om vi tittar lite noggrannare på vad som menas med preferentialism. Som påpekats av Wlodek Rabinowicz, Bertil Ström-

berg och Jan Österberg klyver sig preferentialismen i två huvudgrupper alltefter vad som ses som värdebärare. Enligt den ena huvudgruppen, som vi kan kalla *objektpreferentialism,* är det det önskade eller oönskade objektet som har värde. Ofta lägger man här till ett antagande om värdets uppkomst som säger att det är *genom* att vara önskat eller oönskat som objektet *får* värde. Enligt den andra huvudgruppen, som vi i brist på bättre kan kalla *tillfredsställelsepreferentialism,* är det i stället *sakförhållandet att* en preferens är tillfredsställd eller frustrerad som har värde. Skillnaden mellan dessa huvudgrupper kan enkelt belysas med följande exempel. Anta – för att hämta ett fall från Hägerström – att jag önskar att jag får en tunna sill. Att denna önskan är tillfredsställd innebär att det föreligger ett sammansatt sakförhållande som består av sakförhållandet att jag önskar att jag får en tunna sill *och* sakförhållandet att jag får en tunna sill. Enligt tillfredsställelsepreferentialismen har *hela* detta sammansatta sakförhållande värde medan enligt objektpreferentialismen så är det endast en *del* som har värde, nämligen sakförhållandet att jag får en tunna sill.

4. Objektpreferentialism

Nu tillbaka till frågan om preferentialism och agentrelativitet. Jag vill hävda att *alla* rimliga former av objektpreferentialism och *många* rimliga former av tillfredsställelsepreferentialism är agentrelativa. Jag börjar med objektpreferentialismen.

Anta att jag just nu kan springa ut på torget och köpa en tunna sill. Värdet av att jag får en tunna sill bestäms av preferenser, men vilka? Är vi riktigt generösa så tittar vi på alla *möjliga* preferenser, både faktiska och blott möjliga, som är riktade mot sakförhållandet att jag får en tunna sill. Men detta tycks betyda att sakförhållandet att jag får en tunna sill varken är bra eller dåligt, eftersom man för varje möjligt gillande med en viss styrka av att jag får en tunna sill kan hitta ett möjligt ogillande med samma styrka av att jag får en tunna sill, och det positiva värde som gillandet ger uppvägs av det negativa värde som ogillandet ger. Nu kanske man kan hålla med om att sakförhållandet att jag får en tunna sill varken är bra eller dåligt, men eftersom värderingsmetoden är tänkt att vara generell och det gäller för *varje* sakförhållande p att man för varje möjligt ogillande av p hittar ett möjligt gillande av p som har samma styrka, så får vi det absurda

resultatet att inget sakförhållanden är bra och inget sakförhållande är dåligt. (Detta har påpekats av Wlodek Rabinowicz och Jan Österberg.) Observera att dessa värderingar gäller oberoende av vilken agentposition vi intar; alltså är denna objektpreferentialism agentneutral. Agentneutraliteten har dock köpts till ett alldeles för högt pris. Vi måste alltså begränsa mängden värdegrundande preferenser för att får en rimlig objektpreferentialism.

En radikal begränsning, som föreslagits av bland andra Wlodek Rabinowicz, är att man ska tillämpa en form av *aktualism* och bara utgå från de preferenser som förekommer i den aktuella världen. Värdet av sakförhållandet att jag får en tunna sill bestäms endast av de *aktuella* preferenser som är riktade mot sakförhållandet ifråga. Om vi dessutom inte gör någon skillnad i när en preferens föreligger utan räknar med förflutna, nuvarande och framtida aktuella preferenser undviker vi agentrelativism. Värdet av att jag får en tunna sill fixeras av de faktiskt existerande förflutna, nuvarande och framtida preferenser som är riktade mot sakförhållandet i fråga. Det är klart att detta värde inte varierar över agentpositioner.

Men begränsningen till de aktuella preferenserna ger oss en oacceptabel "aktualistisk diktatur". För plocka ut ett sakförhållande p som någon i den aktuella världen önskar. Tänk på en annan värld där p föreligger och alla personer (eller djur) har så gott som alla preferenser frustrerade. Kontrastera denna helvetesvärld med ytterligare en annan tänkt värld – en paradisvärld – i vilken p inte föreligger men alla personer har så gott som alla preferenser tillfredsställda. Anta dessutom att förutom p så gäller det för varje sakförhållande som någon aktuell person önskar att antingen föreligger detta i både helvetesvärlden och i paradisvärlden eller så föreligger detta varken i helvetesvärlden eller i paradisvärlden. Om vi nu tänker oss en situation där agenten kan välja att realisera helvetesvärlden eller paradisvärlden, så säger den aktualistiska objektpreferentialismen att agenten ska värdera helvetet högre än paradiset. Varför? Enligt denna teori är ju den enda relevanta skillnaden mellan helvetet och paradiset att det finns ett aktuellt önskat sakförhållande p som föreligger i helvetet men inte i paradiset. Det är endast beroende på denna skillnad som helvetet får högre värde än paradiset. Att helvetet verkligen innebär ett helvete för dem som lever där och att paradiset innebär ett paradis för dem som lever där har ingen direkt betydelse för värderingen av dessa världar.

Detta är oacceptabelt. (Observera att jag här inte behöver förutsätta att illvilliga preferenser ska beaktas. *p* behöver inte vara ett sakförhållande att någon lider.) Således är även här priset för agentneutralitet alldeles för högt.

Man skulle kunna invända att värdet av en värld *v* ska bestämmas av de preferenser som är aktuella i *v*. I helvetesvärlden är de aktuella preferenserna de preferenser som existerar där, och värdet av världen ska därför bestämmas av dessa preferenser. I paradisvärlden är de aktuella preferenserna de preferenser som existerar där, och värdet av denna värld ska därför bestämmas av dessa preferenser. Men om "aktuella preferenser" förstås på detta relativa sätt, så är det ju klart att värdet kommer att variera över världar. För anta att i en värld *v* föredrar alla *p* framför *q* och att i en annan värld *v´* föredrar alla *q* framför *p*. Då kommer *p* att vara bättre än *q i världen v,* eftersom alla aktuella preferenser *i världen v* är till *p*'s fördel. Men *q* kommer att vara bättre än *p* i *världen v´*, eftersom alla aktuella preferenser *i världen v´* är till *q*'s fördel. Om man vill undvika att värdet varierar kan man alltså inte tolka de aktuella preferenserna på detta relativa sätt.

En annan invändning är att vi faktiskt här i den aktuella världen bryr oss om andra människors lidande och lycka. Vi önskar inte andra ont, vi önskar dem gott. Bortsett från att detta inte kan gälla alla aktuella personer – alla aktuella personer är ju inte änglar – och att de som bryr sig om andras väl och ve bryr sig betydligt mer om den egna välfärden, så är detta ett misslyckat sätt att försvara objektpreferentialismen. Vi vill ju inte säga att det negativa värdet av helvetesvärlden är avhängigt av om någon *annan* bryr sig om deras lidande. Att personerna i helvetesvärlden har nästan samtliga preferenser frustrerade måste ha någon *direkt* påverkan på värdet av världen.

Vi har alltså sett att vare sig vi når agentneutralitet genom att beakta alla möjliga preferenser eller genom att endast beakta de aktuella preferenserna så är kostnaderna för stora. Jag kan nu inte se hur man skulle kunna uppnå agentneutralitet genom att begränsa mängden av värdegrundande preferenser på något annat sätt utan att vi därmed lämnar preferentialismens centrala idé att värde ska baseras endast på preferenser. Därför tror jag att vi kan slå fast att inte någon av de rimliga formerna av objektpreferentialism kan undkomma agentrelativitet.

5. *Tillfredsställelsepreferentialism*

För att en tillfredsställelsepreferentialist ska kunna bestämma värdet av de alternativa utfallen måste han besvara frågan: Vilka tillfredsställelser och frustrationer ska vara med och bestämma värdet hos utfallen? Svarar han "alla tillfredsställelser och frustrationer som förekommer i utfallet!", så avslöjar han sig vara en agentneutral preferentialist. Värdet av ett sakförhållande varierar inte över agenter i hans teori, eftersom agentpositionen inte kan påverka vilka tillfredsställelser och frustrationer sakförhållandet innehåller. Men många tycker att denne preferentialist visar en alldeles för tillåtande attityd. Vissa typer av preferenser bör diskvalificeras. I det följande ska jag ta upp två kriterier som kan användas för att diskvalificera preferenser. Enligt det första kriteriet, *det temporala kriteriet*, är preferensernas förhållande till *tiden* viktig. Vi bör vara tidspartiska i någon mening. Enligt det andra kriteriet, *det modala kriteriet*, är det viktigt att skilja på preferenser vars existens är avhängig av våra val och de preferenser som existerar oavsett vad vi gör.

Den tidspartiskhet som det temporala kriteriet påbjuder kan yttra sig på lite olika sätt.

(1) Man kan vara tidspartisk vad gäller *preferensens* datering. Här kan vi skilja mellan *futuristen*, som inte kan se några skäl att tillfredsställa preferenser som har funnits men som inte längre finns, och som därför anser att endast nuvarande och framtida preferenser är värdegrundande, och *presentisten*, som går ett steg till och inte ser varför man ska tillfredsställa preferenser som kommer att finnas men som inte finns nu, och som därför menar att endast nuvarande preferenser är värdegrundande.

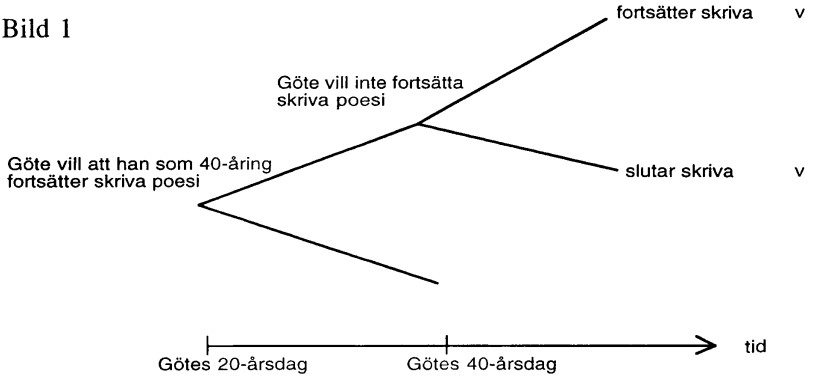
(2) Man kan vara tidspartisk vad gäller *objektets* datering. Även här kan vi skilja mellan presentisten, som inriktar sig på det intresse som finns för nuet och som därför menar att endast preferenser vars objekt föreligger i nuet är värdegrundande, och futuristen som inriktar sig på det intresse som finns för nuet och framtiden och därför menar att endast preferenser vars objekt ligger i nuet eller i framtiden är värdegrundande.

Vad jag nu vill visa är att alla tillfredsställelsepreferentialismer som är tidspartiska vad gäller preferensdatering men som är tidsneutrala vad gäller objektdatering är agentrelativa. Med "alla" menar jag de

presentistiska och futuristiska versionerna. Observera att dessa teorier inte är uppenbart orimliga. De säger något i stil med att vi bör göra det så bra som möjligt för de personer som existerar nu (presentisten) eller de som existerar nu och i framtiden (futuristen). Att göra det bra för en person innebär att tillfredsställa hans nuvarande eller framtida preferenser, oavsett vilken "riktning" i tiden dessa preferenser har.

Jag låter exemplet med poeten Göte illustrera de tidspartiska preferentialismernas agentrelativitet. (Se bild 1).

Bild 1



Göte, en ung poet, vaknar upp på morgonen till sin tjugonde födelsedag. Han håller på att avsluta diktsviten "Det mänskliga träskets sång". Även denna morgon känner han sitt poetiska kall men han vänder nu blicken även framåt och önskar intensivt att han som 40-åring fortfarande skriver poesi. Han är så frälst på poesi att han anser att även om han som 40-åring inte vill skriva så bör han ändå göra det. Det finns inget annat som kan göra livet så värdefullt. Det viktiga är dock inte att man lyckas skriva mästerverk utan att man arbetar med poesi och försöker skriva. Anta att v1 och v2 kan realiseras av både den 20-årige och den 40-årige Göte. v1 är en värld i vilken han mellan 20 och 40 skriver poesi och efter 40 fortsätter att skriva poesi. v2 är en värld där han mellan 20 och 40 skriver poesi men där han som 40-åring slutar skriva poesi. Anta nu att Göte som 40-åring i v1/v2 kommer att sakna all lust att skriva poesi. Han tycker att det är helt OK att andra håller på med det men han kan *själv* inte finna någon glädje i det.

Anta att den framåtblickande preferens Göte har som 20-åring för att skriva poesi när han är 40 är starkare än den preferens Göte har som 40-åring mot att fortsätta skriva poesi. Anta slutligen att vad

gäller alla andra preferenser så är v1 och v2 likvärdiga, dvs för varje annan preferens som förekommer i både v1 och v2 så gäller det att den är tillfredsställd i båda världarna eller frustrerad i båda världarna.

Om vi utgår från Götes agentposition på hans 20-årsdag, så ger en *presentistisk* (vad gäller preferensdateringen) tillfredsställelsepreferentialism resultatet att v1 är bättre än v2. v1 innehåller ju tillfredsställelsen av en present preferens medan v2 innehåller frustrationen av densamma, och förutom den 40-årige Götes motvilja för poesiskrivande som inte ska räknas, så är v1 och v2 likvärdiga för alla andra preferenser. Om vi utgår från Götes agentposition på hans 40-årsdag (i världen v1/v2), så ger dock teorin resultatet att v2 är bättre än v1. För vid Götes 40-årsdag så hör Götes tidigare poesipreferens till det förflutna, och ska därför inte räknas längre. Relativt 40-årsdagen innehåller v2 tillfredsställelsen av en nuvarande preferens medan v1 innehåller frustrationen av densamma och för övrigt innehåller v1 och v2 exakt samma tillfredsställelser och frustrationer. Alltså är denna presentistiska preferentialism agentrelativ.

Vad säger en *futuristisk* (vad gäller preferensdateringen) tillfredsställelsepreferentialism om värderingen av dessa världar? Om vi utgår från Götes agentposition vid hans 20-årsdag, så säger teorin att v1 är bättre än v2. v1 innehåller tillfredsställelsen av en nuvarande preferens och frustrationen av en framtida preferens medan v2 innehåller frustrationen av en nuvarande preferens och tillfredsställelsen av en framtida, men eftersom den nuvarande preferensen är starkare än den framtida och världarna är likvärdiga för alla andra preferenser, så är v1 bättre än v2. Om vi i stället utgår från Götes agentposition vid hans 40-årsdag, ger teorin resultatet att v2 är bättre än v1, eftersom relativt denna agentposition är den unge Götes poesipreferens förfluten och därmed överspelad. Den enda preferens vi har att ta i beaktande här är den äldre Götes preferens för att inte fortsätta skriva poesi, och denna preferens är tillfredsställd i v2 och frustrerad i v1. Alltså är också denna futuristiska preferentialism agentrelativ.

Vi ser alltså att även tillfredsställelsepreferentialismen råkar ut för agentrelativitet om den omfattar en viss typ av tidspartiskhet; en tidspartiskhet som inte alls verkar orimlig.

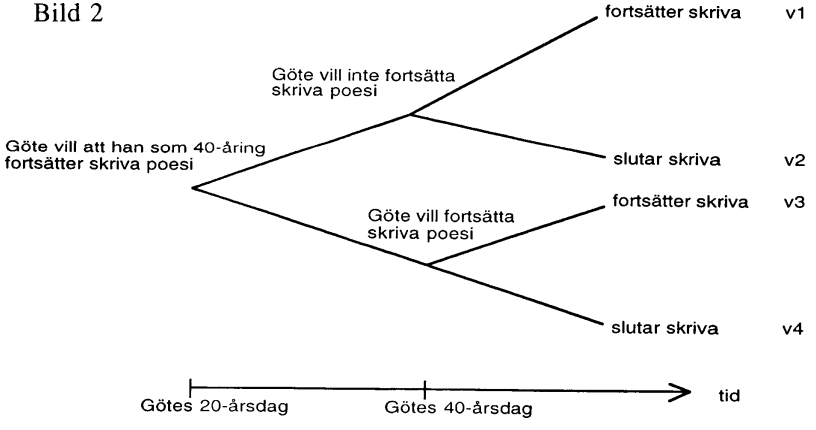
Nu kan man ju ändå tycka att tidspartiskhet är något konstigt och att en rimlig tillfredsställelsepreferentialism ska vara tidsneutral i båda avseendena. Men även en tidsneutral tillfredsställelsepreferentialism

råkar ut för agentrelativitet om den är modalt partisk. Ni minns att det modala kriteriet skiljer ut preferenser alltefter som de är kontingenta eller inte. Gemensamt för de modalt partiska teorierna är att de hävdar att ett tillstånd inte kan vara bättre (sämre) än något annat om det inte är bättre (sämre) *för* någon. En variant av modal partiskhet menar att detta betyder att endast *nödvändiga preferenser*, sådana som är oundvikliga i den meningen att de existerar oavsett vad vi gör, är värdegrundande. Vi gör det inte bättre för någon, och därmed inte heller världen bättre, genom att skapa nya preferenser och tillfredsställa dem. Vi ska istället inrikta oss på att tillfredsställa nödvändiga preferenser.

En annan variant – och en populär sådan bland populationsetiker – är vad man brukar kalla "person-affecting theory", enligt vilken man inte kan göra det bättre för en person genom att skapa henne. Vi kan bara göra det bättre för *nödvändiga personer*, sådana personer som existerar oavsett vad vi gör. Om "göra det bättre för nödvändiga personer" förstås som att tillfredsställa deras nödvändiga preferenser är vi tillbaka till teorin som värnar om oundvikliga preferenser. Men vi kan stoppa denna reduktion genom att säga att vi kan göra det bättre för nödvändiga personer genom att *skapa nya* preferenser åt dem och tillfredsställa dessa. Gemensamt för dessa teorier är dock att vi inte gör världen bättre bara genom att skapa nya lyckliga människor.

Oavsett vilken form av modal partiskhet vi väljer så får vi en agentrelativ teori. Låt oss gå tillbaka till poeten Göte, (se bild 2).

Bild 2



Här har vi ändrat exemplet något genom att anta att Göte på sin 20-årsdag kan se till att han fortfarande på sin 40-årsdag är poesifrälst. Vi kan t ex föreställa oss att han kan åka iväg och bosätta sig i ett kollektiv av brinnande poeter och stanna där i 20 år. Resultatet av denna vistelse skulle bli att han på sin 40-årsdag intensivt skulle vilja fortsätta skriva poesi. Detta betyder att den äldre Götes motvilja för att skriva poesi är en kontingent preferens relativt den yngre Götes agentposition. Motviljan finns i endast två av de alternativa världar som det står i den yngre Götes makt att realisera. Anta också att denna motvilja är starkare än den yngre Götes önskan att som 40-åring fortsätta skriva poesi. (v3 är en värld där Göte har åkt till poesikollektivet och han på sin 40-årsdag intensivt vill fortsätta att skriva poesi och faktiskt fortsätter att göra det. v4 är identisk med v3 förutom att den äldre Göte inte fortsätter att skriva.)

Hur värderas v1 och v2 av en teori som är ute efter att tillfredsställa nödvändiga preferenser? Om vi utgår från den yngre Götes agentposition får vi resultatet att v1 är bättre än v2. Den äldre Götes motvilja är ju kontingent relativt den yngre Götes agentposition och ska därför inte räknas utifrån denna position, och den yngre Götes poesipreferens är den enda nödvändiga preferens som berörs av valet mellan v1 och v2. Om vi däremot utgår från den äldre Götes agentposition i världen v1/v2, så får vi resultatet att v2 är bättre än v1. För relativt denna agentposition är ju nu den äldre Götes motvilja en nödvändig preferens. Och eftersom den yngre Götes preferens också är nödvändig men svagare, så får vi resultatet att alla nödvändiga preferenser tagna i beaktande så är v2 bättre än v1.

Vi får samma resultat om vi istället tillämpar "person-affecting theory". Byt ut den äldre Göte i v1/v2 till en Göte junior, och byt ut den äldre Göte i v3/v4 till en *annan* Göte junior. Anta att Göte som 20-åring har valet att skapa en men inte båda av dessa Göte juniors, och att Göte vill att hans barn ska skriva poesi. Anta också, mycket orealistiskt, att inga andra preferenser är berörda än Götarnas olika poesipreferenser.

Resultatet blir alltså att oavsett vilken modal partiskhet vi antar att den tidsneutrala tillfredsställelsepreferentialismen omfattar så får vi agentrelativitet.

6. Relativiserat värde och konsekventialism

Vi har funnit att varje rimlig form av objektpreferentialism och vissa rimliga former av tillfredsställelsepreferentialism, två tidspartiska och två modalt partiska, alla är agentrelativa i den meningen att de ger olika värderingar av samma sakförhållande till olika agenter. Utifrån en agent sett är ett sakförhållande *p* bättre än ett annat sakförhållande *q*, men utifrån en annan agent sett är *q* bättre än *p*. Men innebär inte detta att dessa preferentialismer implicerar en motsägelse? Det är väl omöjligt att *p* är bättre än *q och* att *q* är bättre än *p*. Nog för att vi ofta är oense om vilka strukturvillkor som gäller för relationen "bättre än" men så mycket kan vi väl ena oss om att den måste vara asymmetrisk?

Det enkla svaret är att vi undviker motsägelsen genom att göra relativiseringen till agentpositioner explicit i värderingen. Med andra ord, enligt dessa preferentialismer är den korrekta formen för en "bättre-än"-värdering inte den tvåställiga "*p* är bättre än *q*" utan den treställiga "*p* är bättre än *q utifrån agentpositionen A*". När vi sedan som preferensutilitarister rör oss från axiologin till det normativa, så är det dessa agentrelativa värderingar som ska bestämma den normativa statusen hos handlingarna. Varje agent bör utföra den handling som leder till det utfall som maximerar värdet relativt den agentposition vid vilken agenten befinner sig.

Hur går dessa värderingar ihop med konsekventialism? Inte alls, om vi utgår från Schefflers definition av konsekventialism som jag citerade i inledningen. Enligt Scheffler är konsekventialismer agentneutrala, men det är ju helt klart att de preferensutilitarismer vi gått igenom inte är agentneutrala.

Notera också att dessa preferentialistiska värderingar inte går ihop med konsekventialism tolkad som en teori enligt vilken den normativa statusen hos en handling bestäms enbart av det *intrinsikala* värdet hos dess utfall (och de alternativa handlingarnas utfall). Även om det är oklart hur man mera exakt ska förstå intrinsikalt värde, så torde man kunna säga att det intrinsikala värdet inte kan variera över tidpunkter inom en och samma värld. Det verkar också rimligt att lägga till att det intrinsikala värdet inte kan variera över världar. Nu är det så, om vi ska vara lite petiga, att det värde som de beskrivna preferentialismerna omfattar inte varierar över tidpunkter eller världar, eftersom relativiseringen inte rör tidpunkter eller världar utan rör agentposi-

tioner. Men om det nu överhuvudtaget är möjligt att relativisera intrinsikalt värde till agentpositioner – något som moralfilosofer i Moores anda skulle sätta sig emot – så verkar det rimligt att här liksom vad gäller tidpunkter och världar kräva invarians. Med andra ord, om ett värde varierar över agentpositioner, så kan inte detta värde vara intrinsikalt.

Enligt de preferensutilitarismer vi gått igenom så bestäms den normativa statusen hos en handling av ett värde som varierar över agentpositioner. Men om ett värde varierar över agentpositioner så kan det inte kan vara intrinsikalt. Alltså får vi resultatet att dessa teorier *inte* är former av konsekventialism! Ett något överraskande resultat eftersom så många moderna utilitarister, utan att förneka att de är konsekventialister, bekänner sig till någon av dessa preferensutilitarismer. Var ligger felet? Hos dessa utilitarister, som vi kanske bör kalla "deontologer", om vi förstår en deontolog som en icke-konsekventialist? Eller hos de föreslagna definitionerna av konsekventialism som båda utgår från att värdet är invariant?

Jag tror att vi måste hålla fast vid att intrinsikalt värde inte kan variera över agenter. Även om vi kan vara oense om många strukturvillkor för intrinsikalt värde och om hur den riktiga definitionen lyder, så borde väl vi enas om att *ett* kännetecken för intrinsikalt värde, det värde något har i kraft av sina intrinsikala egenskaper, är att det inte kan variera över agentpositioner. Att ett sakförhållande ligger i framtiden, nuet eller det förflutna, är *extrinsikala* egenskaper hos sakförhållandet, egenskaper det har i kraft av att de står i relation till någon tänkt tidpunkt. Likaså är det att ett sakförhållande innehåller eller konstitueras av tillfredsställelsen av någon kontingent preferens en extrinsikal egenskap hos sakförhållandet som det har i kraft av att stå i en viss relation till någon agent. Det är därför de föreslagna definitionerna av konsekventialism som vi bör förkasta. En definition som är för snäv utifrån våra förteoretiska intuitioner bör vi inte hålla fast vid.

Hur ska vi då skilja på deontologi och konsekventialism, något som de flesta moralfilosofer förutsätter är möjligt? Kanske visar svårigheten att finna en godtagbar distinktion mellan konsekventialism och deontologi att det egentligen är lönlöst att leta efter formella eller strukturella olikheter med vilka man skulle kunna definiera traditionerna. Gemensamt för definitionsförsöken är ju att man försöker finna skillnader som inte ligger direkt på det substantiella planet. Man hävdar att det

är *genom* sina strukturella egenskaper som konsekventialismen utesluter, och deontologin kan omfatta, vissa substantiella värderingar. Kanske ska man i stället se konflikten mellan konsekventialister och deontologer som en konflikt mellan rent substantiella värderingar. Kort sagt: konsekventialismen och deontologin kan stöpas i samma form. Oenigheten rör inte formen utan vad som ska läggas i formen. (Jag vill tacka Erik Carlson och de andra deltagarna på filosofidagarna i Umeå, 1995, för värdefulla synpunkter på en tidigare version av denna text.)

Litteratur

NAGEL, T, *The View from Nowhere*, Oxford, 1986.
PETTIT, P, (red), *Consequentialism*, Aldershot, 1993.
SCHEFFLER, S, *The Rejection of Consequentialism*, Oxford, 1987.

ERIK CARLSON

Några teorier om egenvärde

En fullständig teori om egenvärde måste naturligtvis först och främst besvara frågan *vad* som har egenvärde. Kandidaterna är som bekant många: lycka, preferenstillfredsställelse, jämlikhet, mänsklig fulländning, kunskap, etc. Därnäst kommer frågan om egenvärde är mätbart och i så fall på vilken typ av skala. Även om dessa fundamentala problem vore lösta, så skulle emellertid andra svårigheter återstå. En sådan svårighet är att utifrån de sakförhållanden vår axiologi betraktar som fundamentala värdebärare ('värdebaser') tilldela egenvärde till andra sakförhållanden. Jag skall här först kritisera en lösning på detta problem som föreslagits av Edward Oldfield, för att sedan presentera och diskutera tre egna alternativa lösningsförslag.

1. Värdebaser

Låt oss anta att det är sakförhållanden som har egenvärde och att det är meningsfullt att tala om konjunktiva, disjunktiva och negativa sakförhållanden. Vidare antar vi att vissa sakförhållanden är 'värdebaser', i den meningen att de har egenvärde på ett mer grundläggande sätt än andra förhållanden. Egenvärdet hos ett förhållande som inte är en värdebas bestäms helt och hållet av värdebasernas egenvärden. Det är möjligt att ha kunskap om en värdebas egenvärde utan att känna till egenvärdet hos något annat förhållande.

Enligt Warren Quinn är värdebaserna sannolikt sakförhållanden som "lokaliserar en specifik kännande individ vid en specifik punkt längs en värderelevant dimension, såsom lycka, dygd, visdom, etc" (Quinn 1974, s 131). Jag skall utgå från denna idé, men göra ett par tillägg. Noga taget tycks Quinns värdebasbegrepp både för vitt och för snävt. Det är för vitt eftersom det inte innehåller några tidsangivelser. En värdebas borde lokalisera en individ vid en punkt längs en värderele-

vant dimension *under ett specifikt tidsintervall*. I avsaknad av en sådan tidsparameter tycks ett sakförhållande som 'Jon är lycklig till graden 2' inte ha något bestämt egenvärde, och därmed inte vara en värdebas, eftersom det inte anger hur länge Jon är lycklig. Det är rimligtvis intrinsikalt bättre att han är lycklig i ett år, än att han är lycklig i en timme. Jag skall därför anta att en tidsreferens finns inbakad i värdebaserna, trots att jag för enkelhetens skull inte kommer att skriva ut den.

Quinns begrepp är för snävt i så måtto att det förutsätter att entiteterna som lokaliseras längs en värderelevant dimension måste vara 'kännande individer'. Om t ex skönhet eller jämlikhet är en värderelevant dimension kunde konstverk eller samhällen höra till de objekt värdebaserna lokaliserar. Mina exempel kommer visserligen uteslutande att röra människors lycka och olycka, men jag vill alltså inte hävda att alla värdebaser involverar människor eller andra kännande varelser.

Låt oss för resonemangets skull anta att lycka är mätbart (på kvotskala) och värdemässigt relevant. En värdebas som innehåller en större kvantitet lycka, eller en mindre kvantitet olycka, har sålunda större egenvärde än en värdebas som innehåller en mindre kvantitet lycka, eller en större kvantitet olycka. (Ingen värdebas innehåller både lycka och olycka.) Att Jon är 'lycklig$_2$', dvs lycklig till graden 2 under ett visst tidsintervall, och att Bea är 'lycklig$_{-4}$', dvs olycklig till graden 4 under ett visst tidsintervall, är exempel på värdebaser enligt denna teori. Vi kan beteckna dessa sakförhållanden 'j_2' respektive 'b_{-4}'. Alla de exempel som följer gäller, tänker vi oss, ett och samma tidsintervall. Två olika värdebaser som involverar samma person, som t ex b_1 och b_5, är därför inbördes oförenliga. (Bea kan inte vara lycklig till två olika grader under hela det relevanta intervallet.) Låt oss också anta att en värdebas egenvärde motsvarar den kvantitet lycka eller olycka den innehåller, och låt 'EV' beteckna egenvärde. Så gäller t ex att EV(j_2)=2 och att EV(b_{-4})= –4.

(Några kanske vill hävda att mindre specifika sakförhållanden som inte involverar någon bestämd individ, t ex 'Någon är lycklig$_2$', har egenvärde på ett mer 'fundamentalt' sätt än Quinns och mina värdebaser. Huvudskälet mot att välja sådana mindre specifika förhållanden som värdebaser är att detta skulle göra det avsevärt svårare att hitta en funktion som bestämmer andra sakförhållandens egenvärden utifrån värdebasernas värden.)

Jag kommer inte att skilja mellan att ha egenvärdet noll och att sakna egenvärde. Vissa skulle kanske vilja göra denna distinktion och hävda att sakförhållanden som j_0 (dvs förhållandet att Jon varken är lycklig eller olycklig under det berörda tidsintervallet) och $j_2 \& b_{-2}$ har egenvärdet noll, medan förhållanden som 'Gräs är grönt' och 'Det finns stenar' saknar egenvärde. Åtminstone i detta sammanhang förefaller dock denna eventuella skillnad oväsentlig.

2. *Oldfields teori*

Vårt problem är nu, givet denna enkla modell, att värdera sakförhållanden som inte är värdebaser. Vilket egenvärde har förhållanden som, t ex, $\neg j_2$, 'Någon är lycklig$_2$', $j_2 \& b_{-4}$, eller $j_2 \vee b_5$? Quinn och Oldfield har föreslagit sinsemellan snarlika teorier om egenvärdet hos sådana icke grundläggande sakförhållanden. En nackdel med deras förslag är att de gör egenvärdet hos vissa förhållanden *världsrelativt*, dvs olika i olika möjliga världar. Denna idé om världsrelativt egenvärde tycks dubiös. Det hävdas ofta att ett sakförhållandes egenvärde bara beror på dess inre egenskaper, eller att dess egenvärde är det värde det har med nödvändighet, dvs i varje möjlig värld. Något sådant villkor verkar behövas för fånga den intuitiva skillnaden mellan egenvärde och instrumentellt värde. En teori som laborerar med världsrelativa egenvärden hotar att utplåna denna distinktion.

Förutom denna generella svårighet finns det också mer specifika problem med Quinns och Oldfields förslag. Eftersom Oldfields teori i mitt tycke är en förbättring av Quinns, så skall jag endast diskutera den förstnämnda. De flesta av de kritiska anmärkningar jag kommer att göra drabbar emellertid även Quinns teori.

Innan han formulerar sin huvudprincip för bestämmandet av egenvärde, ger Oldfield följande två definitioner. Ett sakförhållande p är *irrelevant* om och endast om (omm) p är logiskt oberoende av varje värdebas, av varje negation av en värdebas, samt av varje kontingent konjunktivt sakförhållande, vars samtliga konjunkter är antingen en värdebas, eller negationen av en värdebas (Oldfield 1977, s 237. Jag har förenklat hans definition något. Begreppet irrelevant sakförhållande behövs för att disjunktioner som (j_1 &'Gräs är grönt') $\vee$ (b_2 &'Det finns stenar') skall tilldelas egenvärde.) En mängd M är en *minimal mängd för p i världen w* omm följande villkor gäller:

(i) M är en mängd av värdebaser som alla föreligger i w,
(ii) det finns ett irrelevant sakförhållande c som föreligger i w och som är sådant att konjunktionen av c med elementen i M logiskt implicerar p,
(iii) det finns ingen äkta delmängd av M och något irrelevant sakförhållande som föreligger i w, vars konjunktion logiskt implicerar p (Oldfield 1977, s 239).

Vi kan nu återge Oldfields princip för tillskrivandet av egenvärde, låt oss kalla den 'EVO':

EVO För varje värld w och sakförhållande p, gäller att
(a) om p föreligger i w så är dess egenvärde i w lika med summan av egenvärdena hos elementen i unionen av alla minimala mängder för p i w,
(b) om p inte föreligger i w, så (i) om dess egenvärde är större än eller lika med (mindre än) noll i alla världar där det föreligger, så är dess egenvärde i w lika med det minsta (största) egenvärde det har i någon värld där det föreligger, (ii) om dess egenvärde är större än eller lika med noll i vissa världar där det föreligger och mindre än noll i vissa världar där det föreligger, så är dess egenvärde i w noll (Oldfield 1977, ss 240, 242).

EVO har en del implausibla konsekvenser. Antag att j_1 och b_1 föreligger i w. Om vi låter 'EV_w' stå för 'egenvärde i världen w', så följer att $EV_w(j_1 \vee b_1)=2$, medan $EV_w(j_1 \vee b_{100})=1$. Men $j_1 \vee b_{100}$ kan knappast ha *lägre* egenvärde än $j_1 \vee b_1$, eftersom de två förhållandena har en gemensam disjunkt och den andra disjunkten är *bättre* i det förra förhållandet än i det senare. Likaledes, om j_0 och b_{-1} föreligger i w, så gäller att $EV_w(j_0 \vee b_{-100}) =0$, medan $EV_w(j_{100} \vee b_{-1})=-1$. Det senare förhållandet skiljer sig från det förra därigenom att varje disjunkt är ersatt av en mycket *bättre* värdebas. Därför är det knappast rimligt att hävda att det har *lägre* egenvärde.

Också beträffande universella och existentiella generaliseringar, såväl som negationer, ger EVO egendomliga resultat. Antag att det endast finns tio personer i w, och att de alla är lyckliga_2. Då följer att EV_w('Alla är lyckliga_2')=20, medan EV_w('Alla är lyckliga_{19}')=19. (Eftersom 'Någon är lycklig_{19}' inte föreligger i w är detta förhållandes

egenvärde i w, enligt villkor (b) i EVO, lika med det lägsta egenvärde det har i någon värld där det föreligger. Detta värde är 19, dvs dess värde i en värld där det endast existerar en person, och denna person är lycklig$_{19}$.) En ännu underligare konsekvens är att EV$_w$('Ingen är lycklig$_{19}$')=20, medan EV$_w$('Någon är lycklig$_{19}$')=19. (Den enda minimala mängden för 'Ingen är lycklig$_{19}$' i w innehåller alla värdebaser som föreligger i w. Därför gäller att EV$_w$('Ingen är lycklig$_{19}$')=20.) Vad gäller negationer implicerar EVO att om j_5 föreligger i w, så gäller att EV$_w$($\neg j_n$)=5, för alla tal n≠5. Det verkar dock mycket tveksamt om negativa förhållanden som $\neg j_2$ har något bestämt egenvärde (skilt från noll).

Rörande disjunktionerna $j_1 \vee b_{-20}$ och $j_{-1} \vee b_{20}$ hävdar Oldfield att den senare är 'i en mening' intrinsikalt bättre än den förra. "Denna mening är att det trots att det finns världar w_i, w_j sådana att egenvärdet av $[j_1 \vee b_{-20}]$ i w_i är högre än egenvärdet av $[j_{-1} \vee b_{20}]$ i w_j, så finns det också världar där $j_{-1} \vee b_{20}$ är bättre än $j_1 \vee b_{-20}$ är i någon värld" (Oldfield 1977, s 245). Låt 'EV*' beteckna egenvärde i denna mening. Jag har kritiserat Oldfields princip EVO för att den implicerar exempelvis att EV$_w$ ($j_1 \vee b_1$) > EV$_w$($j_1 \vee b_{100}$), om j_1 och b_1 föreligger i w. Oldfield kunde svara genom att påpeka att EV*($j_1 \vee b_{100}$) > EV*($j_1 \vee b_1$), och att hans teori därför kan fånga intuitionen att $j_1 \vee b_{100}$ är intrinsikalt bättre, eller i varje fall inte sämre, än $j_1 \vee b_1$.

Detta svar är dock av tvivelaktigt värde. Påståendet att det finns en betydelse av 'intrinsikalt bättre' i vilken $j_1 \vee b_{100}$ i w är intrinsikalt bättre än $j_1 \vee b_1$, och en annan betydelse i vilken det omvända gäller, tycks vara ett *ad hoc*-antagande, ägnat att få teorin att stämma bättre överens med våra intuitiva omdömen. Dessutom uttrycker EV* knappast en rimlig idé om egenvärde. Det gäller t ex att EV*($j_1 \vee j_{-100}$) > EV*(j_0) och att EV*('Ingen är lycklig$_{20}$') > EV*('En miljon människor är lyckliga$_{20}$'). (Den sistnämnda värderelationen råder eftersom det finns världar där en minimal mängd för 'Ingen är lycklig$_{20}$' innehåller exempelvis en miljon värdebaser, vilka alla innebär att någon är lycklig$_{21}$.)

Oldfield erkänner faktiskt också ett tredje egenvärdebegrepp, nämligen 'absolut' egenvärde (Oldfield 1977, s 240). Han gör mycket lite för att förklara detta begrepp, men han förefaller benägen att säga att p:s absoluta egenvärde är identiskt med p:s (världsrelativa) egenvärde i en värld där p inte föreligger, i enlighet med villkor (b) i EVO. Hur

som helst anser han att idén om världsrelativt egenvärde har en 'intuitiv grund' och därför är av intresse oberoende av dess relation till absolut egenvärde. Som ett argument för detta påstående uppmanar Oldfield oss att överväga vad 'vårt första, oreflekterade svar' skulle vara, om vi tillfrågades om egenvärdet hos en disjunktion som $j_1 \vee b_{-20}$. "[E]tt sådant svar skulle vara att säga 'det beror helt och hållet på huruvida j_1 eller b_{-20} föreligger'. Med andra ord, vi är böjda att säga att värdet hos $[j_1 \vee b_{-20}]$ varierar från värld till värld" (Oldfield 1977, s 247).

Enligt min mening avspeglar detta svar snarare en oförmåga att förstå begreppet egenvärde. Ett sakförhållandes egenvärde måste vara oberoende av vilka andra förhållanden som råkar föreligga. Om detta antagande överges tycks distinktionen mellan egenvärde och instrumentellt värde kollapsa. Det skulle därför vara en stor vinst om vi kunde hitta en teori som undviker världsrelativa egenvärden, men ändå ger plausibla svar vad gäller egenvärdet hos förhållanden som inte är värdebaser. Jag skall formulera tre sådana förslag, varav det andra mycket liknar (min tolkning av) Oldfields idé om absolut egenvärde.

3. Tre förslag

Mitt första förslag är mycket enkelt. Idén är att ett sakförhållandes egenvärde är en funktion av egenvärdena hos de värdebaser det logiskt implicerar. Så länge vi håller oss till lycka och olycka är det kanske också naturligt att tänka sig att den relevanta funktionen helt enkelt är addition. (Detta sista antagande är dock problematiskt, då det tycks leda till 'den Frånstötande Slutsatsen'. Se Parfit 1984, s 388.) Vi får då:

EV1 Egenvärdet hos p är lika med summan av egenvärdena hos de värdebaser som p logiskt implicerar.

Enligt EV1 saknar disjunktioner som $j_2 \vee j_{-5}$ och $j_2 \vee j_5$ egenvärde. Beträffande den förra disjunktionen tror jag det finns ett ganska bra argument för denna slutsats. Betrakta två världar, w och w', där j_2 respektive j_{-5} är de enda föreliggande värdebaserna. Intuitivt sett finns det inget intrinsikalt *dåligt* sakförhållande i w, och inget intrinsikalt *bra* förhållande i w'. Men om vi antar att $j_2 \vee j_{-5}$ har antingen negativt eller positivt egenvärde, och att egenvärdet hos ett förhållande inte varierar över möjliga världar, så finns det antingen ett intrinsikalt

dåligt förhållande i *w*, eller ett intrinsikalt bra förhållande i *w'*. Alltså, kunde man hävda, saknar 'blandade' disjunktioner som $j_2 \vee j_{-5}$, med en bra och en dålig disjunkt, egenvärde.

Rörande 'rent goda' och 'rent dåliga' disjunktioner, å andra sidan, ger EV1 mer tvivelaktiga resultat. En disjunktion som $j_2 \vee j_5$ implicerar logiskt att åtminstone ett gott sakförhållande föreligger, men inte att något dåligt förhållande föreligger. Därför tycks det rimligt att hävda att $j_2 \vee j_5$ är intrinsikalt gott. Av analoga skäl är då $j_{-2} \vee j_{-5}$ intrinsikalt dåligt. Samma resonemang kan tillämpas på förhållanden som 'Någon är lycklig_2' och 'Tio personer är lyckliga_{-5}', vilka saknar egenvärde enligt EV1. Jag skall nu föreslå en lösning som tar hänsyn till dessa intuitioner.

Låt oss behålla Oldfields ovan återgivna definition av ett irrelevant sakförhållande, och låt oss generalisera hans definition av en minimal mängd, genom att ta bort hänvisningen till en möjlig värld. Alltså, en mängd *M* är en *minimal p-mängd* omm (i) varje element i *M* är en värdebas eller ett irrelevant sakförhållande, (ii) konjunktionen av elementen i *M* är kontingent och logiskt implicerar *p*, och (iii) ingen äkta delmängd av *M* är sådan att konjunktionen av dess element är kontingent och logiskt implicerar *p*. Vidare, en mängd *M* är en *p-mängd* omm (i) *M* är en minimal *p*-mängd eller en union av minimala *p*-mängder, och (ii) konjunktionen av elementen i *M* är kontingent. Låt oss slutligen definiera begreppen 'minimum av *p*' ('Min(*p*)') och 'maximum av *p*' ('Max(*p*)'), på följande sätt: *Min*(*p*) (*Max*(*p*)) är summan av egenvärdena hos elementen i en *p*-mängd *M*, sådan att summan av egenvärdena hos elementen i *M* är åtminstone lika liten (stor) som summan av egenvärdena hos elementen i varje annan *p*-mängd. Vi kan nu formulera vårt andra förslag:

EV2 (a) Ett sakförhållande *p* är intrinsikalt gott omm $\text{Min}(p)>0$, intrinsikalt dåligt omm $\text{Max}(p)<0$ och intrinsikalt neutralt omm det varken är intrinsikalt gott eller intrinsikalt dåligt,

(b) om *p* är intrinsikalt gott så $\text{EV}(p)=\text{Min}(p)$, om *p* är intrinsikalt dåligt så $\text{EV}(p)=\text{Max}(p)$ och om *p* är intrinsikalt neutralt så $\text{EV}(p)=0$.

Enligt EV2 är endast förhållanden som säkerställer något på det hela taget gott intrinsikalt goda, och endast förhållanden som säkerställer

något på det hela taget dåligt är intrinsikalt dåliga. Dessutom motsvarar egenvärdet hos ett förhållande så att säga den kvantitet gott eller dåligt förhållandet säkerställer. Således, EV(j_5&b_{-2})=3, EV($j_1 \vee j_5$)=1, EV($j_{-1} \vee j_{-5}$)=–1, EV('Någon är lycklig$_2$')=2, och EV($j_{-1} \vee j_5$)=0. Dessa resultat verkar rimliga.

Å andra sidan skulle vi kanske vilja hävda t ex att EV($j_0 \vee j_{100}$) > EV($j_0 \vee j_{-100}$), eller att EV($j_{100} \vee j_{-1}$) > EV($j_1 \vee j_{-100}$). Men EV2 implicerar att alla dessa förhållanden har egenvärdet noll. Nu kan det tyckas omöjligt att behålla både idén att endast sådana förhållanden som säkerställer något gott eller dåligt är intrinsikalt goda eller dåliga och åsikten att EV($j_0 \vee j_{100}$) > EV($j_0 \vee j_{-100}$). Bägge dessa teser kan dock upprätthållas, om vi är beredda att acceptera att det finns sakförhållanden som är intrinsikalt 'obestämda', dvs varken goda, dåliga eller neutrala, och att vissa obestämda förhållanden ändå är jämförbara med avseende på egenvärde. Vi kan ge utrymme för obestämda förhållanden i vår teori genom att låta egenvärden representeras av numeriska intervall, i stället för av enskilda tal. Mitt sista förslag blir då detta:

EV3 (a) För varje sakförhållande p, EV(p) = intervallet Min(p)–Max(p) (om Min(p)=Max(p) antar jag att intervallet Min(p)–Max(p) = Min(p) och Max(p)),

(b) p är intrinsikalt gott omm Min(p)>0, intrinsikalt dåligt omm Max(p)<0, intrinsikalt neutralt omm Min(p)=0 och Max(p)=0, och intrinsikalt obestämt omm det varken är intrinsikalt gott, intrinsikalt dåligt, eller intrinsikalt neutralt,

(c) om Min(p) ≥ Min(q) och Max(p) ≥ Max(q), så EV(p) ≥ EV(q),

(d) om EV(p) ≥ EV(q) och EV(q) ≥ EV(p), så EV(p) = EV(q),

(e) om EV(p) ≥ EV(q) och inte EV(q) ≥ EV(p), så EV(p) > EV(q),

(f) om varken EV(p) ≥ EV(q) eller EV(q) ≥ EV(p), så är EV(p) och EV(q) ojämförbara.

EV3 implicerar att EV($j_0 \vee j_{100}$) > EV($j_0 \vee j_{-100}$) (och att EV($j_{100} \vee j_{-1}$) > EV($j_1 \vee j_{-100}$)) och motsvarar därför vissa intuitioner bättre än EV2.

Eftersom dessa fyra sakförhållanden alla är intrinsikalt obestämda så innebär dessa resultat inte heller något brott mot idén att endast förhållanden som säkerställer något på det hela taget gott eller dåligt är intrinsikalt goda eller dåliga.

Också vad gäller värderingen av förhållanden som $j_2 \vee b_2$ och 'Någon är lycklig$_2$' skiljer sig EV3 från EV2. Enligt EV2 har dessa bägge förhållanden egenvärdet 2. Enligt EV3, däremot, är EV($j_2 \vee b_2$) lika med intervallet 2 – 4, och EV('Någon är lycklig$_2$')= intervallet 2 – n*2, där n är antalet personer i den mest 'tätbefolkade' möjliga värld där alla är lyckliga till graden 2 under det berörda tidsintervallet. Även i dessa fall tycks EV3 ge rimligare resultat än EV2. Det finns ju t ex en 'realisering' av $j_2 \vee b_2$, nämligen $j_2 \& b_2$, som är bättre än både j_2 och b_2, tagna var för sig. Därför kan man, vilket Krister Bykvist uppmärksammat mig om, med fog hävda att $j_2 \vee b_2$ är intrinsikalt bättre än j_2 och b_2. (Detta är anledningen till att jag definierat Min(p) och Max(p) i termer av 'sämsta' respektive 'bästa' p-mängder, och inte direkt i termer av 'sämsta' och 'bästa' *minimala* p-mängder. Enligt de senare definitionerna skulle EV3 implicera att EV($j_2 \vee b_2$)=2 och EV('Någon är lycklig$_2$')=2.)

En ytterligare skillnad mellan EV2 och EV3 gäller negationer, såsom $\neg j_1$. Medan EV2 implicerar att EV($\neg j_1$)=0, så implicerar EV3 att EV($\neg j_1$)= –n – n, om vi antar att det finns ett tal n, sådant att –n och n representerar de största möjliga graderna av olycka respektive lycka. Således är $\neg j_1$ ett i extrem grad obestämt förhållande. EV3 är därför förenlig med mitt påstående ovan, att negativa sakförhållanden saknar ett *bestämt* egenvärde (skilt från noll). Det kan ändå förefalla lite egendomligt att egenvärdet hos $\neg j_1$ är ojämförbart med värdet hos de flesta andra (icke-negativa) förhållanden. Vi kanske är böjda att påstå t ex att EV(b_{10}) > EV($\neg j_1$). Men om vi betänker det faktum att $\neg j_1$ har både mycket goda 'realiseringar', såsom j_{100}, och mycket dåliga 'realiseringar', såsom j_{-100}, så förefaller slutsatsen att EV(b_{10}) och EV($\neg j_1$) är ojämförbara acceptabel.

EV3 har en viss likhet med ett förslag som framförts av Sven Danielsson. Hans idé innebär ungefärligen att p är intrinsikalt bättre än q omm Min(p)>Max(q). (Danielsson 1978, s 11.) I mitt tycke är Danielssons förslag i överkant 'försiktigt'. Som jag antytt ovan är jag benägen att hävda att $j_0 \vee j_{100}$ är bättre än $j_0 \vee j_{-100}$. Dessa förhållanden är ojämförbara i Danielssons teori.

Jag skall avsluta denna uppsats med att peka på två egenskaper som delas av EV1, EV2 och EV3. För det första bryter dessa teorier mot en princip, förespråkad av bl a Roderick Chisholm, som säger att en disjunktion aldrig är intrinsikalt bättre eller intrinsikalt sämre än bägge disjunkterna. EV1 bryter mot denna princip, som vi kan kalla '(D)', på ett trivialt sätt, eftersom EV1 implicerar att alla disjunktioner som inte logiskt implicerar någon värdebas har egenvärdet noll. För att inse att EV2 och EV3 bryter mot (D), betrakta disjunktionen $p=(j_1 \vee (b_1 \& d_{-100}))\ (b_1 \vee (j_1 \& d_{-100}))$. EV($p$)=1, enligt EV2 och EV3. (Eftersom p är logiskt ekvivalent med $j_1 \vee b_1$, så är j_1 och b_1 de enda minimala p-mängderna.) Men för varje disjunkt q i p så gäller att EV(q)=0 enligt EV2, medan EV(q) = –99 – 1, enligt EV3. Alltså är p ett exempel på en disjunktion som är intrinsikalt bättre än bägge disjunkterna.

Denna slutsats behöver dock inte genera oss. Oldfield har givit ett övertygande argument mot (D). Betrakta alla konjunktioner av formen $j_1 \& b_i$, där förhållandet att Bea inte existerar ingår bland b_i-förhållandena. Några av dessa konjunktioner är intrinsikalt goda, några är dåliga, och några är neutrala. Låt G, D och N vara disjunktionen av de goda, dåliga respektive neutrala konjunktionerna. Det förefaller som om varken N eller disjunktionen $G \vee D$ är intrinsikalt god. Om (D) gäller, och en disjunktion således inte kan vara bättre än varje disjunkt, så följer att inte heller disjunktionen $N \vee (G \vee D)$ är intrinsikalt god. Men $N \vee (G \vee D)$ är logiskt ekvivalent med det intrinsikalt goda förhållandet j_1, och måste därför vara intrinsikalt god. (Oldfield 1977, s 245.)

Den andra egenskapen värd att notera är att vissa 'holistiska' värdefenomen föreligger, enligt både EV1, EV2 och EV3. Ingen av dessa teorier uppfyller denna princip:

(*) Om p, q och r är logiskt oberoende, och EV(p)=EV(q), så EV($p \& r$)= EV($q \& r$).

För att hitta ett motexempel mot (*), enligt alla tre teorierna, låt $p=j_1 \vee b_2$, låt $q=b_1 \vee j_2$ och låt $r=j_0$. Enligt EV1, gäller att EV(p)=0 och EV(q)=0, enligt EV2 gäller att EV(p)=1 och EV(q)=1, och enligt EV3 gäller att EV(p) och EV(q) båda är lika med intervallet 1 – 3. Alla tre teorierna implicerar emellertid att EV($p \& r$)=2, medan EV($q \& r$)=1.

Om principen (*) däremot inskränks till att gälla endast värdebaser, så är den giltig enligt både EV1, EV2 och EV3. Detta kan uppfattas

som en nackdel med dessa principer, eftersom det förmodligen finns få rimliga axiologier som medger ett sådant enkelt sätt att aggregera egenvärdena hos värdebaser. Det finns dock ingen anledning att tro att någon specifik aggregationsprincip skulle vara förenlig med alla, eller ens de flesta, plausibla axiologier. Detta aggregationsproblem tycks därför inte kunna diskuteras på något fruktbart sätt utan mycket mer detaljerade axiologiska antaganden än den enkla hedonism vi här förutsatt.

Not: Jag vill tacka Erik Olsson, deltagarna vid Filosofidagarna i Umeå och framförallt Krister Bykvist för värdefulla synpunkter. En utförligare engelsk version av denna uppsats kommer att publiceras i *Philosophical Studies*.

Litteratur

SVEN DANIELSSON, 'Another Version of the Harman-Quinn-Oldfield Theory of Intrinsic Value', stencil, Uppsala Universitet 1978.

EDWARD OLDFIELD, 'An Approach to a Theory of Intrinsic Value', *Philosophical Studies* 32 1977, 233–49.

DEREK PARFIT, *Reasons and Persons,* Oxford: Oxford University Press, 1984.

WARREN S. QUINN, 'Theories of Intrinsic Value', *American Philosophical Quarterly* 11 1974, 123–32.

INGMAR PERSSON

Lika för olika

1. Problemet jämlikhet kontra faktisk olikhet

Vi lever i en värld av olikhet och individuell variation. Den innehåller en mångfald av arter, och inom varje art flödar den individuella särarten. Vi människor skiljer oss åt ifråga om intelligens, intressen, fysisk styrka, hälsa och oräkneliga andra egenskaper. Denna långtgående olikhet till trots bär ändå många av oss på en moralisk övertygelse att vi i någon mening är jämlika eller lika mycket värda. Men hur kan denna jämlikhet förenas med den faktiska olikheten? Kan den senare annat än leda till att vi att vårt värde blir olika p g a att *några* skillnader är relevanta för värderingen av oss?

Min avsikt är att presentera en tolkning av idén att alla människor är jämlika som är förenlig med ett erkännande av denna faktiska olikhet. En del ser kanske inget problem här, eftersom de menar att jämlikhet är en *värdering* och som sådan oberoende av fakta. Vi kan därför helt enkelt *bestämma oss* för att värdera människor som jämlika, oavsett deras faktiska olikhet, bestämma oss för att tillämpa t ex Bentham's välbekanta diktum "Everybody to count for one, nobody for more than one', tolkad som ett direktiv att negligera alla faktorer hos individer utom deras *antal.*

Det är riktigt nog att värderingar inte *är härledbara* ur några underliggande sakförhållanden, men dessa kan ändå bilda *nödvändiga förutsättningar* för dem. Till exempel kan vi inte bestämma oss för behandla träd, grepar och stenar som människans jämlikar, dvs lika bra eller dåligt som människor. Skälet är att de inte uppfyller en nödvändig förutsättning för att någon behandling skall kunna vara bra eller dålig för dem: de saknar medvetande och kan därmed inte reagera med positiva eller negativa upplevelser på den. Så värderingar är inte helt fristående från faktiska förhållanden: även om de senare inte är

(logiskt) tillräckliga för värderingar, kan de som synes vara (logiskt) nödvändiga. Farhågan är då att en likhet mer omfattande än den faktiskt förefintliga bland människor är nödvändig för att de skall ha lika värde.

I varje fall är en *fullständig* underliggande likhet *tillräcklig* för likvärdighet:

(1) A och B måste värderas lika om det inte finns någon underliggande skillnad mellan dem.

En skillnad är "underliggande" visavi en värdering om den logiskt kan anföras som skäl för den. Att A och B skiljer sig värdemässigt i ett hänseende kan inte anföras som *skäl* för att de skiljer sig *i detta hänseende*. (1) utesluter däremot inte att en värdeskillnad kan ligga under en annan och gör – i motsats till många teser om värdens "superveniens" – inte anspråk på att kunna särskilja värdeegenskaper från en klass av underliggande, "naturliga" egenskaper.

Principens tillämpning blir emellertid mycket begränsad, ty två distinkta ting är självfallet aldrig lika med avseende på *alla* underliggande egenskaper (åtminstone inte tid-rumsliga). Trots detta kan (1) ha ett visst teoretiskt intresse som exempel på hur underliggande fakta kan närma sig tillräcklighet för värdering. Vi erhåller en mer substantiell princip genom att skjuta in att skillnaden måste vara "relevant":

(2) A och B måste värderas lika om det mellan dem inte finns någon underliggande skillnad som är värderingsmässigt *relevant*.

(2) kan – som Sven Danielsson har påpekat – förstärkas så att den artikulerar också en nödvändig betingelse, men med eller utan denna förstärkning är problemet att utan en angivelse av vilka skillnader som är relevanta blir principen normativt innehållslös.

Det finns en specifikation av (2) som är speciellt intressant i sammanhanget. Den specificerar "värderas lika" som "behandlas lika väl" och hävdar således:

(3) A och B måste behandlas lika väl (eller illa) om det mellan dem inte finns någon underliggande skillnad som är relevant för behandlingen av dem.

Strängt taget kan (3) inte förstärkas till nödvändighet, eftersom det är tänkbart att likvärdig behandling kan rättfärdigas fastän det existerar relevanta skillnader i olika avseenden mellan individer, nämligen om dessa alltid – som genom ett under! – tar ut varandra. Denna möjlighet är dock alltför långsökt för att jämlikhetsförespråkarna skall kunna förlita sig på den. De gör klokast i att betrakta frånvaron av relevanta skillnader också som nödvändig för likvärdig behandling.

Den specifikation som (3) utgör är endast möjlig om A och B är medvetandeutrustade entiteter för vilka något kan vara bra eller dåligt. Därmed blir det möjligt att tala om *rättvisa* och karakterisera (3) som en *rättviseprincip* som stadfäster att *rättvisan* kräver att A och B måste behandlas likvärdigt om det inte finns någon relevant skillnad dem emellan.

(3) är en princip som vi känner igen från Aristoteles' diskussion av "distributiv rättvisa" i femte boken av hans *Nikomachiska etik*. I modern tid har Isaiah Berlin framställt denna princip att relevant lika fall skall behandlas lika som sin tolkning av mänsklig jämlikhet: vi skall behandlas lika, "såvida det inte finns ett tillräckligt skäl att inte göra så" ("Equality", *Concepts and Categories*, Oxford U P, 1980, s 82). Som redan har antytts är nackdelen med dylika formuleringar att de, vilket Berlin också inser, inte förser jämlikhetstanken med normerande innehåll så länge det förblir obevisat att de omtalade mångfaldiga skillnaderna människor emellan är irrelevanta för en rättvis behandling av dem.

2. *De obestridliga olikheterna*

Skall jämlikhetstillskyndarna med stöd i (3) kunna försvara sin tes, måste de följaktligen argumentera för denna irrelevans. Man kan skilja mellan en starkare och en svagare variant av en sådan homogenitet. Den starkare kan lyda så här:

(H+) Alla människor besitter i lika grad alla de egenskaper som är relevanta för en rättvis behandling av dem.

Den svagare kan formuleras:

(H-) Alla människor besitter i lika grad alla de egenskaper som är relevanta för en rättvis behandling av dem *i vissa hänseenden*.

Poängen med (H-) är att människor kan vara lika i de avseenden som utgör fundamentet för vissa grundläggande rättigheter, som rätt till liv, politiska fri- och rättigheter som yttrandefrihet och rösträtt, likhet inför lagen och dylikt.

(Om jämlikheten dessutom skall vara *exklusivt* mänsklig måste dess anhängare också demonstrera att de egenskaper som (H+) och (H-) kvantifierar över särskiljer människor från andra väsen, men denna aspekt lämnas här därhän.)

Bakom (3) och homogenitetsteserna ligger tanken att jämlik behandling består i att alla får *samma tilldelning* eller *samma ranson* (t ex samma rättigheter). Om man tänker sig att jämlik behandling består i att man portionerar ut samma ranson till alla, är det naturligt (men, som vi sett, inte strikt nödvändigt) att basera denna behandling på en frånvaro av relevanta skillnader.

En annan strategi står emellertid till buds: att jämlik behandling anses ha ett visst *mål* – att alla får det "lika bra" – och att den (som medel) tillåter tilldelning av *olika ransoner* för att uppnå detta mål (t ex att man skänker mer åt dem som har större behov). Vi kan tala om *medel-* respektive *målinriktad* jämlikhetsdistribution. Den senare modellen förutsätter inte en förhandenvarande homogenitet som rättfärdigar likheten i utdelade medel, men den är i stället förpliktigad att mot bakgrund av faktiska olikheter rättfärdiga det likställdhetsmål den ställer upp. Innan jag tar upp denna modell vill jag pröva den förra, genom att undersöka sanningshalten av (H-).

En föreslagen kandidat till den sökta likheten, på vilken de grundläggande rättigheterna kan förmodas vara baserade, är egenskapen *att vara människa* tolkad som en biologisk artbestämning: att vara medlem av arten *homo sapiens*. Men enligt det gängse – genetiska – kriteriet på arttillhörighet tillmötesgår denna egenskap inte kravet på (värderings)relevans. Ett tankeexperiment kan klargöra detta. Antag att det på en planet i ett annat solsystem finns en art vilken har en lika högt utvecklad intelligens som vår, vilken lever i en civilisation där konst och vetenskap är lika avancerade som i vår, och vilken kanske också till utseendet liknar oss, men vilken är distinkt från den mänskliga arten på grund av att dess genetiska olikhet är alltför stor för att interfertilitet med oss skall råda. Om vi lyckas etablera kontakt med dessa varelser, vore det då rätt och rättvist att handskas mera lättvindigt med deras väl och ve än med vårt eget, kanske jaga och för-

slava dem? Nej, eftersom deras själutrustning är lika välutvecklad som vår. Alltså verkar det vara denna snarare än arttillhörighet som är moraliskt relevant.

En propå i denna anda är John Rawls' tanke att människor är jämlika "som moraliska personer, som varelser med en föreställning om vad som är bra för dem och med ett sinne för rättvisa" (*A Theory of Justice*, Harvard U P 1971, s 19). Idén kan spåras tillbaka till Kants lära att alla människor är delaktiga i "ändamålens kungarike", dvs har förmåga att skaffa sig en moraliskt god vilja.

Detta antagande är emellertid oriktigt: vissa människor är så gravt mentalt handikappade att de tycks renons på all förmåga att handla moraliskt. Dessutom verkar vissa människor ha en större kapacitet att efterleva moralens krav än andra: det förefaller inte troligt att helgonet och brottslingen skulle ha samma förmåga att leva upp till moralens normer. Varför skulle denna gradskillnad inte avspegla sig i form av ojämlikhet på rättighetsplanet (så att man räddar livet på helgonet snarare än på brottslingen, ge fler röster åt de moraliskt mer fullkomliga etc)? Den allmänna slutsatsen blir således att ifråga om själsförmögenheter är människor så radikalt olika att det inte här kan finnas någon grundval för jämlikhet.

Härav följer att inte bara (H-) utan också (H+) är falsk, att människor är olika med avseende på *vissa* egenskaper som är relevanta vid fördelningen av för- och nackdelar. Ekonomiskt välstånd fördelas exempelvis efter egenskaper som människor besitter i olika omfattning: somliga är onekligen bättre skickade än andra att t ex nå toppjobben med de högsta lönerna.

Ibland hör man kravet att folk bör ges samma sociala start, så att de har lika möjlighet ("equal opportunity") att nå de åtråvärda samhällsposterna. Ingen stat har uppnått jämlikhet i detta hänseende – har t ex eliminerat de ekonomiska skillnader som ger olika möjlighet till skolutbildning och sociala kontakter – men även om så skulle vara fallet, skulle folk inte ha samma chanser att klättra upp till samhällstoppen, ty vissa föds med bättre genetiska anlag än andra, med anlag för att utveckla högre intelligens, mer energi, etc. En del föds rentav med svåra handikapp som inte bara stäcker deras karriärmöjligheter utan också gör dem starkt beroende av hjälp från andra.

Låt oss distingera mellan det värde en människas liv (vid olika tidpunkter) har för andra varelser (för vilka det kan finnas värde) – dess

transitiva värde – från det värde livet har för henne själv – dess *reflexiva* värde. (Dessa termer introducerades i min uppsats "Människovärde", *Filosofisk tidskrift* 1986:1.) Jag skall här utgå ifrån att vad som har värde för en individ kan förstås i termer av vad som tillfredsställer vederbörandes önskningar. Ett livs reflexiva värde består således i den grad i vilken de önskemål, som den levande själv utvecklar, tillfredsställs, medan dess transitiva värde utgörs av den utsträckning i vilken detta liv bidrar till tillfredsställelsen av andras önskemål, dvs till det reflexiva värdet av dessa liv.

Det arbete en människa utför har förhoppningsvis (transitivt) värde för andra samhällsmedborgare, och den lön hon får för det ökar förhoppningsvis hennes livs reflexiva värde. Men för alla ansträngningar som har transitivt värde får man inte betalt i reda pengar – att en kvinna tar hand om sina barn, att man bispringer sina vänner, kan vara illustrationer av detta. De kan vara sin egen belöning – dvs att utföra dem kan höja det reflexiva värdet av ens liv lika mycket som det bidrar till värdet av andras liv (för dem) – eller så kan de kompenseras genom gentjänster från de gynnade.

Detta indikerar möjligheten av vägningar mellan det transitiva och det reflexiva värdet av ett liv. Sådana kommer till uttryck i begreppet *förtjänst*. Som en första approximation kan man säga att det transitiva värdet av någons liv är grunden till vad hon förtjänar i form av reflexivt värde: det senare värdet skall stå i proportion till det förra. Tanken är med andra ord att man genom att ge någon vad hon förtjänar upprättar en *rättvis balans* mellan (det transitiva) värdet för andra av hennes förehavanden och (det reflexiva) värdet för henne själv av hennes liv. Som regel är det nämligen så, att man måste göra vissa uppoffringar för egen del för att kunna bidra till andras väl. Ens förtjänst är grovt uttryckt en rättvis kompensation för denna förlust i reflexivt värde.

Det är uppenbart att såväl det transitiva som det reflexiva värdet av människors liv är olika högt, att människors liv inte är lika mycket värda i dessa bemärkelser. Det är också uppenbart att en fördelning efter förtjänst – en meritisk fördelning – inte befrämjar jämlikhet i den meningen att den tenderar att utjämna skillnader i reflexivt värde mellan olika individers liv. I många fall tenderar den tvärtom att vidga klyftorna av reflexivt värde, eftersom många av dem som kan ge stora bidrag till värdet av andras liv också är väl skickade att sörja för sig

själva, sitt livs reflexiva värde, medan t ex individer med olika handikapp har sämre förmåga att förse sitt liv både med transitivt och reflexivt värde. Om en fördelning efter förtjänst är rättvis, kommer alltså en rättvis distribution inte att gynna jämlikhet med avseende på reflexivt livsvärde.

En del tänkare – t ex franska upplysningsfilosofer som Condillac och Helvetius verkar ha förfäktat att dessa olikheter i förmåga att alstra transitivt och reflexivt värde härflyter ur skillnader i miljö eller yttre omständigheter snarare än ur människors egen natur. Så om allas sociala situation vore densamma, om alla hade föräldrar med lika god ekonomi, fick samma uppfostran och utbildning etc, skulle dessa olikheter försvinna. Man skulle då kunna hävda, för det första, att varje människas *egen* kontribution till sitt livs transitiva och reflexiva värde var densamma och att därför *människors eget* värde – fattat som en värdet av deras egen ursprungliga potential att bidra till sina livs värde – är lika, i motsats till värdet av *deras liv* som delvis är en funktion av yttre faktorer.

För det andra skulle man kunna göra gällande att det inte är *rättvist* att fördela på grundval av det transitiva värdet av *ett liv*, enär detta delvis är ett resultat av externa faktorer över vilka individen inte rår. En rättvis fördelning – som en fördelning efter förtjänst gör anspråk på att vara – måste fördela efter aspekter som individen själv förfogar över och är ansvarig för. Den måste fördela på basis av individens egen natur, konkluderar resonemanget, och alla människors natur har samma värde.

Det finns iögonfallande invändningar mot bägge dessa steg. För det första är det inte sant att om man skalar bort all yttre påverkan av människors natur når man fram till en likvärdig kärna. Om något är ursprungligt hos en människa, är det de genetiska egenskaperna, och dessa är uppenbarligen inte likvärdiga: en del har anlag för grava handikapp som radikalt sänker deras förmåga att befrämja tillfredsställelsen både av deras egna önskemål (det reflexiva värdet) och av andras (det transitiva värdet). Slutsatsen blir att inte bara värdet av människors liv är olika, utan också att deras eget värde är det, dvs deras ursprungliga potential att generera livsvärde.

För det andra har människan uppenbarligen inte ansvar för sina egna genetiska egenskaper. De ligger lika mycket utanför hennes ansvar och kontroll som socialt och miljömässigt inflytande under den tidiga barn-

domen. Om det är orättvist att distribuera reflexivt värde på basis av resultatet av detta inflytande, är det helt klart också orättvist att distribuera sådant värde på basis av hennes ursprungliga, genetiska natur. Eftersom den meritiska fördelningen efter förtjänst begreppsligen gör anspråk på att vara rättvis, underminerar detta resonemang legitimiteten av all slik fördelning. (Jämför min uppsats "Jämlikhet" i *Filosofisk tidskrift* 1991:3.)

Robert Nozick föreställer sig att det här argumentet tar för givet att "grunderna som ligger under förtjänst själva är förtjänta, *hela vägen ned*" (*Anarchy, State, and Utopia*, Basic Books, 1974, s 225) och att man kan bemöta det genom att hävda att vi i stället har naturliga *rättigheter* till några av dessa grunder, t ex att vi har rätt till vår egen kropp med dess inneboende kapaciteter. Men som jag har försökt visat annorstädes ("Har vi en naturlig rättighet till oss själva?", *Filosofisk tidskrift* 1993:3) undergräver inte bara samma, utan också en annan, typ av resonemang appellen till rättigheter.

3. Övervinnandet av de faktiska olikheterna och ett kvarstående dilemma

Om transitivt värde inte duger som rättvis distributionsgrund, verkar det som om en rättvis fördelning av reflexivt värde måste vara anpassad efter den föreliggande nivån av reflexivt värde i individers liv. Mot bakgrund av denna nivås variationer framstår det medelinriktade kravet att alla får samma tilldelning av reflexivt värde som obefogat. Rimligare är en målinriktad jämlikhetsprincip som deklarerar:

(4) Rättvisan kräver att resursfördelning sker i syfte att åstadkomma att alla kommer att leva liv av samma reflexiva värde,

så att t ex de vars liv faktiskt har lågt reflexivt värde får motta proportionellt mer av detta värde. Karl Marx' slogan "åt envar efter behov" är av detta slag (behovstillfredsställelse höjer onekligen det reflexiva värdet).

I förstone kan man tro att Peter Singers princip om "lika hänsyn till intressen" (*Praktisk etik*, Thales, 1990, s 25) sammanfaller med Marx', och han förefaller själv att tro detta (se s 40). Singers princip föreskriver ju att man skall rangordna intressen eller önskningar efter styrkegrad och först och främst tillfredsställa de starkare. Detta kan tyckas innebära att man vid distribution av en resurs enbart skall se till

var önskemålen efter eller behoven av den är som störst, var tillgången av den skulle leda till den största ökningen av reflexivt värde.

Men i ett av hans exempel visar sig "andra faktorer vara relevanta" (s 26) vid resursfördelning: i en olyckssituation tänker han sig att det kan vara berättigat att prioritera lindringen av en läkares smärta om detta sätter henne i stånd att bistå i arbetet med att lindra andra skadades smärta. Singer skriver visserligen att detta gäller "när smärtförnimmelserna är lika" (s 26), men det kan han knappast mena, givet sin utilitarism. Antag att läkarens smärta är lindrigare än andra drabbades, men likväl så svår att hon kan bidra till ett effektivare hjälparbete endast om den elimineras. Då påbjuder den utilitaristiska principen att förtur skall ges åt upphävandet av hennes smärta eftersom detta maximerar den totala smärtlindringen. Läkaren hamnar således längre fram i kön av behövande, tack vare att en "annan faktor" – hennes "nyttighet" eller förmåga att hjälpa till – än hennes behov är relevant.

I klartext innebär alltså detta att man vid resursfördelning tar hänsyn inte bara till människors behov, till nivån av deras reflexiva livsvärde, utan också till deras livs transitiva värde. (Den marxistiska rekommendationen – som i sin helhet lyder "Från envar efter förmåga, åt envar efter behov" – tycks missa detta, att det kan vara nödvändigt att ge åt folk för att få dem att utnyttja sina förmågor.) Eftersom det transitiva värdet av människors liv, som konstaterats, är olika stort leder en sådan distribution inte till största möjliga utjämning av det reflexiva värdet av deras liv, utan till att de mer "nyttiga" relativt sett får det bättre.

Det kan verka som om Singers utilitarism, precis som den ovan granskade meritiska tankegången, vid fördelning beaktar individers *förtjänst*, men det finns en signifikant skillnad. För utilitaristen rättfärdigas denna fördelning noga taget inte av att den mottagande har gjort sig *förtjänt* av den, dvs av att den är *rättvis* relativt det transitiva värdet av vederbörandes liv, utan av att den *maximerar den totala summan av reflexivt livsvärde*. Medan meritister hänvisar till rättvisa, hänvisar utilitarister till maximering av den sammanlagda summan av det reflexiva värdet i alla liv. Den förra modellen blickar *tillbaka* på värdet av vederbörandes bidrag och söker en tilldelning vars reflexiva värde står i proportion till det, medan den andra söker den rätta fördelningen genom att blicka *framåt* mot de (framtida) konsekvenser olika fördelningar skulle få.

Denna skillnad är viktig, eftersom vi har sett att det inte är rättvist att distribuera efter förtjänst emedan de varierande (genetiska och miljöbetingade) förutsättningar som ligger till grund för individers olika förtjänster strös ut utan moralens samtycke. Ur ett *rättvise-perspektiv* måste således människors – olika – förtjänst (och naturliga rättigheter) – eller med andra ord, differenserna dem emellan med avseende på transitivt värde – förkastas som distributionsfundament. En rättvis fördelning kommer då att bestå i att de får ransoner som i största möjliga mån utjämnar rådande differenser i reflexivt värde – dvs den ovannämnda principen (4). Så tänker jag mig alltså att det uppställda jämlikhetsmålet kan rättfärdigas. (Det finns inte här utrymme att diskutera reservationer av (4), varav den väsentligaste är "om de inte *frivilligt avstår* från värde".)

Eftersom man i stället för att säga att människor förtjänar X kan säga att de är *värda* X, kan man också kalla denna princip för en *lika-värdesprincip*, som hävdar att människor är värda att leva liv av samma reflexiva värde. Denna princip är fullt förenlig med att det *föreligger* differenser i deras livs transitiva och reflexiva värde; den uppställer ju ett *ett mål* som skall realiseras. Dessa två livsvärden måste därför hållas isär: det (reflexiva) livsvärde som är rättvisans mål, å ena sidan, och det faktiska (transitiva och/eller reflexiva) livsvärdet, å den andra.

Det kan förefalla som om vi därmed har nått vårt mål – en jämlikhetsprincip som äger giltighet fastän människors faktiska värde skiljer sig åt – men en komplikation återstår: vi har visserligen dragit slutsatsen att människor är värda (=förtjänar) detsamma, men till priset av konstaterandet att vad de är värda är – ingenting! Resonemanget rättfärdigar därför inte intuitionen att det är att föredra att alla människor lever liv av samma *höga* värde framför att de lever liv av samma *låga* värde. Det är inte *orättvist* att ge dem liv av lågt reflexivt värde i stället för högt, om eländet drabbar dem alla utan undantag, och ingen av dem förtjänar något.

En utväg ur denna svårighet öppnar sig när vi betänker att den utilitaristiska principen inte har vederlagts. Ovan förda resonemang implicerar förvisso att den inte duger som *rättvise*princip, men därav följer inte att den har satts ur spel som *moralisk* princip. Moralen behöver inte bestå enbart av rättvisa. Det är plausibelt att tänka sig att den också inkluderar en *värdemaximeringsprincip* som säger att det är

bättre att varelser åtnjuter mer (reflexivt) värde än mindre. (Jag begick ett misstag när jag i "Jämlikhet" bakade in maximeringsövervägandena i en rättviseprincip.) Med denna princip i ryggen kan vi moraliskt rättfärdiga att vi ger människor liv av högt snarare än lågt reflexivt värde.

Denna plausibilitet har emellertid ett pris. Om vi accepterar en moral bestående dels av en egalitär rättviseprincip, dels av en utilitaristisk maximeringsprincip, uppstår rangordningsproblem av detta slag: är ett alternativ där ett mindre värdetillskott helt tillfaller de sämst ställda, och som därmed minskar ojämlikheten, bättre än ett större tillskott som helt eller till större delen tillfaller de välbeställda och som därmed vidgar ojämlikheten?

En del kanske menar att detta är problem som inte uppkommer i praktiken därför att – enligt principen om "decreasing marginal utility" – lyfter ett resurstillskott värdenivån mest där den på förhand är lägst: att låta det tillfalla de sämst ställda är således inte bara jämlikhetsbefrämjande utan också värdemaximerande. Men det transitiva värdets relevans för distribution får inte förbises: med hänvisning till detta värde kan det vara möjligt att rättfärdiga att de välbeställda får mer om de besitter större förmåga att höja levnadsstandarden i hela samhället, och denna belöning är nödvändig för att motivera dem att utnyttja sin förmåga. Vi kan undkomma dylika dilemman endast om det transitiva värdet av människors liv görs lika, men detta inte kan ske med mindre än att vi manipulerar människors gener (på ett vis som f n är tekniskt omöjligt). Så mitt resonemang utmynnar i ett moralsystem med (åtminstone) två grundprinciper som divergerar i vissa omständigheter, men vars prioritetsproblematik inte kan diskuteras här.

JAN ÖSTERBERG

Hur många tolkningar har "Macbeth"?

1. Monism och pluralism

Vissa litteraturteoretiker hävdar att varje litterärt verk har en enda riktig (sann) tolkning. Andra teoretiker hävdar att tolkningar inte kan vara riktiga eller sanna; de är på sin höjd plausibla (rimliga), och det kan, menar man, mycket väl finnas flera, lika plausibla, tolkningar av ett verk. Ett förtydligande: denna motsättning mellan *(tolknings)monism* och *(tolknings)pluralism* gäller naturligtvis den totala (sammanlagda) tolkningen av verket. Vad monisterna hävdar (och pluralisterna bestrider) är alltså att det bara kan finnas en enda riktig *totaltolkning* av ett litterärt verk.

Både monister och pluralister är realister, dvs de antar att ett litterärt verk har objektivt fastställbara egenskaper, egenskaper som föreligger oberoende av läsarnas uppfattningar av och reaktioner inför verket. Men varje realism har sin antirealism – så ock den litterära realismen. Den litterära antirealismen, företrädd av bl a poststrukturalistiska litteraturteoretiker och av s k läsarreaktionsteoretiker, hävdar att verket är något som läsaren konstruerar vid sin läsning. Och olika läsare, i varje fall om de tillhör olika historiska epoker, kan, utgående från samma litterära produkt, konstruera olika verk – och ingen konstruktion är rätt och ingen är fel.

Själv är jag anhängare av litterär realism. Men att försvara den mot antirealistens argument faller utanför ramen för denna uppsats. Den fråga som jag skall försöka besvara är alltså: givet att den litterära realismen är riktig, vem har rätt, monisten eller pluralisten? Jag skall först göra några distinktioner mellan olika slags tolkningar. Sedan skall jag diskutera de huvudargument som har framförts av företrädarna för de två positionerna. Till sist skall jag argumentera för mitt svar på frågan.

2. *Tolkningsbegreppet*

Vad innebär det att tolka ett litterärt verk? Begreppet litterär tolkning rymmer olika slags aktiviteter (och resultaten av dessa). Motsättningen mellan monister och pluralister beror till rätt stor del, anser jag, på att man inte skilt mellan olika slags tolkningar. Jag skall i det följande redogöra för några olika indelningar av tolkningar.

En viktig och användbar indelning av tolkningar har gjorts av Beardsley, som skiljer mellan *explikation*, *elucidation* och *interpretation* (i inskränkt mening). Att *explicera* en text är att fastställa dess språkliga mening. Som Beardsley betonar är detta mycket mer komplicerat i fråga om litterära än i fråga om icke-litterära texter.

> Because it contains deep levels of meaning that are only hinted at through connotation and suggestion, a literary discourse has a kind of semantical *thickness* when compared with mathematical and technical discourse (Beardsley, 1958: 129).

Det andra slaget av tolkande verksamhet, *elucidation*, har som objekt det litterära verkets fiktiva värld. Att elucidera ett verk är att söka fastställa viktiga sakförhållanden i dess fiktiva värld som inte är explicit framställda i verket. Som exempel på frågor vars svar innebär en elucidation anför Beardsley följande:

> Is Hamlet mad? What is Raskolnikov's real motive in killing the old woman? ... What traits are basic to the character of Antigone? Is the "Lydian Stranger" of Euripides' *The Bacchae* ... Dionysus in disguise? (Beardsley, 1958: 242).

Det tredje slaget av tolkning slutligen, *interpretation* (i inskränkt mening), gäller verkets "tema" eller dess "tes": en övergripande idé respektive en generell utsaga om verkligheten. (Beardsley menar att de flesta verk har ett tema men inte nödvändigtvis en tes.)

> We say that the theme of *Wuthering Heights* is the quest for spiritual contentment through harmony with both good and evil forces of nature; of *War and Peace*, the endless rhythmic alternation of youth and age, life and death, ambition and resignation. ... Critics say, for example, that Shakespeare's *Tempest* embodies a mystical view of life; that Upton Sinclair's *The Jungle* is a protest against the injustices suffered by the poor under a freewheeling economic system (Beardsley, 1958: 403 f).

(Beardsleys indelning av tolkningar är inte helt oproblematisk: inte alla slag av litterär tolkning har en självklar plats i Beardsleys schema, och det är inte helt klart vilka indelningsgrunderna är. Men indelningens brister saknar betydelse i det här sammanhanget, och en revidering får anstå till ett annat tillfälle.)

En annan viktig indelning av tolkningar är en indelning med avseende på tolkningens syfte. För att beteckna de två kategorier som jag här intresserar mig för skall jag, i brist på bättre, ta i bruk det gamla slitna begreppsparet *objektiv – subjektiv*. Syftet med en *objektiv* tolkning är att bidraga till en bättre, en mer fördjupad, förståelse av verket. Därför åberopar den som gör en sådan tolkning som stöd för sin tolkning endast (1) (vad han menar vara) sådana språkliga, litterära och andra konventioner samt övertygelser som var allmänt omfattade i den miljö i vilken verket tillkom och/eller (2) författarens intentioner. (Den som inte känner sig övertygad om att endast tolkningar som åberopar (1) och/eller (2) bidrar till en bättre förståelse av verket kan i stället förstå "objektiv tolkning" som liktydigt med "tolkning som endast åberopar (1) och/eller (2)".)

Syftet med en *subjektiv* tolkning är att göra ett litterärt verk estetiskt, idémässigt eller i annat avseende relevant (intressant) för läsaren. Därför åberopar den som gör en sådan tolkning som stöd för sin tolkning även sådana språkliga, litterära och andra konventioner samt övertygelser som inte var omfattade i den miljö i vilken verket tillkom, och han känner sig inte bunden av författarens intentioner. (Det kan noteras att subjektiva tolkningar även kan manifesteras icke-verbalt, t.ex i pjäsuppsättningar. När t ex en regissör som sätter upp ett antikt drama låter kören vara klädd i moderna kläder, är väl tanken att dramat har något viktigt att säga om problem som vi, nutida åskådare, står inför.)

Låt oss, för att åskådliggöra skillnaden, tänka på en freudiansk tolkning av Sophokles' *Kung Oidipus*. Som bekant dödar Oidipus i vredesmod en äldre man och gifter sig senare med drottningen i Thebe, som nyligen blivit änka. Han är inte medveten om att mannen som han dödar är hans far och att kvinnan han gifter sig med är hans mor. Enligt Freuds teori har emellertid Oidipus en omedveten insikt i båda dessa förhållanden, och det är denna insikt som, tillsammans med hans oidipuskomplex, förklarar både dråpet och ingåendet av äktenskapet. Varje man, säger Freuds teori, har ett oidipuskomplex,

dvs en omedveten önskan (drift) att döda sin far och "äkta" sin mor.

Om den freudianska tolkningen av Sophokles' drama är en objektiv tolkning, så tror alltså tolkaren att Sophokles och hans samtida accepterade den freudianska teorin om oidipuskomplexet, om än på det undermedvetna planet. (Detta förutsätter väl att teorin är sann, för det är väl högst implausibelt att de skulle ha gjort det annars.) Om däremot den freudianska tolkningen av dramat är en subjektiv tolkning, så behöver tolkaren inte tro att Sophokles och hans samtida omedvetet accepterade denna freudianska teori eller att den är sann; det räcker att tolkaren tror att *hans* samtida är förtrogna med den och åtminstone finner den intressant.

En subjektiv freudiansk tolkning är ju en subjektiv elucidation. Det finns också gott om subjektiva explikationer och interpretationer. Först ett exempel på en subjektiv explikation. I inledningsdikten till William Blakes *Milton* förekommer frasen "dark satanic mills". Den engelske kritikern John Wain föreslår en tolkning av denna fras enligt vilken den refererar till de rökutspyende fabriker som den industriella revolutionen begåvade den engelska landsbygden med, fastän (något som Wain är medveten om) Blake av historiska skäl inte kan ha syftat på dessa. Det skäl för denna tolkning som Wain ger är att den är "manifestly superior... to the hotch-potch of altars, legends and repressionsymbols that Blake 'really meant'" (Wain m fl, 1952; cit efter Juhl, 1986: 124).

Som exempel på en subjektiv interpretation kan anföras Olof Lagercrantz' symboliska tolkning av paradiset i Dantes *Divina Commedia*:

> Vad är alltså till sist paradiset? Paradiset är inte en ort långt ovan jorden och kanske är den inte heller en belöning till själarna efter döden. Paradiset är här och nu. ... Paradiset är ett tillstånd av kärlek och den som nått dit vet att landsflykten är över, att hon är hemma, att hon inte längre har behov av ett uppbrott (Lagercrantz, 1985: 176 f).

Ingenting i Dantes verk tyder på att paradiset enbart är en symbol för ett jordiskt tillstånd eller att *Komedins* avslutningsord om den kärlek "som rör sol och andra stjärnor" syftar på människors kärlek. Men Lagercrantz' tolkning är naturligtvis en utmärkt subjektiv tolkning – med förmåga att göra Dantes verk intressant för en nutida sekulariserad läsekrets.

Av subjektiva tolkningar finns det flera olika slag, men att skilja dessa från varandra är inte så viktigt i detta sammanhang och får anstå

till en annan gång. Viktigare i detta sammanhang är att betona den stora skillnaden mellan, å ena sidan, subjektiva tolkningar av alla slag och, å andra sidan, objektiva tolkningar – en skillnad väl så stor som den mellan endera av dessa och det slags tolkning som vi sägs göra automatiskt vid all perception.

Innan jag lämnar diskussionen av tolkningsbegreppet vill jag tillägga att man med hjälp av distinktionen objektiv – subjektiv tolkning nog kan slita följande tvist rörande tolkningars ontologiska status. Vissa litteraturteoretiker menar att tolkningssatser uttrycker propositioner. Andra, t ex Charles Stevenson (Stevenson, 1950), har i polemik mot denna uppfattning hävdat att tolkningssatser är ateoretiska och uttrycker föreskrifter. Enligt min mening är den förra uppfattningen riktig vad gäller objektiva tolkningar, den senare vad gäller subjektiva tolkningar.

3. Argument för pluralism

Jag skall nu övergå till att redogöra för argument som framförts till stöd för tolkningspluralism. Enligt en av de mest kända företrädarna för denna ståndpunkt, Joseph Margolis, råder det en avgörande skillnad mellan att beskriva och att tolka ett konstverk:

> In description, the emphasis is upon an object independent of any particular effort of description, an object that has or has not the properties attributed to it. ... In interpretation, the emphasis is upon performance, on what is added beyond the mere materials provided (Margolis, 1965: 71. Margolis har senare anammat en radikalare uppfattning och är numera antirealist; se Margolis, 1980 och 1995).

Vad den litteräre uttolkaren gör enligt Margolis är att framlägga vad Margolis kallar en "myt". En myt är

> a schema of the imagination which is independent of the scientific status of the propositions it may subtend, is capable of effectively organizing our way of viewing portions of the external world in accord with its distinctions. (Margolis, 1965: 89)

Det spelar enligt Margolis ingen roll om myten är sann eller inte, och det spelar heller ingen roll om uttolkaren eller verkets upphovsman accepterar myten eller inte. Det räcker att den "is so much part of our general culture that the imagination both of artists and of semieducated

persons is saturated with it" (Margolis, 1965: 87).

Som exempel på sådana tolkningsmyter anger Margolis kristna, marxistiska och freudianska teorier eller teorifragment. Som ett kuriosum nämner han sitt eget försök att tolka Tennessee Williams' pjäs *Linje lusta* som en inversion av Platons grottmyt (Margolis, 1965: 95).

Eftersom varje konstverk i princip kan tillskrivas flera inbördes oförenliga myter, finns det enligt Margolis inga sanna eller riktiga ("correct") tolkningar, endast mer eller mindre plausibla. Och detta, menar Margolis, är också något som litteraturtolkare i praktiken accepterar; det som utmärker litterära tolkningar är "its tolerance of alternative and seemingly contrary hypotheses" (Margolis, 1965: 87).

Det är väl uppenbart att Margolis med tolkningar avser (vad jag kallar) subjektiva tolkningar, i första hand subjektiva elucidationer och på dessa baserade subjektiva interpretationer. Och det är väl också uppenbart att när det gäller subjektiva tolkningar har tolkningspluralisterna rätt: sådana tolkningar är inte riktiga eller oriktiga utan mer eller mindre rimliga (plausibla).

Vad gäller Margolis' argument att tolkningspraxis uppvisar tolerans gentemot alternativa, till synes motsatta tolkningar, så äger nog detta sin riktighet i fråga om subjektiva tolkningar – och något annat vore väl också anmärkningsvärt. Men när det gäller (vad jag kallar) objektiva tolkningar är bilden en annan. Här argumenterar man friskt mot varandras tolkningar och försöker visa att de är oriktiga: att de kommer i konflikt med vad som uppenbart står i texten eller att de ger en sämre förklaring av detta än vad den egna tolkningen gör. Ett exempel bland många som kan illustrera denna brist på tolkningstolerans är den debatt som förts bland svenska litteraturvetare och kritiker om Stagnelius' dikt "Endymion". Vem åsyftas med frasen "rosenkransade brud på saffransbädden", mångudinnan Diana eller morgonrodnadens gudinna Aurora?

En annan förespråkare för tolkningspluralism är Peter Jones. I sin *Philosophy and the Novel* hävdar han att tolkningar kan kritiseras som "'implausible', 'pointless' or 'far-fetched'" men inte (annat än i ytliga avseenden) som oriktiga. Att även Jones med tolkningar avser subjektiva tolkningar framgår tydligt på flera ställen i hans bok. Visserligen nämner Jones på ett ställe existensen av (vad jag kallar) objektiva tolkningar, men hans beskrivning av dessa är en karikatyr. Om man ägnar sig åt att tolka men inte är intresserad av tolkningar av det slag

som Jones är intresserad av, är man (sägs det) "interested primarily in the historical question of how the author actually conceived the significance of his text, in the light of the precise context in which he saw himself" (Jones, 1975: 184).

Men detta är uppenbart fel. Få, om ens någon, utövare av eller förespråkare för objektiv tolkning är anhängare av en sådan stark version av intentionalism, och flera, t ex Beardsley, är ju uttalade antiintentionalister. De flesta objektivister skulle säkert acceptera Herman Melvilles försäkran att när han skrev *Moby Dick* hade han en vag uppfattning att verket kunde ges en allegorisk tolkning; vari denna närmare bestod gick dock upp för honom först sedan han tagit del av de tolkningsförslag som hans författarkollega Nathaniel Hawthorne hade framlagt (se Barnes, 1985: 57).

För att sammanfatta diskussionen så här långt: Förespråkarna för pluralism, menar jag, har nog till övervägande del haft subjektiva tolkningar i tankarna när de argumenterat för sin tes. De har visat – något som dock knappast behöver visas – att pluralismen är riktig när det gäller subjektiva tolkningar. Men de har inte visat att så är fallet också när det gäller objektiva tolkningar. Men detta innebär givetvis inte att monismen är riktig när det gäller objektiva tolkningar. Är den det? Låt oss se vilka argument som har framförts av monisterna. (De har nog utan undantag haft objektiva tolkningar i tankarna.)

4. Argument för monism

Ett argument för monismen lyder: När det finns flera möjliga tolkningar av en text, väljer i allmänhet en uttolkare en av dessa, ibland även när, på basis av den evidens som texten ger, flera tolkningar är *prima facie* lika möjliga eller plausibla (Juhl, 1986: 208).

Men detta argument visar väl inte att monismen är riktig. På sin höjd visar det att kritiker *tror* att monismen är riktig. Men visar det ens det? Att uttolkare ibland väljer en bland flera *prima facie* lika plausibla tolkningar visar väl bara att de inte accepterar den starka och föga plausibla form av pluralism som säger att det *aldrig* finns en och endast en plausibel tolkning av ett verk. Det visar väl inte att uttolkare förkastar den svagare och mer plausibla form av pluralism som säger att det *ofta* inte finns en och endast en plausibel tolkning av ett verk.

Att uttolkare ibland finner flera elucidationer *prima facie* lika plausibla men väljer att argumentera för en av dem såsom den riktiga

kan ju ges en plausibel psykologisk förklaring. När det gäller tolkningar av mänskligt beteende i vår vanliga värld, tänker vi – med rätta – att det finns en enda riktig tolkning, även om vi ofta är osäkra på vilken den är. Den verklighetsillusion som det litterära verket skänker har nog en tendens att suggerera något motsvarande om tolkningar av händelser i den litterära världen.

Till detta kommer att personliga och kulturella särdrag nog kan ha som effekt att för tolkare *A* framstår en av de möjliga tolkningarna som mest spännande, intressant eller psykologiskt trovärdig, för tolkare *B* någon annan av dessa tolkningar. (Att personliga och kulturella särdrag influerar valet av tolkning kan vi se i fallet med den anonyme deltagaren i Gunnar Hanssons läsarundersökningar som gått till historien genom att tolka Runebergs dikt "Den enda stunden" som diktjagets möte med en älg (Hansson, 1970: 274 f).

Ett annat av Juhls argument (Juhl, 1986: 203 ff) lyder: Pluralism är rimlig endast om man föreställer sig meningen hos en litterär text som analog med meningen hos en sats *in abstracto*. Men detta, menar Juhl, är en felaktig föreställning; i stället är meningen hos en litterär text analog med meningen hos ett yttrande av en sats i en viss kontext, dvs en språkhandling.

Juhl illustrerar med hjälp av följande exempel. Betrakta satsen:

Cary saw the girl laughing at John.

Denna sats är mångtydig; den kan betyda detsamma som:

(i) Cary saw the girl as she was laughing at John,

eller detsamma som:

(ii) Cary saw the girl who is laughing at John.

Satsen har uppenbart bägge dessa betydelser, och den som valde en av dessa och framhärdade att den har denna och inte den andra betydelsen skulle bara visa sin bristande förtrogenhet med engelska språket. Om däremot en person yttrar denna sats, är det rimligt att tolka hans yttrande som antingen liktydigt med (i) eller (ii).

Jag tror att Juhl har rätt i att det finns litteraturtolkare som uppfattar meningen hos en litterär text som analog med meningen hos en sats *in abstracto*. Jag tror också att de har fel: meningen hos en text är mer analog med meningen hos ett yttrande av satsen. Men jag tror inte att

pluralism är oförenlig med denna uppfattning. Även yttranden kan vara mångtydiga. Knappast någon behärskar sina språkliga uttrycksmedel så väl att vad han/hon yttrar alltid är entydigt – inte ens givet kontexten. Och när det gäller längre yttranden (texter) med lång produktionstid tillkommer ytterligare en faktor. Även om varje del är entydig, kan ändå helheten vara mångtydig. Författaren kan ändra uppfattning under arbetets gång, och, eftersom vi är dåliga läsare av våra egna texter, kan det, trots alla hans bemödanden att skapa enhetlighet, ändå kvarstå motsägelser. Flera olika tolkningar av texten är då möjliga.

Enligt pluralister är ju tolkningar inte sanna eller riktiga utan på sin höjd "rimliga" ("reasonable") eller "plausibla" ("plausible"). Ett argument för monism, ett *argumentum ad hominem* riktat mot pluralister, som framförts av bl a Beardsley, går ut på att om man hävdar att tolkningar kan vara rimliga eller plausibla, så har man *eo ipso* gått med på att de kan vara sanna. Beardsley skriver:

> For I do not see how an interpretation could be reasonable unless reasons can be given to show its superiority to some alternatives; and I do not see how the reasons could count unless they are reasons for thinking it true. ... But plausibility is at least an appearance of truth based upon some relevant evidence, and any statement that is plausible must be in principle capable of being shown to be true or false (Beardsley, 1970: 42 f).

Mot detta skulle en pluralist kunna invända att man mycket väl kan hävda att en tolkning är rimlig (reasonable) utan att man därför hävdar att det finns skäl att tro att den är sann. En nonkognitivist på moralens område kan mycket väl anse att vissa moraliska övertygelser är rimliga, t o m rationella – och det är väl ingalunda självklart att han har inkonsistenta övertygelser. (Och motsvarande gäller om "plausibel".)

5. Min teori

Jag anser att det är ställt utom allt tvivel att det förekommer fall där vi har flera, lika eller nästan lika, plausibla tolkningar av ett verk. Vi har, för det första, fall där författaren förmodligen avsiktligt skapat mångtydighet. Ett sådant exempel är Henry James' berättelse *The Turn of the Screw*, där det med (i varje fall nästan) samma grad av plausibilitet kan göras gällande att berättelsens huvudperson verkligen ser spöken som att hon bara inbillar sig detta.

Men även om författaren inte, som i detta fall, avsett att flera olika elucidationer av ett verk är lika plausibla, kan det ändå vara så. Texten medger helt enkelt att verket tolkas på flera, inbördes oförenliga, plausibla sätt. Även om den intentionalistiska tesen, dvs tesen att författarens intentioner är relevanta för tolkningen av verket, är riktig, så hjälper den oss inte här. För författarens intentioner kan vara otillgängliga för oss. Och även om de vore tillgängliga, skulle det nog ofta visa sig att de inte har tillräckligt djup för att kunna slita tvisten: författaren har helt enkelt inte ägnat saken en tanke. Och även om han gjort det, är det nog ofta fallet att han själv varit odeciderad eller haft motstridiga uppfattningar, samtidigt eller successivt.

Ett exempel på hur en text har tolkats på olika och, såvitt man kan avgöra, lika plausibla sätt anförs av Torsten Pettersson (Pettersson, 1988). Det gäller Grays berömda "Elegy on a Country Churchyard". Jag anser att Pettersson har rätt i att det finns flera lika plausibla, inbördes oförenliga, tolkningar av Grays dikt. Jag anser också att detsamma gäller om flera andra litterära verk. Men detta kan inte visas här och nu. Jag får nöja mig med att hänvisa till Petterssons bok.

Det är väl klart att om vi har flera lika plausibla, inbördes oförenliga, helhetstolkningar av ett verk kan inte alla dessa vara sanna (riktiga). Man kan t ex inte hävda både att det förekommer och att det inte förekommer spöken i *The Turn of the Screw*. Inte heller kan man med fog hävda om en viss av dessa tolkningar att den är sann. Vad man kan hävda, och vad jag vill hävda, är att disjunktionen av dessa plausibla tolkningar är sann, trots att ingen disjunkt är sann. I (vad jag menar vara) standardfallet, fallet när det endast finns en plausibel tolkning av ett verk, är denna tolkning också sann. Och så länge som det finns åtminstone en plausibel tolkning av ett verk, finns det också exakt en sann tolkning av det, bestående av en eller flera disjunkter. Den ståndpunkt jag vill förespråka i kontroversen mellan monism och pluralism är en version av monism som kan kallas *svag* monism.

Denna svaga monism skiljer sig från pluralismen i följande avseende. Givet en mängd av lika plausibla tolkningar av ett diktverk säger pluralismen att var och en av dessa är en acceptabel tolkning. (Det är helt legitimt att någon omfattar någon av dessa som *den* acceptabla tolkningen.) Den svaga monismen däremot säger att, eftersom varje enskild tolkning har lika plausibla konkurrerande tolkningar, är ingen enskild tolkning, endast disjunktionen av dessa, acceptabel – och riktig (sann).

Litterära tolkningar liknar i mångt och mycket hypoteser i vetenskap och vardagsliv. Särskilt gäller detta elucidationer – och det är sådana jag tänker begränsa mig till i resten av min framställning. Den stora skillnaden är väl att när det gäller lika plausibla men oförenliga hypoteser om den empiriska verkligheten är inte bara deras disjunktion (sannolikt) sann. Också någon av disjunkterna, dvs någon av hypoteserna, är (sannolikt) sann – även om vi inte vet vilken. Åtminstone är det i många fall rimligt att tänka sig det.

Men när det gäller litterära elucidationer är läget i många fall annorlunda: vissa verk innehåller "ontologiska luckor". Vissa verk – t ex *The Turn of the Screw*, "Elegy on a Country Churchyard", Kafkas *Processen* – framställer fakta som inte får någon förklaring i verket självt. Vid tolkningen av ett sådant faktum kan man ibland ge flera, lika plausibla, men inbördes oförenliga förklaringar. Låt F_1, F_2, ..., F_n vara de lika plausibla men inbördes oförenliga förklaringarna av ett sådant faktum X. Förklaringen av X ges då av disjunktionen F_1 eller F_2, ..., eller F_n. Men ingen av disjunkterna, dvs de enskilda förklaringarna, är (sannolikt) sann. Det litterära verkets värld består av de objekt och de fakta som framställs i verket samt av de fakta som ger bästa förklaringen till de fakta som framställs i verket. När flera bästa förklaringar kan ges, kan man inte säga att någon av dessa, men vi vet inte vilken, trots allt är den sanna förklaringen, den som är sann om den diktade världen. För det finns inget faktum här: den diktade världen har här en genuin ontologisk lucka.

Det finns en parallell till detta. Vi accepterar en tolkning som riktig därför att vi anser att den har ett bättre induktivt stöd än vad alternativa tolkningar har. Det finns andra påståenden om ett litterärt verk som vi accepterar, inte för att de ges ett induktivt stöd, utan därför att de ges ett slags deduktivt stöd av verket. Jag tänker på sådana påståenden som impliceras av vad som sägs i verket, i förening med vissa generella fakta och/eller litterära konventioner. Det sägs t ex aldrig i Stendhals *Rött och svart* att huvudpersonen, Julien Sorel, har lever och njurar, men det impliceras väl av det som sägs om honom i romanen, i förening med vissa empiriska fakta om mänsklig fysiologi samt de litterära konventioner som gäller för realistiska romaner. Även i fråga om sådana påståenden uppträder liknande fenomen som i fråga om tolkningar: ibland är det påstående som impliceras av verket en disjunktion, och det är inte fallet att någon bestämd av dessa disjunkter,

vi vet bara inte vilken, är sann. Även i dessa fall finns det ontologiska luckor. (Detta har dock ifrågasatts; se Skalin, 1991.)

Ett klassiskt exempel på detta ges i den kända uppsatsen "How Many Children Had Lady Macbeth?" Lady Macbeth hade barn, eftersom det i skådespelet talas om hennes "children". Det sägs visserligen inte hur många barn hon hade, men om en person har barn, har han eller hon ett bestämt antal. Så exakt hur många barn hade Lady Macbeth – ett, två, tre, ...? Svaret är naturligtvis att här finns en genuin lucka i Shakespeares text. Vad som är sant är att ladyn hade antingen ett eller två eller tre eller ... barn. Men det finns inget tal *n* sådant att, trots att vi inte kan veta det, hon faktiskt hade *n* barn. Varken texten eller språkliga och litterära konventioner implicerar ett bestämt antal. Och även om författarens intentioner är relevanta, så är inte detta till någon hjälp i det här fallet. Vi kommer aldrig att få reda på Shakespeares åsikt i denna fråga – och det troliga är att han inte hade någon.

(En skillnad mellan tolkningar och den här sortens påståenden om ett litterärt verk är följande: den som hävdar att en disjunktion av tolkningar är sann avser i standardfallet en bestämd uppsättning disjunkter, medan den som accepterar en disjunktion som (i ovan mening) impliceras av verket nog av empiriska skäl vill begränsa disjunktionens omfång, men är osäker på exakt hur många disjunkter som skall tas med. Lady Macbeth hade med säkerhet inte mer än tusen barn, med rätt stor säkerhet inte mer än tio barn och troligen inte mer än fem barn. Stort mer kan man inte säga.)

Som kanske någon noterat, grundar sig min teori om litterära tolkningars sanning på en idé om sanning som påminner om van Fraassens teori om *superevalueringar* – (se t ex van Fraassen, 1960) – ja man kan säga att den generaliserar van Fraassens idé i en viss riktning. Även för van Fraassen är utgångspunkten frågan om vilket sanningsvärde man skall tillskriva sammansatta satser vars komponenter saknar sanningsvärde. van Fraassens förslag är att man skall tillskriva en sådan komplex sats det sanningsvärde som alla "klassiska" evalueringar skulle tillskriva den, om det finns ett sådant värde, och annars inget värde. Eftersom alla klassiska evalueringar, dvs både de som tillskriver *p* värdet sant och de som tillskriver *p* värdet falskt, tillskriver *p v -p* värdet sant, så tillskriver också superevalueringen denna komplexa sats värdet sant. Men eftersom vissa klassiska evalue-

ringar tillskriver $p \vee q$ värdet sant och andra tillskriver den värdet falskt, tillskriver superevalueringen den inget sanningsvärde alls.

Enligt van Fraassens teori om superevaluering är alltså en komplex sats vars komponenter saknar sanningsvärde sann om och endast om den är en tautologi. Min tolkningsteori ställer ett svagare krav: en disjunktion av tolkningssatser är sann om och endast om disjunktionen är empiriskt nödvändig i den fiktiva världen i fråga. Om alltså p_1, p_2, ..., p_n är mängden av alla plausibla tolkningar av ett verk, så är disjunktionen av dessa empiriskt nödvändig och sann, medan disjunktionen av elementen i en äkta delmängd av denna mängd inte är empiriskt nödvändig och alltså inte sann.

Så, för att sammanfatta mitt föredrag: Hur många tolkningar har *Macbeth*? Svar: legio. Men hur många *riktiga* tolkningar har *Macbeth*? Svar: en, troligen av disjunktiv natur.

Litteratur

ANNETTE BARNES, On Interpretation: A Critical Analysis (Blackwell, 1988).
MONROE C BEARDSLEY, Aesthetics: Problems in the Philosophy of Criticism (Harcourt, Brace & World, 1958).
MONROE C BEARDSLEY, The Possibility of Criticism (Wayne State University Press, 1970).
GUNNAR HANSSON, Dikten och läsaren (Prisma, 1970 (1959)).
PETER JONES, Philosophy and the Novel (Clarendon Press, 1975).
P D JUHL, Interpretation: An Essay in the Philosophy of Literary Criticism (Princeton University Press, 1986 (1980)).
OLOF LAGERCRANTZ, Från helvetet till paradiset (Wahlström och Widstrand, 1985 (1964)).
JOSEPH MARGOLIS, Art and Philosophy (Humanities Press, 1980).

JOSEPH MARGOLIS, "Plain Talk about Interpretation on a Relativistic Model", *The Journal of Aesthetics and Art Criticism*, Vol. 53 (1995).
JOSEPH MARGOLIS, *The Language of Art and Art Criticism: Analytic Questions in Aesthetics* (Wayne State University Press, 1965).
TORSTEN PETTERSSON, *Literary Interpretation: Current Models and a New Departure* (Åbo Academy Press, 1988).
LARS-ÅKE SKALIN, *Karaktär och perspektiv* (Acta Universitatis Upsaliensis: Historia Litteraria, 1991).
CHARLES STEVENSON, "Interpretation and Evaluation in Aesthetics", Max Black (utg), *Philosophical Analysis: A Collection of Essays* (Cornell University Press, 1950).
BAS VAN FRAASSEN, "Singular Terms, Truth-Value Gaps, and Free Logic", *The Journal of Philosophy*, Vol. 63 (1960).
JOHN WAIN m fl, "'Intention' and Blake's *Jerusalem*", *Essays in Criticism*, Vol 2 (1952).

PETER PAGIN

Offentlighet och obestämdhet hos språklig mening

Denna uppsats är ägnad åt en ganska begränsad fråga: Är obestämdhet hos översättning en konsekvens av offentlighet hos språklig mening?

Så som jag har uppfattat W.V. Quine, anser han att svaret är ja. Jag kommer att ge ett visst stöd för denna tolkning. Själv anser jag att svaret är nej, vilket också är rimligt, eftersom jag gillar offentlighet och ogillar obestämdhet. Jag tänker emellertid inte här försöka visa att svaret är nej. Jag vet inte hur det ska göras, eller ens om det är möjligt.

Vad jag ska göra är i stället att undersöka en speciell tankegång hos Quine, ett sätt att resonera från offentlighetstesen till obestämdhetstesen. Jag kommer att rekonstruera tankegången till ett explicit argument, och försöka visa att ett sådant argument inte kan vara framgångsrikt. I vilken bemärkelse det inte kan vara framgångsrikt är inte helt lätt att säga i förväg, eftersom det beror på egenheter i argumentets struktur, men det kommer att bli klart senare.

1. Offentlighet hos språklig mening

Tanken att språklig mening är offentlig är central under en mycket lång period av 1900-talets språkfilosofi. Den förekommer i olika versioner hos Wittgenstein, hos Quine, Donald Davidson, Michael Dummett och andra. Det gemensamma hos dessa ligger i första hand i avvisandet av en viss bild av språklig mening, enligt vilken privata, mentala faktorer, epistemiskt otillgängliga för andra än talaren själv, bestämmer vad dennes ord betyder. Avvisandet av denna bild tar sig ibland uttryck i teser som uttryckligen handlar om kunskap om språklig mening, ibland i teser om väsendet hos språklig mening.

Den definition Fredrik Stjernberg slutligen väljer i sin avhandling (s 18) är följande:

Meaning is accessible, given suitable capacities, constraints, and (empirical) evidence

Grundtanken är alltså, om jag inte missuppfattar honom, att fakta om mening hos språkliga uttryck är epistemiskt *tillgängliga* för andra än användaren. Kvalifikationerna som hänvisas till i formuleringen motiveras så här. För det första säger offentlighetstesen inte mycket om vi inte anger vilka förmågor som krävs av dem för vilka mening är tillgängligt. För det andra måste vi ange ett antal "restriktioner" för att försäkra oss om att den teori vi kommer fram till är en teori om *mening*, och inte om något annat. För det tredje skall kunskapen om mening grundas på empirisk evidens.

Jag är inte ense med Stjernberg om behovet av dessa tillägg. Man kan visserligen behöva säga något om vilka förmågor som *inte* skall räknas in, t ex telepatisk förmåga, men offentlighetstesen blir knappast innehållslös om vi *inte* säger vilka förmågor som *krävs*. Det är stor skillnad mellan att mening är tillgänglig och att mening inte är tillgänglig, för andra, givet att dessa andra t ex inte har telepatisk förmåga.

Restriktionerna kan vi också lämna därhän, om vi inte avser att göra offentlighetstesen till del av en redogörelse för vad språklig mening är. Och även om vi gör det, så är det inte klart att några restriktioner behöver inkluderas i offentlighetsprincipen själv. Själv är jag intresserad av vilka metafysiska slutsatser, dvs om väsendet hos språklig mening, som kan dras ur kunskapsteoretiska premisser, och föredrar därför en mer renodlat kunskapsteoretisk version av offentlighet. Jag vill här skilja mellan två principer om offentlighet hos mening, en modal, som talar om vad som är möjligt att veta, och en faktuell, som talar om den kunskap som talare faktiskt har. Låt oss kalla den första *tesen om fundamental offentlighet* (FO), och den andra *tesen om normal offentlighet* (NO):

(FO) Vad en talare menar med sina ord är *möjligt* för andra att veta
(NO) Talare vet *normalt* vad andra talare menar med sina ord

Detta är intuitiva formuleringar. Mycket kan sägas om hur de kan tolkas och preciseras, och om hur de förhåller sig till varandra, men jag får lämna detta till ett annat tillfälle, dels av utrymmesskäl, dels

för att dessa enkla formuleringar duger för syftet i denna uppsats.

Jag tror att bägge principerna är viktiga för Quine, men den senare, tesen om normal offentlighet, är den viktigare, eftersom Quine argumenterar utifrån antaganden om faktisk språkinlärning. När jag i fortsättningen diskuterar Quines argument kommer jag att tala om offentlighet kort och gott.

2. *Från offentlighet till obestämdhet på Quines vis*

Ett argument från offentlighet till obestämdhet är, som jag ser det, ett argument från en kunskapsteoretisk premiss till en metafysisk slutsats. Som jag tolkar honom har Quine ett sådant argument. Följande passager från *Word and Object* (sid ix) respektive *Pursuit of Truth* (s 37–38) är skrivna med mer än trettio års mellanrum:

> Language is a social art. In acquiring it we have to depend entirely on intersubjectively available cues as to what to say and when. Hence there is no justification for collating linguistic meanings, unless in terms of men's dispositions to respond overtly to socially observable stimulations. An effect of recognizing this limitation is that the enterprise of translation is found to be involved in a certain systematic indeterminacy.

> Critics have said that the thesis [of indeterminacy of translation] is a consequence of my behaviorism. Some have said that it is a reductio ad absurdum of my behaviorism. I disagree with this second point, but I agree with the first. In psychology one may or may not be a behaviorist, but in linguistics one has no choice. Each of us learns his language by observing other people's verbal behavior and having his own faltering verbal behavior observed and reinforced or corrected by others. We depend strictly on observable behavior in observable situations. [...] There is nothing in linguistic meaning beyond what is to be gleaned from overt behavior in observable circumstances.

Strukturen i resonemanget, som jag uppfattar det, är densamma i bägge passagerna:

1. Talare lär sig språket av andra talare (OFFENTLIGHET)
2. Talares enda källa till kunskap om andra talare är observation av deras beteende i observerbara situationer (EMPIRISM)

3. Alltså finns det ingenting i mening utöver vad som är bestämt av talares observerbara beteende i observerbara situationer (SPRÅKLIG BEHAVIORISM, ur 1 och 2)
4. Alltså är riktighet hos översättningsmanualer obestämt (OBESTÄMDHET, ur 3)

Detta sätt att uppfatta Quines skäl för obestämdhetstesen förekommer såvitt jag vet inte i litteraturen. Dagfinn Føllesdal har poängterat relevansen hos offentlighet för obestämdhet, men utan att precisera offentlighetsidén och utan att presentera ett klart argument. Närmast min egen uppfattning, men ändå klart annorlunda, är Dorit Bar-On.

Som framgår har jag försett dessa steg i argumentet med beteckningar. Med avseende på dessa ska jag gå igenom teserna Jag gör det i omvänd ordning, eftersom den första är den mest problematiska i exegetiskt avseende.

Obestämdhetstesen formuleras på följande sätt i *Word and Object* (s 27) respektive *Pursuit of Truth* (s 48):

> The thesis is this: manuals for translating one language into another can be set up in divergent ways, all compatible with the totality of speech dispositions, yet incompatible with one another. In countless places they will diverge in giving, as their respective translations of a sentence of the one language, sentences of the other language which stand to each other in no plausible sort of equivalence however loose. [...] It is in this last form, as a principle of indeterminacy of translation, that I will try to make the point plausible in the course of this chapter.

> A manual of Jungle-to-English translation constitutes a recursive, or inductive, definition of a translation relation together with a claim that it correlates sentences compatibly with the behavior of all concerned. The thesis of indeterminacy of translation is that these claims on the part of two manuals might both be true and yet the two translation relations might not be usable in alternation, from sentence to sentence, without issuing in incoherent sequences. Or, to put it another way, the English sentences prescribed as translation of a given Jungle sentence by two rival manuals might not be interchangeable in English contexts.

En rimlig rekonstruktion är följande:

En översättningsmanual mellan två naturliga språk är *riktig* om den stämmer med talarnas språkliga beteende.

Två översättningsmanualer mellan samma språk är *inbördes förenliga* om de för varje sats i det ena språket ger intutitivt ekvivalenta översättningar i det andra, vilket också skall innebära att de, med bevarad diskursiv koherens, kan användas omväxlande.

Det finns, eller kan konstrueras, två eller flera översättningsmanualer mellan samma naturliga språk som alla är riktiga men som är inbördes oförenliga.

Vad gäller språklig behaviorism är sats 3 helt enkelt en formulering av denna doktrin.

Sats 2 är inte en formulering av empirismen, dvs uppfattningen att all kunskap om verkligheten grundas på sinneserfarenhet, men en konsekvens av den. Strängt taget krävs en extra premiss för att räkna bort sådant som är observerbart men som inte räknas till vare sig beteende eller omständigheter för beteende, men behovet bortfaller om vi är tillräckligt flexibla ifråga om vad som skall räknas till omständigheter.

Att ange sats 1 som en premiss som innefattar offentlighet, i den här definierade bemärkelsen, är emellertid inte uppenbart riktigt. Anser Quine verkligen att språkinlärning är en process i vilken en talare tillägnar sig kunskap om vad andra talares ord betyder? Quine skriver om språkinlärning och språkutveckling på många ställen, och ofta tycks språkinlärning inte vara någonting utöver utvecklandet av egna dispositioner till språkligt beteende. I vissa passager inkluderas emellertid kognitiva komponenter, som i 'Ontological relativity' (s 28):

> The semantic part of learning a word is more complex than the phonetic part, therefore, even in simple cases: we have to see what is stimulating the other speaker.

Här är det uppenbart att någon slag kunskap om andra talare är inblandad i inlärningsprocessen.

Viktigare än sådana passager är att ett argument från språkinlärning och empirism till språklig behaviorism skulle vara meningslöst om språkinlärningen inte innefattade tillägnandet av kunskap om vad andra menar. Om språkinlärning inte bestod i något *annat* än att

tillägna sig språkliga dispositioner, och därmed språkkunskap bestod i att ha språkliga dispositioner, så skulle vi ha antagit språklig behaviorism redan i beskrivningen av språkinlärning, och argumentet vore då cirkulärt. Och om språkinlärning bestod i att utveckla språkliga dispositioner, och utöver detta i att utveckla något mer fast inte kunskap om vad andra menar så skulle språklig behaviorism uppenbart inte följa. Det vore då fullt godtagbart att meningen hos dina ord bestäms av dina privata, för mig otillgängliga, mentala egenheter, och att min språkinlärning bestod i att, på grundval av observation av ditt språkliga beteende, utveckla egna språkliga dispositioner och dessutom i mitt inre associera mina ord med mina privata meningar. Vi skulle inte ha någon restriktion som uteslöt detta. Endast om vi lägger till att språkinlärning innefattar tillägnandet av kunskap om vad andra talare menar erhåller vi en restriktion: det faktum att vi får kunskap om vad andra talare menar, tillsammans med empirismens restriktioner för hur kunskap kan erhållas, och därmed vilken kunskap som kan erhållas, ger oss restriktioner för vad mening kan vara. Endast med denna tolkning av vad Quine menar med språkinlärning har vi ett intressant argument överhuvud taget, och därför väljer jag den.

Det andra bekymret med denna tolkning, eftersom det är Quine den tillskrivs, är att termen "mening" används för att formulera den. Men såvitt jag kan se är detta tillåtet. För att förstå Quines uppfattning om mening måste vi skilja mellan två bemärkelser av "mening", objektiv, eller objektivt grundad, mening, å ena sidan, och intuitiv mening, å den andra. När Quine säger sådant som att "there is nothing in linguistic meaning beyond what is to be gleaned from overt behavior in observable circumstances" används "mening" i den objektiva bemärkelsen. Det är också mening i den objektiva bemärkelsen som bevaras av sådana översättningsmanualer som uppfyller villkoren att passa talarnas språkliga dispositioner.

Mening i den intuitiva bemärkelsen, å andra sidan, är mening i enlighet med våra intuitiva omdömen om likhet och skillnad i mening, dvs synonymi och icke-synonymi. Den intuitiva bemärkelsen behövs för att formulera obestämdhetstesen. I Quines formuleringar ovan förekommer fraserna "stand to each other in no plausible sort of equivalence however loose" respektive " not be usable in alternation, from sentence to sentence, without issuing in incoherent sequences". I bägge fallen är det intuitiva omdömen om ekvivalens mellan satser, respektive koherens-

bevarande substitutioner av satser för satser i diskurser, som åberopas för att formulera obestämdhetstesen. Dessa intuitiva omdömen är intuitiva omdömen om likhet och skillnad i mening. Och om vi väl har tillgång till bägge begreppen, om objektiv och om intuitiv mening, så skulle vi kunna formulera obestämdhetstesen genom att säga att objektiv mening underdeterminerar intuitiv mening: vi gör finare diskrimineringar ifråga om intuitiv mening än vad som kan grundas på skillnader i objektiv mening.

3. Strategin mot Quines argument

Quines argument har alltså två premisser, offentlighet och empirism. Om vi tar empirismens riktighet som bakgrund, så har vi ett argument från offentlighet till obestämdhet, via språklig behaviorism. Argumentet har alltså två huvudsteg, ett från offentlighet till språklig behaviorism, och ett från språklig behaviorism till obestämdhet:

Offentlighet ⇒ Språklig Behaviorism
Språklig Behaviorism ⇒ Obestämdhet

Detta är argumentet jag ska undersöka. Jag skall betrakta argumentet som ett mycket starkt argument, intuitivt med anspråket att slutsatsen *måste* vara sann, under antagande att premisserna är sanna. Quine har aldrig presenterat sina övervägande på detta sätt, och anser obestämdhetstesen som en obevisad hypotes (därmemot tycks han anse sanningen hos sin språkliga behaviorism som etablerad). Att angripa ett starkare argument är lättare, men i detta fall också av principiellt intresse, eftersom jag, om jag lyckas, visar att Quines övervägande i princip inte kan förstärkas till ett konklusivt argument.

Problemet med Quines argument i två steg är följande. Det finns olika former av språklig behaviorism, vissa starkare och andra svagare. Formuleringen ovan av språklig behaviorism täcker alla. Jag ska försöka visa att det visserligen går att argumentera för språklig behaviorism utifrån offentlighet, men att men därvid endast kan sluta sig till en *svag* form av språklig behaviorism. Jag ska också försöka visa att det går att argumentera från språklig behaviorism till obestämdhet, men att det därvid krävs en *stark* form av språklig behaviorism. Jag ska försöka visa att den svaga formen, som är en konsekvens av offentlighet, är *svagare* än den form som krävs för att visa riktigheten hos obestämhetstesen. Schematiskt har vi alltså följande situation:

Offentlighet ⇒ Svag Språklig Behaviorism
Stark Språklig Behaviorism ⇒ Obestämdhet
Icke: Svag Språklig Behaviorism ⇒ Stark Språklig Behaviorism

Jag ska försöka visa att Stark Språklig Behaviorism kan vara falsk även om Offentlighet, och därmed Svag Språklig Behaviorism, är sann. Detta är inte att visa att falskheten hos Obestämdhet också är förenlig med sanningen hos Offentlighet, men det visar omöjligheten av att ge ett konklusivt argument från offentlighetstesen till obestämdhetstesen *via* språklig behaviorism.

4. Från offentlighet till språklig behaviorism

Låt oss skriva om argumentets första steg, genom att förena offentlighet och empirism till en premiss, och göra formuleringarna lika. Då erhåller vi följande lilla argument:

1. Vad en talare menar med sina ord är känt av andra på grundval av observation av hans beteende i observerbara situationer.
2. Alltså gäller att vad en talare menar med sina ord är bestämt hans observerbara beteende i observerbara situationer.

Såvitt jag förstår är detta ett giltigt argument. Det är en variant av följande form:

> X vet att p. Ur p sluter sig X till q, och erhåller därmed kunskap att q. Alltså är q en konsekvens av p (p en grund för q).

Denna form håller jag för giltig, eftersom X:s trosuppfattning att q inte korrekt skulle klassificeras som kunskap, givet X:s skäl för den, om inte q vore en konsekvens av p.

Alltså finns ett giltigt argument från offentlighet, och empirism, till språklig behaviorism. Men det finns versioner av språklig behaviorism, vilka innefattar starkare teser om relationen mellan språkligt beteende och mening, vilka inte följer ur dessa premisser. En tes som inte följer ur premiss 1 är att det finns ett konstant samband mellan språkliga dispositioner och mening, invariant mellan talare, samt, vad gäller en och samma talare, över tid.

Följande är nämligen möjligt, eller åtminstone förenligt med premiss 1. Kalle och Eva har precis samma språkliga dispositioner, dvs dispositioner till språkligt beteende, eller åtminstone så lika som det alls är möjligt för två personer att ha. Lars har fullständig kunskap, grundad

på observation av deras respektive beteende, om bägges språkliga dispositioner, eller åtminstone så god kunskap som det är möjligt att ha, på grundval av observation. Likafullt tolkar han Kalle och Eva olika. Han tolkar bägge på grundval av observation av deras beteende, och deras beteende är lika, men Lars tolkar samma satser på olika sätt, beroende på om de är yttrade av Kalle eller av Eva. Han anser dessutom att det sätt på vilket han tolkar Kalle inte korrekt kan tillämpas på Eva, utan i hennes fall skulle ge felaktigt resultat, och omvänt. Dessutom har Lars rätt. Hans tolkningar är riktiga, och han vet vad Kalle och Eva menar med sina ord.

Hur kan detta vara möjligt? Varför skulle Lars alls tolka Kalle och Eva olika? Att han tolkar dem olika beror på att han är disponerad att tolka dem olika. Hans dispositioner för att tolka Kalle är annorlunda än hans dispositioner att tolka Eva. Denna skillnad i dispositioner kan vara medfödd. Den kan också vara förvärvad, beroende på t ex kausal inverkan av faktorer i bakgrunden, eller på kausal inverkan direkt från Kalle respektive Eva, av ett slag som Lars inte med nödvändighet är medveten om. Vi kan tänka oss många möjligheter. Det väsentliga är vi kan vara berättigade att tillskriva Lars kunskap om mening på grundval av dessa dispositioner. Det kan vara så att Lars är en mycket tillförlitlig tolkare: närhelst, eller näranog närhelst, han har fasta övertygelser om vad någon menar med sina ord, så har Lars rätt. Det kan också vara så de uppfattningar om vad Kalle och Eva menar som Lars utvecklat, har genererats av en mycket tillförlitlig process: närhelst, eller näranog närhelst, en övertygelse om vad någon menar med sina ord, har genererats av denna process, så är övertygelsen riktig. I kraft av sådan *tillförlitlighet* ifråga om tolkning kan Lars tillskrivas kunskap om vad Kalle respektive Eva menar.

Jag har alltså här åberopat uppfattningen om kunskap i kraft av tillförlitlighet (reliabilitet). Detsamma har Edward Craig tidigare gjort i liknande sammanhang (1982 s 554–55, 1986 s 174–76). Om det är berättigat att göra detta, så kan vi inte ur premiss 1, och därmed inte heller ur den svaga språkliga behaviorism som är en konsekvens av premiss 1, sluta oss till att det måste råda ett invariant samband mellan språkliga dispositioner och mening. Är det berättigat att här åberopa tillförlitlighet? Det kan i varje fall inte här avvisas utifrån en uppfattning om ett starkare samband mellan beteende och mening, eftersom det var ett sådant samband som skulle visas. Om det är

berättigat måste istället bero på vad som kan anses som ett adekvat begrepp om kunskap vad gäller språklig mening. Jag återvänder till en diskussion av detta i avsnitt 6.

En annan fråga är om motexemplet med Lars, Kalle och Eva är rimligt. I första hand vill jag säga att jag inte är ute efter att bedöma rimligheten hos det, utan att framhålla dess förenlighet med premiss 1. I andra hand skulle jag emellertid vilja säga att det verkligt orimliga i sammanhanget är antagandet om en fast, och kanske skarp, avgränsning mellan icke-semantiska fakta som är, respektive inte är, relevanta för språklig mening. Själv är jag ense med Quine att *om* det finns någon skillnad i mening mellan en sats som yttrad av Kalle och samma sats som yttrad av Eva, så måste det också finnas en naturlig, eller åtminstone icke-semantisk, skillnad mellan dem som ligger till grund för denna semantiska skillnad (någon form av superveniens). Jag är inte ense med Quine om möjligheten att skilja ut en totalitet av sådana relevanta icke-semantiska grunder, vilka dessutom skall vara relaterade till mening på ett konstant sätt. Min poäng är att man *inte* kan hänvisa till möjligheten för en talare att ha kunskap om mening hos andra talares ord, som en grund för att skilja ut en sådan totalitet av relevanta fakta. Skälet är att de fakta som räknas som irrelevanta ändå kan påverka talaren och få honom att dra *andra* slutsatser än annars ur de fakta som räknas som relevanta.

Det är fortfarande så att mening hos en talares ord är en funktion av hans språkliga dispositioner. Men detta är förenligt med att det är olika funktioner för olika talare, eller för samma talare vid olika tidpunkter, beroende på skillnader ifråga om *andra* faktorer än språkliga dispositioner.

För att kopplingen till obestämdhetstesen tydlig ger jag nu en definition av vad jag kallar Meningsbestämmande Behavioristisk Egenskap, MBB-egenskap, som jag ska återknyta till i nästa avsnitt:

> Egenskapen F är en *Meningsbestämmande Behavioristisk Egenskap* omm det dels gäller att huruvida ett uttryck a som använt av en talare X vid en tidpunkt t har F består i X:s (behavioristiskt specificerade) dispositioner till användning av a vid t, och dels gäller att om ett uttryck a, använt av en talare X vid tidpunkt t, har F, och ett uttryck b, använt av talare Y vid tidpunkt t', har F, så har a och b samma objektivt bestämda mening, och kan korrekt översättas till varandra enligt objektivt giltiga kriterier.

Vad som skall observeras är att existensen av MBB-egenskaper *inte* följer ur premiss 1, och därmed inte ur Svag Språklig Behaviorism. För om det hade funnits MBB-egenskaper, så skulle det också finnas ett invariant samband mellan språkliga dispositioner och mening, och vi har sett att detta inte följer ur premiss 1. Därför gäller också att *ingen* form av språklig behaviorism som implicerar existensen av MBB-egenskaper är en konsekvens av premiss 1. Det är detta jag skall använda i nästa avsnitt.

5. *Från språklig behaviorism till obestämdhet*

För ett direkt och konklusivt bevis för obestämdhet hos översättning krävs tre saker: för det första en uppsättning villkor för riktighet hos översättningsmanualer, för det andra ett bevis för att dessa villkor tillsammans *underbestämmer* riktighet hos översättningar, dvs ett bevis för att flera inbördes oförenliga översättningsmanualer kan uppfylla villkoren, och, för det tredje, ett bevis för att dessa villkor är adekvata och att uppsättningen är *fullständig*, dvs innehåller alla adekvata villkor.

Hur plausibel man finner obestämdhetstesen beror alltså på hur rimliga man finner föreslagna villkor för riktighet hos översättning, samt hur rimligt det verkar att dessa villkor underbestämmer riktighet. I *Pursuit of Truth* (s 43) framställer Quine villkoren för riktighet på följande sätt:

> A pioneer manual of translation has its utility as an aid to negotiation with the native community. Success in communication is judged by smoothness of conversation, by frequent predictability of verbal and non-verbal reactions, and by coherence and plausibility of native testimony. It is a matter of better or worse manuals rather than flatly right or wrong ones.

Såvitt jag förstår innebär detta att kriteriet för att en manual är riktig är att det förefaller användaren av den att han kommunicerar framgångsrikt med dess hjälp. Konversationen flyter, och användaren av manualen får i hög grad sina förväntningar om de främmande talarnas reaktioner bekräftade. Oavsett hur rimliga dessa villkor kan tyckas vara, är problemet i detta sammanhang att de inte ger något stöd för tesen om obestämdhet. Alla möjliga slags faktorer, som frekvensen av konsonanter i översättningarna, eller auktoriteten hos den som före-

slagit manualen, skulle kunna spela en roll för att göra en konversation mer lättflytande, eller för att styra användarens förväntningar på sin samtalspartner. Om vi inte drar någon mer principiell gräns mellan vad som är relevant och vad som inte är relevant för manualens riktighet, så har vi inte något gott skäl att tro att två eller flera manualer skulle kunna vara precis lika bra: vilken till synes irrelevant detalj som helst skulle få en användare att föredra den ena, och om denna detalj bidrar till konversationens lätthet, så bidrar den också, enligt detta förslag, till manualens riktighet. De kriterier Quine föreslår i *Pursuit of Truth* ger inte något stöd för obestämdhetstesen.

Det förhåller sig på annat sätt med de kriterier Quine föreslog i *Word and Object*. Dessa grundade sig på begreppet stimulusmening. Quine definierade stimulusmeningen för en sats, och en talare, som ett ordnat par av positiv respektive negativ stimulusmening, där den positiva stimulusmeningen är mängden av alla de stimuleringar (retningar av sinnesceller) som, inom ett bestämt tidsintervall, gör talaren disponerad att bejaka satsen, och den negativa stimulusmeningen är motsvarande mängd knuten till dispositionen att förneka satsen (s 32–33).

I termer av stimulusmening definierade Quine sedan begreppet observationssats, stimulussynonymi (likhet i stimulusmening), samt stimulusanalyticitet (egenskapen att bejakas oavsett stimulering) (s 42, 46, 55). Quine förklarade också hur logiska konnektiv i det främmande språket skall identifieras i termer av stimulusmening. Till exempel kännetecknas negationen av en sats att dess positiva stimulusmening är den negerade satsens negativa stimulusmening, och vice versa (s 57).

Ovanpå detta införde Quine begreppet analytisk hypotes. En analytisk hypotes om ett främmande språk är en hypotes om vilka ord dess satser innehåller, och hur dessa skall paras samman med ord i det egna språket (s 68). Ur de analytiska hypoteserna kan man sedan härleda översättningar av hela satser.

I dessa termer kunde Quine sedan formulera kriterier på en godtagbar, eller korrekt, översättningsmanual. Den skall innehålla analytiska hypoteser som inte är alltför komplicerade, och vara sådana att de

1. parar samman observationssatser (från respektive språk) med samma stimulusmening

2. översätter logiska konnektiv i enlighet med deras bidrag till stimulusmening
3. parar samman stimulusanalytiska med stimulusanalytiska satser, liksom stimuluskontradiktoriska med stimuluskontradiktoriska, så långt möjligt
4. parar samman stimulussynonyma satspar (i ena språket) med stimulussynonyma satspar (i det andra) (s 68–69).

Quine presenterar inte detta som en fullständig lista över villkor för korrekthet hos manualer, och tycks dessutom tänka sig villkoren som ideal snarare än strikta villkor. Men låt oss anta dessa fyra villkor faktiskt är strikta och listan fullständig: en manual som uppfyller dem är korrekt, en som inte uppfyller dem inkorrekt. Under detta antagande är det lätt att visa obestämdhetstesens riktighet.

Vi kan utan problem anta att det finns en övre gräns för längden på satser som talare överhuvud taget har dispositioner till. Låt oss symboliskt anta att denna längd är 3 000 ord. Och låt oss anta att vi har en översättningsmanual G, i formen av en funktion från satser i det ena språket till satser i det andra, som uppfyller villkoren 1 till 4. Utifrån G definierar vi nu den alternativa manualen H:

H(s) = G(s) för satser upp till 3000 ord
H(s) = "det är icke fallet att"^G(s) för satser på mer än 3000 ord

Det är klart även H uppfyller villkoren, eftersom H sammanfaller med G för alla satser för vilka talarna har några dispositioner alls. Det är också klart att H är oförenlig med G eftersom den, för oändligt många satser, som översättning ger negationen av den översättning som G ger. Det är dessutom klart att H kan specificeras rekursivt eller kompositionellt (översättningar av sammansatta uttryck definieras i termer av översättningen av deras delar) om G kan: det är bara att lägga till definitionen av H till den rekursiva definitionen av G.

Vi har därmed bevisat att dessa villkor underbestämmer korrekthet hos översättningar: flera oförenliga manualer kan uppfylla villkoren (om någon kan, fast vi vet att identitetsmanualen, från t ex svenska till svenska (G(s) = s för varje sats s), uppfyller dem). Och givet att vi har en uppsättning villkor som antas fullständig är det rimligt att tro att något trick som detta kan användas för att bevisa underbestämning.

Det finns ingen garanti att varje uppsättning behavioristiska villkor underbestämmer manualers riktighet. Men omvänt gäller att vi *inte* kan

visa obestämdhetstesens riktighet, ens genom verkliga exempel, med mindre än en fullständig uppsättning villkor. Och det är inte svårt att visa att det är *omöjligt*, utifrån premisserna om offentlighet hos språklig mening och empirism, att sluta sig till riktigheten och fullständigheten hos en uppsättning behavioristiska villkor. För antag att vi har en sådan uppsättning villkor. Antag, för enkelhets skull, att våra villkor 1–4 är adekvata och tillsammans fullständiga. I så fall kan vi enkelt definiera ekvivalensrelationen *likhet ifråga om meningsbestämmande behavioristiska egenskaper*:

> a och b har samma meningsbestämmande behavioristiska egenskap omm a kan översättas med b enligt en manual som uppfyller villkoren 1–4.

Detta är en acceptabel definition, eftersom villkoren är formulerade i behavioristiska termer. Och eftersom det finns en väldefinierad relation om likhet ifråga om MBB-egenskaper, så finns det också MBB-egenskaper.

Lägger vi samman resultaten från detta och föregående avsnitt får vi följande:

i) obestämdhetstesen kan visas följa ur språklig behaviorism endast om denna är så stark att den innefattar existensen av en fullständig uppsättning adekvata behavioristiska villkor för riktighet hos översättningsmanualer
ii) existensen av MBB-egenskaper följer ur existensen av en fullständig uppsättning adekvata behavioristiska villkor för riktighet hos översättningsmanualer
iii) existensen av MBB-egenskaper följer inte ur premisserna offentlighet och empirism
iv) alltså följer ingen form av språklig behaviorism, vilken är så stark att obestämdhetstesen kan visas utifrån den, ur offentlighet och empirism.

Med andra ord: inget framgångsrikt argument kan ges, utifrån offentlighet och empirism, via språklig behaviorism, för obestämdhetstesen.

6. Kunskap om mening

Slutsatsen ovan står sig under förutsättning att det var berättigat att åberopa uppfattningen om kunskap som tillförlitligt sann tro i avsnitt 4. Denna fråga är nära förknippad med frågan om att skilja mellan faktorer som är relevanta respektive inte relevanta för riktighet hos

översättningsmanualer. Det är en fråga som Quine diskuterat tidigare. I *Word and Object* (s 74) ger Quine en lista över orsaker till varför lingvister undgår att lägga märke till obestämdhet:

> [A fifth cause is] that linguists adhere to implicit supplementary canons that help to limit their choice of analytical hypotheses. For example, if a question were to arise over equating a short native locution to 'rabbit' and a long one to 'rabbit part' or vice versa, they would favor the former course, arguing that the more conspicuously segregated wholes are likelier to bear the simpler terms. Such an implicit canon is all very well, unless mistaken for a substantive law of speech behavior.

Vissa dispositioner, eller implicita hjälphypoteser, hos lingvister och andra, kan alltså räknas som goda praktiska hjälpmedel, men är utan teoretisk betydelse. Varför? Såvitt jag har förstått är skälet till detta just språklig behaviorism, såsom motiverad av överväganden om offentlighet och empirism. Stimulusmeningen hos ett språkligt uttryck är behavioristiskt relevant, men längden på uttrycket är det inte. Problemet är att det inte är klart vilket stöd denna uppfattning kan erhålla just av empirismen och offentlighetsprincipen.

Men en tänkbar invändning mot denna kritik av Quine är just att den innefattar en alltför lätttvindig syn på *kunskap*. För att ha kunskap räcker det inte med att vara disponerad att välja rätt. Man måste också ha *evidens*, och när det kommer till att anföra empirisk evidens för vårt val av manual, så skulle vi finna att de konkurrerande manualerna har lika gott stöd i den evidens som finns. Detta är den relevanta omständigheten. Våra dispositioner som talare att välja den ena av manualerna är utan teoretisk relevans.

Mitt svar är detta: Vi är helt enkelt inte i besittning av oberoende kriterier för vad som skall räknas som evidens för manualers riktighet. Enligt vilka kriterier skulle t ex längden på språkliga uttryck inte kunna räknas till evidensen för respektive emot olika översättningsmanualer? När det gäller att komma fram till kriterier kan vi *utgå* ifrån att vi som talare normalt vet vad andra talare säger. Det som är evidens för normala talare som erhåller kunskap om vad andra talare menar, är precis det som skall *räknas* som evidens för riktigheten hos tolkning och översättning. Om talare faktiskt använder satslängd som evidens, så har vi därmed inga skäl att förkasta det.

Invändningen kan emellertid drivas längre. Kunskap, kan man säga, hör samman med möjligheten till *rättfärdigande*. På grund av detta kan försvararen av obestämdhetstesen hävda att vi visst är i besittning av åtminstone vissa kriterier för vad som kan räknas som evidens för manualers riktighet. I den mån vi har kunskap om vad andra säger, på grundval av observation, så måste det vara fråga om kunskap som grundas just precis på konstanta och specificerbara samband. I annat fall skulle vi hoppa från att åberopa vissa saker som evidens i vissa fall, och andra saker i andra fall. Detta vore *ad hoc*, och skulle stå i strid mot kravet på rättfärdigande. Relationen mellan språkligt beteende och mening måste alltså vara invariant mellan talare och över tid. Det måste alltså *finnas* MBB-egenskaper.

Svaret på detta är att vi inte har några *a priori* skäl för kravet att kunskap i detta fall skall kunna ges fullt rättfärdigande. Det är inte i förväg givet att jag själv ska ha insikter i den kognitiva process som leder mig till att tolka andra talare som jag gör. Och empiriskt är det klart att vi mycket sällan är kapabla att rättfärdiga våra tolkningar. Jag tolkar andra som det förefaller mig naturligt, utan att kunna ge något annat rättfärdigande än just att det förefaller mig naturligt. Och om vi faktiskt är ytterst tillförlitliga som tolkare av varandra, och näranog genomgående tolkar varandra riktigt, vore det mycket egendomligt att säga att vi ändå inte vet vad andra egentligen säger så länge vi inte kan ge explicita rättfärdiganden. Kravet på rättfärdigande tjänar mer till att undergräva tesen att vi vet vad andra menar, dvs tesen om normal offentlighet, än att undergräva det berättigade i att åberopa tillförlitlighetsuppfattningen om kunskap vad gäller mening.

Det är alltså rimligt att åberopa denna uppfattning här, och därmed står sig min slutsats från avsnitt 5. Detta är emellertid fullt förenligt med att Quine har rätt i att obestämdhet är en konsekvens av offentlighet, och empirism, och att detta kanske även kan visas genom något annat slags argument än det jag undersökt här.

Not

Jag har många att tacka för kommentarer som givit upphov till förbättringar. Uppsatsen gavs som föredrag i Umeå, och senare även vid konferensen "Questions from Quine", Karlovy Vary, september 1995. Mest hjälp har jag fått av Lars Bergström, som hörde mig båda gångerna och dessutom kommenterat ett tidigare utkast. Andra vars kommentarer varit av värde inkluderar Roger Gibson, Sten Lindström, Dag Prawitz, W V Quine, Wlodek Rabinowicz och Fredrik Stjernberg.

Litteratur

BAR-ON, DORIT, 'Semantic Verificationism, linguistic behaviorism, and translation', *Philosophical Studies* 66, 1992, sid 235–59

CRAIG, EDWARD, 'Meaning, use and privacy', *Mind* 91, 1982, sid 541–564

CRAIG, EDWARD, 'Privacy and rule-following', i Jeremy Butterfield (red), *Language, Mind and Logic*, Cambridge University Press, Cambridge 1986

FØLLESDAL, DAGFINN, 'Indeterminacy and mental states', i Robert Barrett och Roger Gibson (red), *Perspectives on Quine*, Basil Blackwell, Oxford 1990

FØLLESDAL, DAGFINN, 'In what sense is language public', i Paolo Leonardi och Marco Santambrogio (red), *On Quine*, Cambridge University Press, Cambridge 1995

QUINE, W. V. O., *Word and Object,* MIT Press, Cambridge, Mass. 1960

QUINE, W. V. O.,*Ontological Relativity and Other Essays*, Columbia University Press, New York 1969

QUINE, W. V. O., *Pursuit of Truth*, 2a uppl., Harvard University Press, Cambridge, Mass. 1992

STJERNBERG, FREDRIK, The Public Nature of Meaning, Almqvist & Wiksell, Stockholm 1991.

INGAR BRINCK

Indexikal kunskap och social mening: om ordet 'jag'

Finns det någon gåta bakom ordet 'jag'? Det är förmodligen den fråga som ytterst driver många personer som sysslar med detta ords semantik. Man hoppas att analysen av 'jag' ska bereda vägen för en metafysisk slutsats om jagets natur. Men i stället för att direkt ställa denna svåra fråga brukar man börja med en annan, som åtminstone vid första anblicken tycks betydligt enklare att besvara: Hur fungerar ordet 'jag' i vanligt språkbruk? Jo, det pekar ut talaren, den som yttrade ordet (utan att några andra egenskaper förutsätts hos denne). Men det gäller att även förklara *hur* ordet pekar ut talaren. Det finns en mängd olika teorier som säger hur det bör göras. Man kan dela in dem i tre huvudgrupper.

Den *första* teorin säger att ordet 'jag' *indikerar* talaren. Härmed avses att 'jag' riktar åhörarnas uppmärksamhet mot talaren på ungefär samma sätt som ett ångestfyllt skrik låter åhörarna förstå att någon är i nöd, och att denna varelse befinner sig på den plats skriket kom ifrån. 'Jag' lokaliserar talaren för åhörarna, men ger ingen ytterligare information om den som yttrat ordet. Man kan alltså inte sluta sig till något beträffande talarens natur utifrån ordet 'jag'. Talaren själv använder 'jag' som ett rent expressivt uttryck, ett utrop utan kognitivt innehåll.

Den *andra* teorin hävdar att ordet 'jag' *refererar* till, eller står för, talaren sedd som en bestämd fysisk entitet. Det finns en direkt förbindelse mellan talaren och 'jag' som gör att detta ord kan fungera som en ersättning eller 'stand-in' för den fysiske talaren i språket. Förbindelsen, som alltså bestämmer ordets referens, kan vara kausal eller omedelbar och kognitiv eller en kombination av dessa två. Att den är omedelbar och kognitiv innebär att talaren har en insikt om vad 'jag' står för genom sinnena och att denna insikt varken är förmedlad av andra sinnesintryck eller av propositionell kunskap.

Den *tredje* teorin menar också att 'jag' refererar till talaren, men påstår att *referensrelationen inte* kan vara *direkt*. Det måste finnas någon sorts beskrivning av vad en talare är för något som styr ordet 'jag's funktion. Vilken typ av beskrivning som behövs är omtvistat. Ibland hävdas det också att referensen hos enskilda yttranden av 'jag' bestäms separat. I sådana fall behövs en beskrivning av varje enskild talare som yttrar 'jag'.

När man ska bedöma vilken av dessa huvudteorier som kan vara riktig är det en bra strategi att utgå från exempelsatser. Man ser på olika satser och försöker utröna vilka villkor som måste vara uppfyllda för att ordet 'jag' ska peka ut talaren i satserna. Bakom strategin ligger ytterst målet att finna sanningsvillkoren för satsen "Ordet 'jag' pekar ut talaren'. Det gäller också denna artikel. I stället för att var och en för sig diskutera de tre ovan nämnda teorierna om hur 'jag' fungerar, kommer jag att gå igenom exempelsatser direkt. Utifrån resultatet av denna diskussion hoppas jag kunna sluta mig till vilken av teorierna som är riktig. Innan vi börjar med detta vill jag emellertid nämna en möjlig skillnad i perspektiv mellan hur talare och åhörare använder och förstår 'jag'.

Den lingvistiska regeln att 'jag' pekar ut talaren gäller, med vissa kvalifikationer, i alla sammanhang. Men tillämpningen av denna regel tycks skilja sig åt beroende på om man fäster avseende vid talarens eller åhörarens förståelse av ordet 'jag'. Det som talaren behöver veta för att förstå 'jag's användning tycks vara annorlunda än det som åhöraren måste känna till. Till exempel verkar det som om talaren måste veta vem hon är i någon djupare bemärkelse för att kunna identifiera sig själv som 'jag' i vissa kontexter, medan det i många fall räcker att åhöraren känner till den grundläggande språkregeln att ordet helt enkelt pekar ut talaren, vem denne än är, för att identifiera talaren som kontextens subjekt. Detta tyder på att ordet 'jag' har en speciell kognitiv roll för talaren. Vilken denna roll skulle kunna vara förklaras bäst genom exempel.

Först ett exempel som i sin ursprungliga form emanerar från John Perry i uppsatsen "The Problem of the Essential Indexical". Jag lägger till några detaljer. En man, säg Smith, är och handlar i en livsmedelsaffär. Han upptäcker en lång sträng strösocker på golvet och sluter sig till att någon i affären spiller socker på golvet. Så får han i en av speglarna intill grönsaksdisken syn på en man som har en sönderriven

sockerförpackning i sin vagn. Smith tänker att det är denne man, iklädd brun kavaj och gul hatt, som spiller sockret. Men det är inte förrän han inser att det är sig själv han ser i spegeln, som han vidtar åtgärder för att hindra sockret att rinna ut. Tanken 'Det är *jag* som spiller sockret' får Smith att handla på ett riktigt sätt. Exemplet visar att endast indexikaler är handlingsmotiverande. Dessutom blir det tydligt att *om* referensen hos 'jag' bestäms genom en kausal länk, så är det inte genom den typ av länk som Smith har till sig själv genom att se sig i en spegel. För att kunna identifiera sig själv som mannen i spegeln krävs något annat. Faktum är att Smith förmodligen skulle kunna tala om sig själv i första person singularis utan att ha någon deskriptiv kunskap om sig själv alls. Att sådan kunskap, som beskriver talaren utifrån ett tredje persons perspektiv, inte är nödvändig för självreferens kan belysas med ett annat av Perrys exempel, i uppsatsen "Frege on Demonstratives".

En man vid namn Lingens är instängd i ett bibliotek efter stängningsdags och kan inte hitta ut. Lingens har drabbats av total minnesförlust. Han vet inte vem han är eller var han är. Han kan inte ringa till någon och fråga vem han är, och kan inte heller be någon att hämta honom eftersom han inte vet var han befinner sig. Han sätter sig ned vid ett bord för att fundera på vad han ska göra. På bordet ligger en bok och en karta. Han börjar läsa boken och vet snart allt om en spion vid namn Lingens. Han studerar också kartan som är över biblioteket i Stanford. Den nya faktakunskap han förvärvar hjälper honom dock inte i hans prekära situation, trots att just han är Lingens, spionen, och råkar befinna sig i Stanfordbiblioteket. Han vet inte att *han är Lingens i Stanford.* Han saknar kunskap för att kunna identifiera beskrivningarna av Lingens och Stanfordbiblioteket med den egna kontexten. Däremot vet att han att det är *han* som är vilse; han kan tala om sig själv som 'jag' trots att han inte kan knyta beskrivningarna av Lingens till sig själv.

Fakta som inte är bundna till någon bestämd kontext är alltså inte tillräckliga för självidentifikation – något annat måste till: indexikal kunskap. I princip kan vem som helst som känner till språkregeln att 'jag' pekar ut talaren använda ordet för sig själv, men vissa ytterligare villkor måste vara uppfyllda för att man faktiskt ska kunna tillämpa denna regel. Frågan är om dessa villkor tillhör semantiken hos 'jag' eller om de är förberedande eller pre-semantiska. Hur man ställer sig

till denna fråga beror på hur man ser på semantiken. Detta är ju delvis en definitionssak. Semantik är läran om teckens betydelse. Men vad är betydelse? För min del anser jag att en teori med målet att redogöra för hur ordet 'jag' fungerar bör försöka förklara ordets kognitiva roll, det vill säga, den roll som insikten om att man är ett 'jag' spelar för förmågan att kunna använda ordet vid lämpliga tillfällen.

Låt oss nu titta på hur ordet 'jag' fungerar i några exempelsatser.

(1) Jag fryser.
(2) Jag har fått jobbet som redaktör.
(3) Jag är Naima Wifstrand.

Vad måste man känna till om sig själv för att kunna yttra satserna? I det första fallet behöver ingen deskriptiv kunskap vara inblandad. Det är tillräckligt att man har upplevelsen av att frysa. Det vore absurt att tänka sig att en person fryser och samtidigt tvivlar på *vem* det är som fryser. (Däremot kan man tvivla på att man faktiskt fryser. Kanske inbillar man sig bara?) Själva upplevelsen ger direkt upphov till medvetandet om att det är man själv som har den. I de andra två satserna är situationen annorlunda. I dem krävs det att talaren kan identifiera sig själv som något annat eller någon annan. Detta innebär att det i första satsen krävs en *direkt* självidentifikation, i de två andra en *indirekt*. Men i en identifikation måste det väl finnas två entiteter som är identiska – vilka entiteter skulle det vara i dessa fall?

Generellt sedd innebär självidentifikation att man blir varse att man, så att säga, är den man är. Man blir medveten om sig själv som en enskild varelse med förmågan att påverka omgivningen genom sina handlingar. Direkt självidentifikation resulterar i en insikt om att man har en viss sorts egenskap, en egenskap som är sådan att om den verkligen är instantierad, så är det man själv som har den. Man kan inte vara *direkt* medveten om denna sorts egenskap hos andra varelser, däremot på ett indirekt sätt, genom observation eller genom att man får det berättat för sig. Den typ av egenskap det gäller är upplevelser av glädje, värme, smärta, tryck mot huden, *et cetera*. Förekomsten av denna typ av självidentifikation togs upp av Sidney Shoemaker i artikeln "Self-Reference and Self-Awareness". Därefter har bland andra Gareth Evans diskuterat ämnet i *Varieties of Reference*. Indexikal kunskap, som jag använder uttrycket, består i sådan självidentifikation.

Indirekt självidentifikation, å andra sidan, kräver att man har någon

sorts medvetande eller föreställning om sig själv från början och därefter uttryckligen identifierar denna första beskrivning med en andra beskrivning, som exempelvis i satserna (2) och (3). Satser av denna typ tycks kräva att användaren ser sig själv som en oberoende varelse som kan instantiera en mängd olika egenskaper av både fysisk och psykisk natur. Ett sådant självmedvetande inbegriper alltså att man kan vara medveten om sig själv *som* instantierande en viss egenskap – inte bara som i (1) att man är medveten om att man instantierar en egenskap. (2) och (3) placerar också referenten i ett bestämt socialt rum, där det åtminstone pragmatiskt impliceras eller förutsätts att referenten har vissa egenskaper, sådana man brukar tillskriva personer: självmedvetande, uppfattning av tid och rum, begrepp om förflutet och framtid, förmåga att se på andra som på sig själv, ett etiskt medvetande, och så vidare.

Som jag ser det är det inte nödvändigt att man kan reflektera över sig själv som en medveten eller tänkande varelse för att man ska ha ett jagbegrepp. Man kan vara medveten om sig själv som en levande varelse, någon som har upplevelser, utan att veta vilken sorts varelse man är. Självmedvetande kräver inte begreppslig kunskap om vem man är eller explicita metarepresentationer. Det räcker att en varelse kan identifiera sig själv direkt för att den ska ha ett begrepp om sig själv.

Självidentifikation inbegriper alltså inte sammanförandet av två entiteter utan av två beskrivningar, eller i vissa fall upplevelser, av samma entitet. Det är viktigt att inte hamna i en infinit regress beträffande identifikationen. Om Eva till exempel säger att hon är identisk med sin kropp och vet detta genom att hon direkt upplever kroppen som sin, så kan man ifrågasätta om det verkligen är *hennes* kropp hon upplever. Identifikationen får ju inte förutsätta att kroppen redan är identifierad som hennes. Ett liknande problem uppstår om Eva säger att hon är identisk med den golfspelare som vann tävlingen i Falsterbo 1995. Vi kan undvika en infinit regress genom att visa att det finns ett sätt man kan ha kunskap om sig själv på som inte behöver rättfärdigas och alltså inte kan ifrågasättas. Detta sätt förekommer i det som jag ovan kallat direkt självidentifikation. Vissa upplevelser, till exempel glädje eller lust, är sådana att om man har dem så kan man inte tvivla på att det är man själv som har dem. Sedan kan andra sorters självkunskap knytas till denna fundamentala kunskap.

Låt oss gå vidare och se på ytterligare några exempelsatser.

(4) En gång i tiden var jag golfspelare.
(5) Om några år ska jag flytta till Grönland.

Dessa satser skiljer sig från (1)–(3) genom att de förutsätter talarens identitet över tid, (4) historiskt sett och (5) in i framtiden. De förutsätter alltså att talaren har ett begrepp om personlig identitet. Vad personlig identitet består i har diskuterats grundligt under 1900-talet inom analytisk filosofi. De förslag som diskuterats mest hänvisar antingen till kroppslig eller själslig kontinuitet eller en blandning av de två som grund för personens identitet. Med själslig kontinuitet avses sådant som att talaren har minnen från tidigare perioder av sitt liv, och med kroppslig kontinuitet att inga vitala kroppsdelar (till exempel hjärnan) delvis eller helt bytts ut. Jag tänker inte gå in på dessa teorier här. De går tillbaka på metafysiska idéer om vad en person är. För att förklara hur ordet 'jag' kan peka ut talaren som en entitet vilken är självidentisk över tid räcker det, som strax ska förklaras, att hänvisa till rent pragmatiska eller konventionella överväganden.

Det är riktigt att föreställningar om saker och tings identitet påverkar språket. Den yttersta grunden till dessa föreställningar kan man söka på två håll – i tingens natur så till vida vi känner till den, och i hur vår psykofysiska konstitution påverkar vår uppfattning av tingen. Ovanpå denna grund läggs dels kulturella föreställningar om ting, dels förväntningar i fråga om hur dessa ting i vanliga fall brukar uppträda. Tillsammans bildar dessa faktorer något vi kan kalla social mening: en blandning av explicita påståenden och omedvetna förutsättningar. Denna mening är normativ i betydelsen att den sätter en norm för hur ett ord bör användas. Normen har en förankring i objektets natur och i hur människor, förutsatt att de fungerar normalt, upplever detta objekt. Normen har därigenom en stabil bas. De föreställningar som tillkommer kan å andra sidan skifta mellan olika språkgrupper eller över tiden. Den förståelse av personlig identitet som behövs för att en talare ska kunna behärska jagreferens inbegriper inte metafysiska teorier om vad en person är, utan vilar på en blandning av kulturell, i många fall så kallad tyst, kunskap och den omedelbara och direkta upplevelse varje talare har av sig själv -- av sin kropp och sina medvetandetillstånd.

Den omedelbara jagupplevelsen bildar alltså en sorts grund för förståelsen av personlig identitet. Man upplever sig själv som existerande genom tiden, i normala fall både psykiskt och fysiskt. Denna jagupplevelse sätter en norm för hur man tänker sig andra människors identitet, som bekräftas och späds på av de föreställningar om personer som finns i den kultur man lever. Låt oss säga att den sociala meningen bildar en konvention om vad en person är. Denna konvention ligger till grund för förståelsen av satser som (4) och (5). I situationer då det uppstår tveksamheter (dessa existerar än så länge mest i science fiction-artade tankeexperiment) är det inte lönt att vända sig till semantiken eller metafysiken. Pragmatiska överväganden får lösa tvisten om vem 'jag' refererar till.

Följande sats

(6) Jag kunde ha varit proffsboxare.

vilar också på en uppfattning om identitet, men inte över tid utan genom en serie samtidiga möjliga situationer. Först identifieras talaren i yttrandekontexten dels genom indexikal kunskap, dels genom social mening. För att kunna använda och förstå (6) räcker det sedan att man inser att man som talare kan uppfylla olika beskrivningar, och att det som faktiskt gäller för talaren kunde ha varit annorlunda under vissa omständigheter. Från (6) tycks steget inte långt till följande satser.

(7) Om jag vore Ronald Reagan skulle jag inte ställa upp som president.
(8) Jag kunde ha varit Sylvester Stallone, eller till och med en dator!

Skillnaden är dock ganska stor. Här tänker man sig inte att talaren bara kunde haft andra egenskaper än hon eller han faktiskt har, utan att talaren kunde ha varit en annan. Vad kan detta innebära – annat än att talaren kunde ha haft vissa egenskaper som, enligt exemplen ovan, Reagan, Stallone eller en dator, har? Satser som (7) och (8) kan inte tas bokstavligt. Om de tas bokstavligt framstår de som anomalier. Det går inte att försätta sig i en annan människas situation helt och hållet – man kan inte bli den andre. Däremot kan man föreställa sig att man själv har vissa av de egenskaper som den andre har. I satser som (7) och (8) är detta just vad som sker. Reagan, Stallone och datorn får stå för några egenskaper som är relevanta i det sammanhang satsen ut-

talas; de får representera något annat, eller kanske borde jag skriva mindre, än de är.

Sats (9) verkar först ge upphov till större problem än (7) och (8):

(9) Jag tror jag är Mike Tyson.

Den som yttrar denna sats och verkligen tror på vad han säger identifierar sig inte bara med vissa egenskaper hos Tyson, utan tror faktiskt att han är Tyson. Hur ska man förklara att talaren, säg det är Smith igen, lyckas peka ut sig själv i detta fall? Jo, Smith lyckas med det genom att han (i) följer den lingvistiska regeln för 'jag'; (ii) identifierar sig direkt; och (iii) känner till den sociala meningen hos 'jag'. Däremot kan han inte korrekt identifiera sig indirekt som en viss person, det vill säga Smith. Han talar om sig själv, men tror att han är någon annan.

Jag har ganska ensidigt uppehållit mig vid självidentifikation och talaren när jag gått genom exempelsatserna. Vad gäller egentligen för åhöraren? I början av artikeln visade jag på skillnaden i perspektiv mellan talare och åhörare i fråga om hur ordet 'jag' används. Åhöraren, skrev jag, behöver inte känna till mer än den lingvistiska regeln att 'jag' pekar ut talaren, medan talaren måste ha indexikal kunskap om sig själv för att korrekt referera till sig själv. Detta framstår nu som en sanning med modifikation. Ovan har jag nämligen sagt, att för att förstå vissa användningar av 'jag', som i satserna (4)–(6), måste man känna till ordets sociala mening. Den sociala meningen, i sin tur, bygger delvis på individens omedelbara upplevelse av sig själv och av hur hon interagerar med omvärlden. Detta innebär att indexikal kunskap ligger till grund för förståelsen av ordet 'jag' oavsett om det gäller talar- eller åhörarperspektiv. Även åhöraren knyter ordet 'jag' till indexikal kunskap, men inte om talaren, naturligtvis, utan om sig själv. Detta sker genom att den indexikala kunskapen generaliseras från det egna fallet till att även gälla andra. På så sätt bildar den en bas för social mening.

Låt mig nu återknyta till de tre teorierna om hur 'jag' pekar ut talaren. Utifrån hur förståelsen av ordet 'jag' förklarats i exempelsatserna tycks den indirekta teorin vara mest adekvat. Denna teori säger att 'jag' refererar till talaren, och att referensen förmedlas genom någon sorts beskrivning. Indexikal jagkunskap och social mening har här axlat rollen som beskrivning. Den teori som säger att 'jag' endast

indikerar talaren är felaktig, eftersom 'jag' refererar till en bestämd entitet, talaren, i en bestämd kontext, till exempel Jones vid tid *t* och på plats *p*. 'Jag' har dessutom en speciell kognitiv roll. Den direkta teorin, å sin sida, är inte tillräcklig. En kausal och/eller en omedelbar kognitiv förbindelse mellan ord och referent räcker inte till för att peka ut talaren. Referenten måste pekas ut *som* någonting för att ordet ska kunna sägas vara om denna referent. Referenten, i detta fall talaren, presenteras alltså på ett visst sätt genom att omtalas som 'jag', och denna presentation styr användningen av ordet.

Till sist vill jag återvända till den inledande frågan: finns det någon gåta bakom 'jag'? Tyvärr måste jag göra dem som läst denna text för att finna gåtans lösning besvikna. En semantisk redogörelse för hur en term fungerar får inte alltid en metafysisk knorr. Gåtan kvarstår.

Not

Jag vill tacka Peter Gärdenfors, Sten Lindström och Fredrik Stjernberg för värdefulla kommentarer.

Litteratur

EVANS, G, *Varieties of Reference*, Oxford University Press, 1982.

PERRY, J, *The Problem of the Essential Indexical and Other Essays*, Oxford University Press, 1993.

SHOEMAKER, S, *Self-Knowledge and Self-Identity*, Cornell University Press, 1963.

JOHN CANTWELL

Pågående handlingar

1. Introduktion

En formell behandling av handlingsteori kan givetvis, genom sin precision, ge värdefulla insikter i de filosofiska problem som omger handlingar. På senare år har det visat sig att en sådan behandling också kan bidra till lingvisternas projekt att ge en formell semantisk analys av satser i mänskliga språk. Eftersom så mycket av våran vardag handlar om att förstå och diskutera våra egna och andras gärningar, så verkar lingvisternas intresse för handlingsteori motiverat. Deras ambitioner är vidsträckta, kanske vidlyftiga, men en aspekt som bör intressera även den logiskt intresserade är deras ambition att ge en analys som korrekt återger det mönster av slutledningar – den *logik* – som omger vårt tal om handlingar.

Ur filosofisk synpunkt är problemet med en formell behandling av ett begrepp som *handling* eller, för den delen, det sätt vi pratar om handlingar ofta att den ger mer av en laboratoriemodell av de fenomen man försöker beskriva än en fullödig realistisk beskrivning av alla de komplexiteter som omger ett så vitt begrepp. Varje formell behandling blir således fattig i jämförelse med studieobjektet. Men detta bör inte hindra den som är intresserad att med formella metoder studera de aspekter som trots allt lämpar sig för en sådan behandling. En sådan aspekt är den logiska, en annan är den som i någon mån motiverar den logiska: den semantiska eller modellteoretiska, i bästa fall en representation av viktiga delar av den verklighet som satser om handlingar försöker beskriva.

En *pågående* handling exemplifieras av satsen:

(A) Stina är på väg till affären

Eller, helt enkelt:

(B) Stina är på väg

I engelskan finns det ett speciellt tempus – det *progressiva* eller *kontinuerliga* – som markerar att en handling pågår:

(A') Stina is walking to the store
(B') Stina is walking

Detta ger en viss flexibilitet i den språkliga behandlingen av pågående handlingar som i svenskan bara kan uppnås genom omskrivningar:

(A*) Stina är på väg till affären, hon går.
(A#) Stina håller på att gå till affären.

(B*) Stina är på väg, hon går

Ibland kan det således vara praktiskt att hålla den engelska formen i minnet, eftersom det ger ett lite ledigare språk, men principiellt så bör inga problem uppstå genom att titta enbart på den svenska formen.

Problemet med pågående handlingar kan enkelt illustreras med följande sats:

(C) Stina var på väg till affären när hon blev överkörd av en lastbil

Stina var på väg till affären, antag att hon gick till fots, hon kom aldrig dit utan hamnade på sjukhuset. Problemet är bara: eftersom hon aldrig kommer fram till affären kan man fråga sig med vilken rätt vi hävdar att hon överhuvudtaget var på väg till affären. Jämför med:

(D) Stina promenerade när hon blev överkörd av en lastbil

Helt klart är att vid, eller strax innan, tidpunkten för (D) så promenerar Stina, promenaden får ett något abrupt slut, men enligt (D) så har Stina promenerat, hon har också blivit överkörd. Om (C) däremot så kan vi inte säga att Stina gick till affären och, vidare, att hon blev överkörd, hon når ju aldrig fram till affären. Ändå tycks (C) kunna beskriva ett sant förhållande: folk är dagligen på väg någonstans när olyckor drabbar dem och vi finner inget konstigt i att beskriva detta med satser som (C). Problemet är: hur kan man yttra något sanningsenligt eller ens meningsfullt med en sats som (C)?

2. *Ett enkelt språk*

Infinitivformen av handlingsverb är den grundläggande formen i svenska språket för att tala om handlingar. Verbet (att) *gå*, till exempel, kan sägas namnge en viss handlingstyp, likaledes kan den något mer komplicerade verbfrasen (att) *gå till affären* också sägas namnge en viss handlingstyp (idén att ta infinitivformen som grundläggande förekommer både i von Wright (1983) och i Kamp (1993)). Infinitivformen av verb förekommer i satser: "Det är fel *att stjäla*", "Stina har för avsikt *att resa bort*" och i den mån man är intresserad av att studera dylika satser i ett formellt språk så verkar det rimligt att titta på språk där infinitivformen ges en explicit representation, men normalt är vi intresserade av olika aspekter i utförandet av en specifik handling: handlingen pågår, handlingen har slutförts, och så vidare. Dessa aspekter kommer här att omnämnas som olika *handlingsmodus* av ett handlingsverb.

Vi kan också prata om handlingars placering i tiden: handlingen utfördes igår, en handling föregick en annan, och så vidare. Självklart finns det många andra aspekter av handlingar som är intressanta, till exempel så är vi normalt intresserade av vem som har utfört handlingen (och vilka som drabbas av den), men då dessa aspekter inte är centrala för de problem som diskuteras här så kommer de att ignoreras: en anonym agent antas stå för alla handlingar.

Inom ramen för en sådan här uppsats går det inte att ge en fullständig redogörelse för dessa olika fenomen, utan här så kommer endast tre olika handlingsmodus att diskuteras tillsammans med enkla operatorer att relatera handlingar tidsmässigt.

Det logiska språket L kommer att bestå av en mängd InfV av infinitivformen av verb v_1, v_2,..., tre handlingsmodus *prog, culminate*, och *terminate*, de satslogiska konnektiven $\vee$, $\wedge$, $\neg$, $\rightarrow$, $\leftrightarrow$ och två satsoperatorer P och F som kommer att utläsas "vid någon tidpunkt i det förgångna...", respektive, "vid någon tidpunkt i framtiden...".

De välbildade satserna definieras induktivt:

(i) om $v \in$ InfV, så är v en *term*

(ii) om v är en term, så är *prog* v, *culminates* v och *terminates* v, välbildade satser

(iii) om ϕ och ψ är välbildade satser så är även $P\phi$, $F\phi$, $\phi \vee \psi$, $\neg\phi$, osv, välbildade satser.

Avsikten är att *prog* v skall utläsas "agenten (ospecificerad) håller på att v" och att *culminates* v skall utläsas "agenten har precis utfört (inte nödvändigtvis *slut*fört) handlingen att v". En handlingstyp som "att gå" kommer att kulminera, i varje ögonblick den utförs (varje del av en promenad är en promenad), en handlingstyp som "att gå till affären" däremot kommer att kulminera först när agenten har nått affären. *terminates* v skall utläsas "agenten har precis avbrutit (möjligtvis avslutat) handlingen att v". Givet dessa informella tolkningar så kan man fråga sig vilka satser som bör vara logiskt sanna. Till exempel, så bör följande vara logiskt sann:

culminates gå till affären → P *prog* gå till affären

Om agenten har gått till affären, så bör det gälla att agenten har varit på väg (till fots) till affären. Däremot bör inte följande vara *logiskt* sanna:

prog v → *culminates* v
prog v → F *culminates* v

Bara för att agenten håller på att gå till affären så kan vi inte sluta oss till att agenten *har* gått, eller kommer att ha gått, till affären.

Det här enkla språket verkar således vara rikt nog att representera de typer av satser och logiska samband mellan satser som diskuterades i förra avsnittet. Men frågan kvarstår, vad betyder en sats som *prog* v? Den informella tolkningen "agenten håller på att v" är inte vidare hjälpsam då betydelsen av "håller på att..." i sig är oklar.

3. Tillstånd, historier och spår

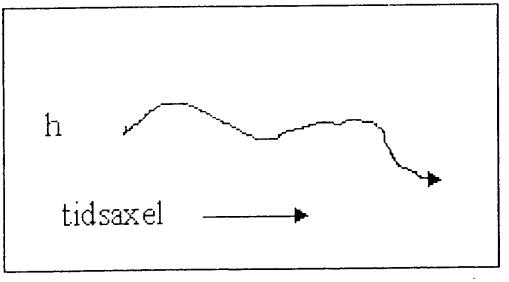
I ett visst ögonblick så befinner sig världen i ett visst tillstånd, under ett tidsintervall så kommer världen att passera igenom en sekvens av tillstånd. Ett tillstånd som det uppfattas här är något *totalt,* det vill säga att varje relevant saktillstånd är bestämt. Det är också *statiskt*, det vill säga, det går inte att ur ett givet tillstånd utläsa vad som har hänt, vad som kommer att hända eller vilka processer det är som pågår, för det krävs tillgång till världens tillstånd vid andra tidpunkter och, för processer, någon form av tidsligt intervall av tidpunkter runt det nuvarande tillståndet. Detta betyder inte att satser av formen "processen p pågår vid tidpunkten t" är meningslös, i varje fall inte om man med en sats mening avser dess sanningsvillkor. Det är fullt möjligt att låta

sanningsvillkoren för en sådan sats ta hänsyn till ett intervall av kringliggande tidpunkter, även om satsen bara nämner en tidpunkt.

Mängden av olika möjliga tillstånd benämns med U. Tid kommer att representeras av mängden heltal I och den sedvanliga linjära ordningen $\leq$ på I. Denna representation väljs av praktiska skäl, det finns ingen principiell anledning att inte använda de reella talen till exempel.

En världs passage genom olika tillstånd benämns dess *historia*. Här blir en historia en funktion från I till U, vid varje tidpunkt i så är världen i ett visst tillstånd. I någon mån kan man tänka sig en historia h som en *möjlig värld*, men detta synsätt kommer inte att användas här. Mängden av olika historier benämner vi H. Ett par (h, i) där $h \in H$ och $i \in I$, kan sägas utgöra en *punkt*.

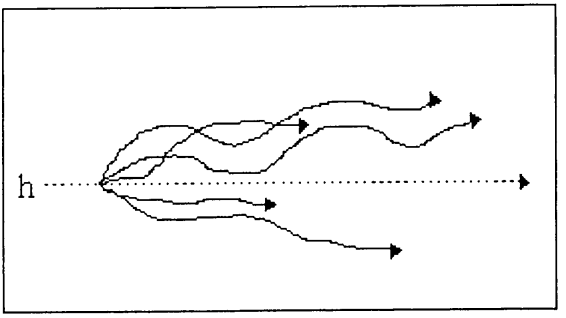
I det som följer kommer det att bli praktiskt att ge enkla bilder av annars ganska otympliga begrepp. En historia kan illustreras så här:



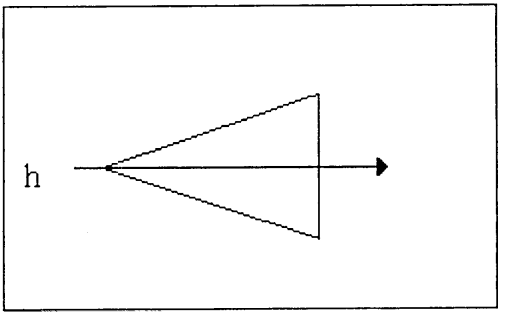
Figur 1: Schematisk beskrivning av en historia som passerar genom olika tillstånd

Antag nu att världen är i ett tillstånd där Stina står utanför sitt hus. Det verkar meningsfullt att fråga sig på vilka olika sätt världen kan fortskrida för att Stina skall ta en promenad till affären. Givetvis så kommer det finnas många olika sätt. Stina kan, under sin promenad, välja mellan olika gångvägar, hon kan välja att växla några ord med grannen på vägen, hon kan välja att se sig för innan hon går över vägen, eller så kanske hon väljer att springa över vägen utan att titta. Samtidigt så kanske Kalle i en helt annan stad beslutar sig för att gå till affären så en sekvens av tillstånd måste innefatta även Kalles promenad, men i en annan sekvens av tillstånd så ångrar sig Kalle. Kalles eventuella promenad är helt irrelevant för vår uppskattning av Stinas promenad men även hans handlingar påverkar världen. I praktiken skulle

det vara omöjligt att specificera ens en sådan sekvens av handlingar, men antag att det går att specificera inte bara en utan alla relevanta sekvenser. En annan sak är att världen kan fortskrida endast på ett sätt, och ingen av dessa sätt behöver sammanfalla med de sätt som utgör en promenad av Stina till affären. Detta illustreras i figur 2 och 3.



Figur 2: De möjliga sekvenser av tillstånd som skulle utgöra en promenad till affären och världens faktiska framfart



Figur 3: Samma som föregående men mer schematiskt

Om man för varje agent och varje tillstånd kunde specificera alla sådana sekvenser så har man i viss mån specificerat *meningen* hos uttrycket *att gå till affären.*

Dessa begrepp kan göras mera precisa. Ett *spår* är en sekvens $<u_1, u_2,...>$ av tillstånd $u_1, u_2,... \in U$, möjligtvis tom. p(1) benämner det första elementet av spåret p (om det finns något), p(2) det andra, och så vidare. Om p är ändlig, så benämner |p| antalet element i p. Det tomma spåret benämns med ∅. Låt P benämna mängden av spår i U. En mera utförlig beskrivning av spår ("paths") och historier finns i Segerberg (1985, 1992), flera av idéerna som presentera i detta avsnitt

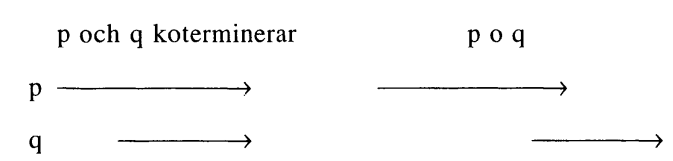
är försök att utveckla tankegångarna i Segerberg (1985, 1992), bland annat så finns där en mera utförlig beskrivning av spår – "paths" – och historier.

Två olika former av överlappning av spår kommer att användas, för varje spår p och q:

p och q *koterminerar* omm $\forall i_0$, om $1 \leq i_0 \leq |q|$, så $q(i_0) = p((|p| - |q|) + i_0)$

p o q omm $\forall i_0$, $1 \leq i_0 < |q|$, i_1, om $1 \leq i_1 \leq i_0$, så $p((|p| - i_0) + i_1) = q(i_1)$

Dessa kan enkelt åskådliggöras med diagram:



Dessa begrepp kan användas även för historier begränsade upp till en tidpunkt i:

h och q koterminerar^i omm $\forall i_0$, $1 \leq i_0 \leq |q|$, $q(i_0) = h((i - |q|) + i_0))$

h o^i p omm $\forall i_0$, $0 \leq i_0 < |q|$, i_1, $1 \leq i_1 \leq i_0$, $q(i_1) = h((i - i_0) + i_1)$

Med dessa begrepp kan man ge en preliminär modell-teoretisk tolkning av uttrycken i L. Låt V vara en funktion som för varje v InfV tilldelar en mängd av spår. En sådan tolkning kommer att vara en del av en större struktur M, en modell för L, men tillsvidare så får den vara ospecificerad.

Sanningsvillkoren för satser i L kommer att vara beroende av en historia och en tidpunkt. Ett första försök att ge sanningsvillkoren för *culminates* v lyder:

$M \models_{h,i}$ *culminates*$_0$ v omm $\exists p \in V(v)$, h o^i p

Enligt detta så skulle *culminates* v, agenten har precis slutfört handlingen att v, vara sann i en historia h vid en tidpunkt i, om något spår i v sammanfaller med något segment av h som slutar vid tidpunkten i. Världen har alltså, från någon tidpunkt tidigare än i fram till i gått igenom en sekvens av tillstånd som exakt beskriver en sekvens av tillstånd som skulle utgöra ett utförande av v.

Detta verkar vara en rimlig ansats men om man istället betraktar *prog* v, så bör det vara uppenbart att de begrepp som har införts är otillräckliga för att genomföra en analys av pågående handlingar. Detta blir klarare om man tittar på två otillräckliga analyser:

$M \models_{h,i} prog_0$ v omm $\exists p \in V(v)$, h o^i p
$M \models_{h,i} prog_1$ v omm $\exists p \in V(v)$, (h o^i p & $\exists i_0 > i$, h och p koterminerar^i)

$prog_0$ v försöker analysera en pågående handling som extrapolering, agenten har fram till tidpunkten i handlat på ett sätt som är förenligt med att utföra v, alltså håller agenten på att utföra v. Men betrakta följande situation: precis bortom affären nere på torget finns ett postkontor, så när agenten håller på att gå till affären så kommer agenten samtidigt att handla på ett sätt som är förenligt med att gå till posten. Enligt $prog_0$ så innebär detta att agenten både är på väg till affären och på väg till posten, vilket i allmänhet leder till orimligheter.

$prog_1$ är ett försök att eliminera problemen behäftade med $prog_0$, det räcker inte längre att agenten handlar på ett sätt som är förenligt med v, agenten måste också slutföra v. Men detta innebär att om agenten håller på att utföra v så kommer agenten att lyckas slutföra v, vilket är orimligt.

4. Avsikter och rutiner

Hur kan satsen "Stina var på väg till affären när hon blev överkörd" vara sann? Den mest uppenbara förklaringen är att Stina *avsåg* att gå till affären, och att hon handlade i enlighet med den avsikten när hon blev överkörd. Satser om pågående handlingar är till stor del satser om avsikter (se Dowty (1979) för en annan syn på pågående händelser).

Problemet med en sådan analys är att avsikten att utföra v utgör vare sig ett nödvändigt eller tillräckligt villkor för att sanningsenligt kunna säga att agenten håller på att utföra (har utfört) v, i varje fall inte enligt en ganska vanlig syn på avsikter.

Det är sant att om Stina håller på att promenera och avser med denna promenad att träffa Lars utanför affären, så är det korrekt att säga "Stina är på väg till affären". Men antag att Lars, tvärtemot vad han lovat Stina, inte är vid affären, det vore i sådan fall fel att säga "Stina är på väg att möta Lars", även om hon tror att hon är det. I

detta fall så utgör avsikten inget tillräckligt villkor för att göra den motsvarande satsen om den pågående handlingen sann.

Antag nu att Stina är blind och att affären är röd, vidare att Stina inte känner till att den är röd. I den ovan beskrivna situationen verkar det korrekt att säga "Stina är på väg till en röd byggnad", trots att hon inte har någon avsikt, *de dicto*, att gå till en röd byggnad. Avsikt verkar inte ens vara ett nödvändigt villkor.

En vanlig tanke är att avsikt är en nödvändig komponent av handlingsbegreppet. En handling kan ha oavsiktliga konsekvenser, men utan en avsikt så är det ingen handling.

Oavsett hur man ställer sig till ett sånt påstående så bör man medge att avsikter (som begreppet har använts här) inte räcker för att klargöra problemen med pågående handlingar.

En agent har förmågan att utföra stereotypiska rörelser med kroppen. Vissa rörelser är karakteristiska för det vi brukar kalla *att gå*. Att en agent utför sådana rörelser är inte tillräckligt för att man skall kunna säga att agenten går, om till exempel agenten befinner sig i vatten när han utför sådana rörelser, så verkar det inte korrekt att säga att han *går,* snarare att han trampar vatten. Givetvis så kommer det finnas situationer där agenten inte kan utföra dessa rörelser, till exempel om han är bunden.

Man kan tänka sig en agents förmågor som att agenten har olika *rutiner* tillgängliga som kan utföras närhelst agenten så vill och omständigheterna är passande. Givetvis så finns det många aktiviteter där kroppsrörelser inte är av intresse, mentala handlingar som att addera två tal, att fatta ett beslut och så vidare, men begreppet rutin bör kunna utvidgas till att gälla även sådana handlingar.

Mera problematiskt är komplicerade handlingar såsom att gå till affären eller att skriva en bok. Det verkar inte realistiskt att tänka sig att en agent har en fix och färdig rutin för att utföra sådana handlingar, det verkar heller inte speciellt meningsfullt att tänka sig en sådan rutin som ett rörelseschema, det finns för många sätt att utföra sådana handlingar på. I dessa fall verkar det mera rimligt att tänka sig att agenten har ett övergripande mål som kan uppdelas i delmål när behov uppstår, vilka i sin tur kan uppdelas i delmål, och så vidare. Att tänka sig även sådana färdigheter som utslag av rutiner är kanske att sträcka begreppet *rutin* för långt, alternativt skulle det behövas en betydligt mera ingående analys av begreppet.

Här så kommer jag utan vidare att ta mig friheten att betrakta komplexa handlingstyper som komplexa rutiner. Sådana antaganden kan ibland verka klargörande, ibland är de mera ett resultat av omfattande begreppsförvirring. Till exempel så kan man hävda att begrepp som *mål* och *delmål* svårligen kan förstås utan tillgång till begrepp som avsikt och, möjligen falska, trosföreställningar. Det verkar som dessa begrepp har smugit sig in bakvägen i analysen. Detta är förvisso sant, men inget annat har påstått. Avsikt är rimligen ett centralt begrepp i förståelsen av komplexa handlingar, men dess förklaringsvärde är högst indirekt. Vi behöver begrepp som slår en brygga från avsikter till handlingar och möjligtvis kan rutiner utgöra en sådan brygga.

Man behöver inte anta att agenten har en rutin kallad "att gå till affären", det räcker att påstå att agenten har olika rutiner varav vissa kan framkalla ett beteende som sanningsenligt kan påstås vara en promenad till affären. Rutiner analyseras därför inte som lingvistiska eller "propositionella" entiteter, många av problemen med avsiktsförklaringar bör därför kunna kringgås.

Vid en given tidpunkt så kommer agenten att ha en uppsättning olika rutiner att välja bland. Valet av rutin blir ett centralt begrepp. Här kommer ett *val* av rutin att förstås, inte som ett beslut att påbörja en rutin, utan som påbörjandet av rutinen, det första steget i rutinen. Det är även rimligt att tänka sig att agenten väljer att fortsätta en redan pågående rutin, eller, för den delen att avbryta en pågående rutin. Även om denna distinktion inte kommer att få någon direkt representation. Man kan givetvis tänka sig att de val som agenten gör är resultatet av avsikter, önskemål, och så vidare, men inte heller detta kommer med i den formella analysen.

Valet av rutin kan ses som ett *åtagande*. Genom att välja en rutin så har agenten, vid tidpunkten för valet, åtagit sig att fullfölja rutinen. Agenten kan vid ett senare tillfälle backa från detta åtagande genom att avbryta rutinen, vidare så kan externa faktorer som lastbilar brutalt sätta stopp för en rutin, men det ursprungliga valet av rutin, tillsammans med upprätthållande val, utgör ett så starkt åtagande att man kan säga att agenten håller på att utföra det som skulle utgöra ett lyckat utförande av rutinen.

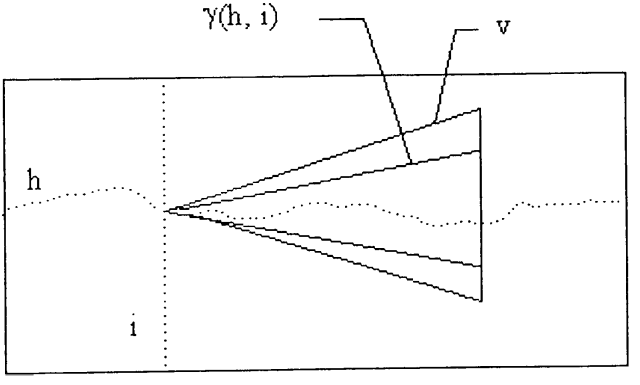
Här kommer det att antas att man kan förstå rutiner tillräckligt bra för att kunna säga vad som skulle utgöra ett lyckat utförande av

rutinen, det vill säga, givet att världen är i ett visst tillstånd, så kan man avgöra vilka fortsättningar som är förenliga med utförandet av rutinen. Detta antagande påminner väldigt mycket om det antagande som gjordes om handlingstyper. På samma sätt så antar vi att en rutin kan identifieras med de spår som utgör lyckade utföranden av rutinen.

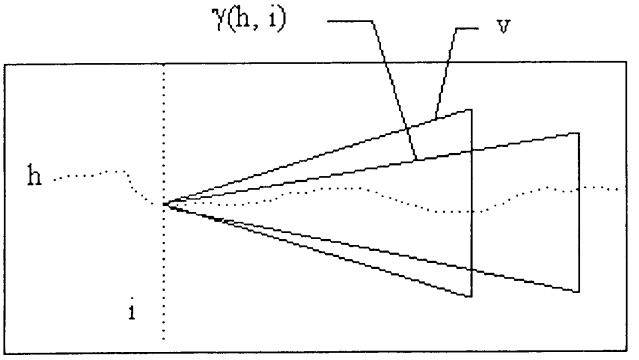
Valet av rutin, i sin tur, blir en funktion från den historia i vilken valet görs och den tidpunkt vid vilken valet görs, till de spår av den valda rutinen som har sitt ursprung i det tillstånd som världen är i vid den tidpunkten och i det ögonblicket. Detta blir enklare att uttrycka formellt. En rutin är en mängd spår, låt R vara mängden rutiner som agenten har till sitt förfogande, låt vara en funktion $H \times I \rightarrow \wp P$, sådan att:

(i) $(h, i) \in P$,
(ii) R, sådan att $p \in \gamma(h, i)$ omm $p \in \alpha$ och $h\ o^i\ p$

$\gamma(h, i)$, valet agenten utför vid h och i, kommer alltså att vara de spår av som är möjliga vid h och i. Handlingsbeskrivningar och rutiner är två olika begrepp som här har fått en likartad representation. Det förra representeras som de spår som är förenliga med det beteende som associeras med handlingstypen, det senare som de spår som utgör lyckade utföranden av rutinen. De två olika begreppen kan kombineras. Antag att världen befinner sig i ett visst tillstånd, om varje lyckat utförande av en rutin också är ett sätt att utföra en handlingstyp v, det vill säga, om varje spår av en rutin från en viss punkt också är ett spår av v, så skulle varje lyckat utförande av rutinen innebära att agenten har utfört handlingen att v. Detta kan illustreras på följande vis:



Faktum är att det räcker med att varje spår av γ(h, i) inkluderar ett spår av v:



Formellt blir detta:

M $\vDash_{h,i}$ *prog* v omm $\forall p \in \gamma(h, i), \exists q \in V(v)$, h o^{i} q & (p o q eller p och q koterminerar)

5. Culminates och terminates

Sanningsvillkoren givna för $culminates_0$ involverar bara agentens *beteende*, inte valet av rutiner. Detta innebär att följande sats kan falsifieras:

$culminates_0$ v $\rightarrow$ P *prog* v

Eftersom agenten för att hålla på med v måste välja en rutin som utför v, så innebär det, teoretiskt, att agenten utan tillstymmelse till val kan bete sig som om han har utfört v, utan att han någonsin har hållit på att utföra v. Det kan vara intressant att behålla distinktionen mellan att utföra v, och att bete sig *som om* man utför v, men det är å andra sidan inte uppenbart hur man skall göra den distinktionen. Antag att vid tidpunkten t_0 så bestämmer sig Stina för att ta av sommardäcken på sin bil och inget mer, vid tidpunkten t_1 så blir detta klart, men då inser Stina att hon lika gärna kan sätta på vinterdäcken, vilket hon gör, detta blir klart vid t_2. Vid t_2 verkar det korrekt att säga att Stina har bytt däcken på sin bil (det vill säga, har tagit av de gamla däcken och satt på nya). Det verkar också rimligt att säga att mellan t_1 och t_2 så *håller Stina på* att byta däcken. Mellan t_0 och t_1 håller hon däremot

inte på att byta däcken, även om hennes beteende är förenligt med den handlingen.

Detta och liknande exempel visar på svårigheterna att säga hur *håller på att v* och *har gjort handlingen att v* är relaterade. Den definition som ges här kommer bara att ta hänsyn till några av svårigheterna:

M $\models_{h,i}$ *culminates* v omm $\exists$p $\in$ V(v), h och p koterminerar^{i} & $\exists i_0 <$ i, $\forall$i', $i_0 \leq$ i' $<$ i, p $\in$ γ(h, i'), $\exists$q $\in$ V(v), h o^{i}q & (p o q eller p och q koterminerar)

Detta är ekvivalent med: :

M $\models_{h,i}$ *culminates* v omm $\exists$p $\in$ V(v), h och p koterminerar^{i} & $\exists i_0 <$ i, $\forall$i', $i_0 \leq$ i' $<$ i, M $\models_{h,i'}$ *prog* v

Slutligen:

M $\models_{h,i}$ *terminates* v omm $\exists i_0 <$ i, $\forall$i', $i_0 \leq i_1 <$ i,
M $\models_{h,i'}$ *prog* v & icke M $\models_{h,i}$ *prog* v

Det finns inte utrymme här att ge en mer ingående analys av det formella språk som har definierats, det kan räcka med att påpeka att följande satser blir logiskt sanna:

T	(tautologier)
culminates v $\rightarrow$ P *prog* v	
terminates v $\rightarrow$ P *prog* v	
prog v $\rightarrow$ *terminates* v	

Och att följande är falsifierbara:

(1) *prog* v $\rightarrow$ *culminates* v $\vee$ F *culminates* v
(2) *terminates* v $\rightarrow$ *culminates* v
(3) *culminates* v $\rightarrow$ *terminates* v

(1) och (2) bör vara ganska klara. (3) däremot förtjänar en kort kommentar. I varje naturlig modell så bör:

culminates gå till affären $\rightarrow$ *terminates* gå till affären

och

prog gå till affären $\rightarrow$ $\neg$*culminates* gå till affären

vara sanna. Om agenten utfört handlingen att gå till affären, så har han slutfört denna. Och om agenten är i färd med att gå till affären, så har han ännu inte utfört denna handling. Samtidigt bör:

prog gå → *culminates* gå

vara sann. När man håller på att gå så har man utfört ett gående.

culminates gå → *terminates* gå

däremot är falsk, ty, när man har utfört ett gående, så behöver detta ej vara slutfört. Det går att karaktärisera dessa olika fenomen semantiskt, men utrymmet förbjuder detta.

Litteratur

DOWTY, DAVID (1979), *Word Meaning and Montague Grammar*, Dordrecht: Reidel

KAMP, HANS, UWE REYLE, (1993) *From Discourse to Logic*, Kluwer Academic Publishers, Dordrecht.

SEGERBERG, KRISTER, (1985) "Routines", *Synthese* 65, pp.211–233.

SEGERBERG, KRISTER, (1992) "Representing Facts", i *Knowledge, Belief and Strategic Interaction*, Cristina Bicchieri and Maria Lousia Dalla Chiara (eds.), pp. 239 – 256, Cambridge University Press.

VON WRIGHT, GEORG HENRIK (1983), *Practical Reason*, Blackwell, Oxford 1983.

BERTIL ROLF

Om trovärdighet

1. Typer av trovärdighetsproblem

Det finns olika typer av problem som förknippas med trovärdighet. Kunskapsproblem handlar om skäl för att lita på information vi får genom expertis och vittnen. Hur kan man veta så mycket om världen, trots att man själv sett en så begränsad del av den?

En annan typ av problem har behandlats av spelteoretiker och ekonomer. Problemen gäller trovärdighet hos löften eller hot. Är exempelvis en ekonomisk politik mer trovärdig med en oberoende riksbank?

En tredje typ av problem avser social och politisk ordning. I "Making Democracy Work" (1993) studerar Robert Putnam två sociala ordningar i Italien i ett försök att knyta samman tillit ("trust"), mellanmänskliga normer och ekonomisk-politiska institutioner. En svensk studie över problematiken är Per Molanders vackra bok "Akvedukten vid Zaghouran" (1994). Till denna grupp hör också studier över den minskande tillit till folkvalda politiker som visar sig i flertalet västliga länder.

En fjärde typ av problem handlar om moralisk auktoritet och trovärdighet. Ibland förutsätts att talesmän för en moralisk ståndpunkt har en "trovärdighet" som bl a innebär att de efterlever normer som de pläderar.

I det följande skall jag diskutera de två första typerna av trovärdighetsproblem. Jag skall visa att trovärdighet av löften eller hot till stora delar kan behandlas med samma slags begrepp som trovärdighet av information.

2. Det s k "genetiska felslutet" och bevisvärdemodellen

Begreppet "trovärdighet" kom jag i kontakt med för många år sedan då jag höll kurser i argumentation, källkritik och vittnespsykologi.

Läroböckerna i argumentationsanalys dömde ut argument ad hominem. Vittnespsykologi och källkritik tycktes systematiskt tillämpa något som liknar argument ad hominem.

I stället för termen "ad hominem" talar argumentationsteoretiker mer precist om "det genetiska felslutet". Det förmodade felslutet består i att "man angriper personen *istället* för hans påståenden". Men vittnespsykologi och källkritik bygger på att det inte finns någon motsättning mellan granskning av person och granskning av sak. Båda har utvecklat systematiska rutiner för att studera hur en utsaga uppkommit. Utsagans uppkomst anses väsentlig för den tillit vi bör fästa till den. Det tycks ligga något rimligt i att en person som har starka intressen att få oss att tro att p kanske inte är den mest tillförlitliga källan rörande p.

Det s k "genetiska felslutet" kan analyseras med metoder som hämtas från bevisvärdering. I början av 1970-talet utvecklades en teori om juridisk bevisvärdering vars grundläggande tankar återfinns hos Per-Olof Ekelöf (1964), Sören Halldén (1973) och Martin Edman (1973).

Ett centralt problem avser förhållandet mellan frekvens och evidens. Låt säga att en anlagd brand bryter ut i det lilla samhället. På orten bor en person som tidigare dömts för mordbrand och haft tillfälle och kunskaper att utföra dådet. Antag att återfallsfrekvensen är 90% och att det är ytterst osannolikt att någon annan haft nog lokalkännedom för att anlägga branden. Varför inte gripa och döma honom för mordbranden? Det rimliga svaret är att personen inte "kan bindas till" brottet, trots att vi av allmänna samband kan sluta oss till en hög sannolikhet för att han begått dådet.

Vad betyder då denna fras "kan bindas till"? Det som binder en person till brottet är evidens i form av fingeravtryck på en bortkastad fotogendunk eller vittnesuppgifter om möten med den tilltalade i närheten av brandplatsen, utsändande en stark röklukt.

Bevisvärdeteorin går ut på att domstolens bedömning vilar på evidens för att något slags "indiciemekanism" eller "informationskedja" (termerna är Halldéns), eventuellt i form av en kausalkedja, finns mellan den anlagda branden och de senare vittnesmål eller den tekniska bevisning som förevisas domstolen. Domstolen överväger inte (eller bör inte överväga) den direkta sannolikheten för att X begick H. Domstolen överväger istället sannolikheten för att vittnesmål och teknisk bevisning beror av att X begick H. Evidensen för "Per tände

på Påls lada" stärks via evidensen för "Pers antändning av Påls lada är kausalt länkat till förekomsten av Pers fingeravtryck på den fotogendunk som undersökts av rättsteknikerna".

Bevisvärdeteorin löser problemen med det s k "genetiska felslutet". Bakom en utredning som uttalar sig om tobaksrökningens oskadlighet finns urval, metoder och procedurer. Om utredningen sponsrats av en cigarettfabrikant kan man inte utesluta att rapportens utsaga om oskadligheten har annat upphov än den påstår. Uppkomsten är relevant för vår tilltro. Argumentationsteoretiker som dömt ut det s k "genetiska felslutet" tycks ha trott att det genetiska betraktelsesättet försökt argumentera från "undersökningen har sponsrats av Philip Morris" till "rökning är skadligt" vilket givetvis vore ett felslut.

3. Generalisering av teorin till trovärdighet av information

Under 1980-talets början kom Nils-Eric Sahlin (1986) och jag själv (Rolf 1983) att, oberoende av varandras insatser, generalisera bevisvärdeteorin till en generell modell för kunskap eller kunskapsöverföring. Med viss förenkling kan modellen uttryckas:

> (K) X vet att p om och endast om (1) X tror att p; (2) Det finns en fungerande informationskedja mellan p och X:s tro att p och (3) X har evidens för (2).

Medan reliabilitetsteorier för kunskap nöjer sig med villkor (1) och (2) ställer (K) också krav på evidens. Motivet för krav (3) framgår genom fabeln om pojken och vargen. För att lura byborna ropade vallpojken "Vargen kommer" nio gånger utan att vargen kom. Den tionde gången kom vargen men byborna hade slutat lita på pojkens rop. Byborna har evidens mot (2) och pojken kan inte längre överföra sin kunskap till dem, trots att informationskedjan faktiskt fungerar i det tionde fallet.

Genom kombinationen av evidens och reliabilitet på ett sätt som icke är ad hoc, kringgår (K) Russell-Gettier problematiken. Om det finns falska mellanled mellan personens övertygelse och sakförhållandet p, så är villkor (2) inte uppfyllt.

Idag känner jag tveksamhet inför (K) som ett bidrag till samtida kunskapsteori. Dels förstår jag inte hur genren relaterar sina teorier till preteoretiska kunskapsproblem och kunskapsbegrepp eller hur man använder data, argument och metoder. Dels vill jag fokusera de procedurer och institutioner som levererar information snarare än det slut-

resultat "kunskap" som eventuellt uppkommer i den som mottar informationen och då till följd av ett flertal andra faktorer, exempelvis bakgrundsinformation. Jag föreslår i stället en modell som fångas i följande definition:

> DEF: Låt A vara en person eller institution som i situationen s utfärdar påståendet p som mottas av B. A:s utsägande av p i s skall sägas vara "trovärdigt" för B om och endast om B har evidens för att det finns en tillförlitligt fungerande informationskedja mellan p och A:s utfärdande av p i s

I förenklande syfte bortser modellen från gradering av trovärdighet.

4. Tillförlitliga informationskedjor

Ett problem gäller vad som skall menas med att en informationskedja eller indiciemekanism fungerat "tillförlitligt" och hur man etablerar evidens för det. En vanlig tanke är att bestämma tillförlitlighet genom en kontrafaktisk villkorssats: Vittnesmålet om p skulle inte förelegat såvida inte p varit fallet.

Denna lösning kan medföra problem. Säg att en notorisk skattefuskare till taxeringsmyndigheterna varje år deklarerar uppgift D(p) om nollinkomst. Att nollinkomst förelegat kallar vi "p". Men 1994 gjorde han stora spekulationsförluster och hade faktiskt nollinkomst. Den kontrafaktiska teorin tycks då medföra att informationskedjan fungerat endast om:

> (S) Skattefuskaren skulle inte ha deklarerat nollinkomst såvida han inte haft nollinkomst.

Men (S) är falsk, trots att informationskedjan fungerat. Det kan rent av vara så, att taxeringsmyndigheten har evidens för att informationskedjan fungerat vid 1994 års taxering, trots att skattefuskaren alltid tidigare bluffat. Han kan vid denna taxering förete korrekt bestyrka yrkanden på avdrag.

Problemet med den kontrafaktiska kopplingen är att utan p kan något annat komma att ge upphov till D(p). Den lösning som förefaller mig mest rimlig är att betrakta uppgiften D(p) som en rigid designator: just *denna* deklarationsuppgift jämte underlag, D(p), som uppger nollinkomst, skulle inte ha förelegat såvida inte nollinkomst förelegat. Vi kan då precisera:

DEF: Säg att det föreligger information om p, I(p). En informationskedja bakom I(p) sägs ha fungerat "tillförlitligt" om och endast om i varje tillstånd där *denna* informationskedja föreligger, föreligger också p.

Säg att bevismaterial för p i form av en blodig handske företes tillsammans med undersökningsprotokoll av handsken jämte protokoll om att handsken hittats i mordoffrets bil. För domstolen är detta yttersta ändpunkten av en informationskedja vars närmare egenskaper man inte känner. Man överväger då om *denna* informationskedja som levererat handsken skulle kunna finnas i en värld där p är falsk. Frågan man bör ställa sig är *inte*: "Skulle detektiven ha planterat bevismaterial om han inte funnit något". Utan den för bevisningen relevanta frågan är: "Planterade han *detta* bevismaterial *den här* gången?" Svaret på frågan beror av de möjligheter som är förenliga med *denna* indiciemekanism.

5. Trovärdighet av utfästelser

Det finns trovärdighetsforskning som sysslar med ett annat begrepp "trovärdighet". En gren härstammar från Thomas Schelling "The Strategy of Conflict" (1963). Bakgrunden är terrorbalansen och den amerikanska strategi som i fackkretsar förkortades MAD (Mutual Assured Destruction). Strategin gick ut på att USA skulle bemöta ett sovjetiskt angrepp på Västeuropa med en kärnvapenattack. Detta antogs leda till en okontrollerad eskalation av kärnvapenkriget. Ett sådant hot är svårt att göra trovärdigt. Hotet har ungefär samma logik som att försöka skrämma bort inbrottstjuvar i grannens hus genom att hota spränga sig själv i luften.

Liknande problem kan uppkomma vid överenskommelser. Ett stort byggföretag kontrakterar en hantverkare som utför en installation. Efter att installationen har utförts kan byggföretaget underlåta att betala eftersom hantverkaren inte gärna kan göra sitt arbete ogjort. Kan byggföretaget med trovärdighet utfästa sig att betala?

Det finns två slags analyser av varför utfästelser förknippas med trovärdighetsglapp. En analys (Coleman 1990, Persson 1987) går ut på att det finns ett tidsglapp mellan utfästelse och verkställan. Generellt sett har många utfästelser, dvs löften och hot, formen:

A utfäster sig vid t_1 [A ser till att p vid t_2] (där t_1 föregår t_2)

A kan ändra sina preferenser mellan t_1 och t_2. Odysseus skulle efter att ha hört sirenernas sång ha övermannats av sin längtan att spendera sin återstående tid i deras sällskap. I fallet MAD uteslöt USA:s preferenser både vid t_1 och vid t_2 att man skulle komma att utlösa ett storskaligt kärnvapenkrig.

En andra typ av analys (Schelling 1963, Elster 1984) vill förstå trovärdighetsglappet som ett kontrollglapp. Jag tror detta är riktigt. Mer generellt kan man säga att trovärdighetsproblemen för utfästelser beror av följande underliggande form:

A utfäster sig i s_1 [Någon ser till att p vid s_2] (där s_1 och s_2 är olika situationer)

Ofta finns komplexa antaganden bakom en utfästelses trovärdighet. Ibland vet man inte exakt vem som utför den handling man åtar sig. Om en städfirma åtar sig ett uppdrag, åtar man sig bara att se till att någon utför städningen. Trovärdighetsglapp kan då uppkomma genom att fallissemang hos någon annan än den som gjort utfästelsen.

Tidsglapp kan vålla glapp i trovärdighet. Men det beror då av att agenten som gör utfästelsen själv kan komma in vid t_2 samt att han inte vid t_1 kan binda besluten vid t_2. Själva tidsglappet är en konsekvens. Trovärdighetsglapp uppkommer genom oförmåga att binda instanser som kommer mellan utfästelse och verkställighet.

6. Tillförlitlighet och trovärdighet av utfästelser

Vi har diskuterat trovärdighet hos informationskedjor. Man kan använda en symmetrisk begreppsapparat för att tala om trovärdighet hos vad jag vill kalla *utfästelsekedjor*. Bevisvärdeteorin och mer generellt, trovärdighetsteorin, handlar om sådan evidens som kan "binda" en person till ett faktum. Trovärdighet av löften eller hot handlar om att en person skall "binda sig" till en framtida händelse. Vi kan tala om tillförlitlighet och trovärdighet enligt nedan:

DEF: Säg att det föreligger en utfästelse om q, U(q). En utfästelsekedja bakom U(q) sägs fungera "tillförlitligt" om och endast om i varje tillstånd där *denna* utfästelsekedja föreligger, föreligger också q.

DEF: Låt A vara en person eller institution som i situationen s utfärdar utfästelsen q som mottas av B. A:s utfästelse av q i s, U(q)

skall sägas vara "trovärdig" för B om och endast om B har evidens för att det finns en tillförlitligt fungerande utfästelsekedja bakom U(q).

Som framgår finns en stark symmetri mellan trovärdighet/tillförlitlighet hos information och hos utfästelser. Det finns en skillnad i tempus, men av skäl jag anfört tycks tidsskillnaden vara irrelevant.

7. Institutionella villkor för informations- och beslutshantering

Symmetrin mellan de två teorierna om trovärdighet ger verktyg för att analysera institutioner. Halldén (1980) anger en del praktiska tumregler för nyhetshantering som kan återföras till matematiska egenskaper hos sammansatt evidens. Multiplikationseffekter talar för att man bör ha korta informationskedjor med hög evidens för att varje länk korrekt överför information. Additionseffekter talar för att man hamnar i en bättre situation om man har oberoende informationskedjor.

Motsvarande tumregler gäller för utfästelser. Mängden av led mellan beslutsfattare och verkställare i en byråkratisk pyramid tenderar att dra ner trovärdigheten av utfästelser. Om det mellan den centrala myndigheten och verkställandet finns motsträviga instanser ute i provinsen, kan den totala sannolikheten för att utfästelsekedjan leder till verkställan inte överstiga sannolikheten för att den svagaste länken kommer att agera lojalt. Ett företag kan öka trovärdigheten av sina utfästelser genom att förlita sig på underleverantörer som är oberoende med avseende på råvaror och avtal med arbetskraften.

Trovärdighet beror inte av den evidens man själv har utan av den evidens mottagaren har. I "4.15 från Paddington" låter Agatha Christie sin hjältinna Mrs McGillicuddy dåsa till i kupén när tåget lämnar stationen. Ett annat tåg kör parallellt med hennes tåg, en gardin flyger upp och Mrs McGillicuddy ser en man strypa en kvinna på det andra tåget. Hon tillkallar konduktören. När denne kommer in i kupén ser han en gammal dam som i upprivet tillstånd berättar om ett mord hon sett. Konduktören noterar att på sätet där den gamla damen dåsat ligger en tidning uppslagen. Den visar en bild där en man stryper en kvinna. Konduktören är misstrogen mot damens uppgifter, i synnerhet som hon medger att hon dåsat till. Trots att Mrs McGillicuddy (och läsaren) vet att ett mord har begåtts kan hon inte överföra sin kunskap till konduktören. Utan trovärdighet blir hon en epistemisk ändhållplats.

Trovärdighet av information beror ofta av insyn i den process som är upphov till informationen. När man tillsätter en utredning om missförhållanden inom en samhällssektor, kan det vara exempelvis vara viktigt att utredningen medger insyn från dem man vill övertyga.

Trovärdighet av utfästelser beror av andras evidens. Är svensk neutralitetspolitik trovärdig? Det är möjligt att politiken är starkt trovärdig i förhållande till den evidens vi själva besitter om vår orubbliga sinnesförfattning och fasta beslutsamhet. Mer intressant är kanske frågan om neutralitetspolitiken är trovärdig i förhållande till den evidens som ryska bedömare kan besitta. Dels kan de rimligen inte veta hur orubbliga vi är i våra själar. Dels kan det hända att ryska principer för att vikta evidens mer ser till yttre faktorer: den svenska neutralitetspolitiken under andra världskriget, långvarigt svenskt samarbete med NATO på olika nivåer, ökade svenska åtaganden i Baltikum, minskade försvarsanslag och inrikespolitisk oenighet huruvida svenskt försvar har adekvat styrka för att säkra utrikespolitiska mål i en kärvare utrikespolitisk miljö.

Det är lättare att vara trovärdig inför sig själv än inför andra. En analys av vilken evidens andra använder för att bedöma ens trovärdighet kan ha relevans för beslut och öka den egna förståelsen av omgivningens reaktioner.

8. Vad är poängen med trovärdighet?

Vittnespsykologi och källkritik handlar om vad slags utsagor som får utgöra underlag för slutledningar. Detta ger en ledtråd till poängen med trovärdighetsbegreppet. Jag är böjd att tro att trovärdighet ofta är ett redskap för grovsållning av det brus av information och erbjudanden som man möter i samhället. Det som saknar trovärdighet hamnar under den nivå där signalerna har betydelse för vad vi sätter lit till.

Vad är det en person har förlorat när han förlorat sin trovärdighet? Fabeln om pojken och vargen vill varna oss för att förbruka vår trovärdighet. Den som förbrukat sin trovärdighet kan inte längre informera rationella åhörare. Den affärsman eller politiker som förlorat sin trovärdighet kan inte längre träffa bindande avtal med rationella parter.

Fabeln om pojken och vargen tycks antyda att trovärdighet är en egenskap hos en person som denne genom egen förtjänst kan vinna eller förlora. Så är i allmänhet inte fallet. Trovärdighet kan man förlora genom strukturella sociala förändringar av de kedjor som

binder information och utfästelser. Om dagens finansminstrar är mindre trovärdiga än finansminstrarna på 1950-talet, så beror detta av att positionen förlorat många bindningsmöjligheter, inte av personliga brister hos dess innehavare.

Trovärdighet är inte monadisk egenskap utan ytterst en flerställig relation. Bristande trovärdighet beror alltid på relationer och endast ibland på personliga egenskaper. Den som hamnat i fel slags relationer är, eller uppfattas som, en del av det otillförlitliga brus som inte informerar och som inget åstadkommer.

Jag vill tacka Martin Edman, Nils-Eric Sahlin, Wlodek Rabinowicz och deltagarna i filosofiseminariet i Lund för synpunkter och frågor som undanröjt en del av mina tidigare grumligheter.

Litteratur

COLEMAN, J S (1990): *Foundations of Social Theory*, Cambridge, Mass, The Belknap Press.

EDMAN, M (1973): "Adding independent pieces of evidence", i *Modality, Morality, and Other Problems of Sense and Nonsense: Essays dedicated to Sören Halldén*, Lund.

EKELÖF, P O (1964): "Free Evaluation of Evidence", *Scandinavian Studies in Law.*

ELSTER, J (1984): *Ulysses and the Sirens*, Cambridge University Press.

HALLDÉN, S (1973): "Indiciemekanismer", *Tidsskrift for rettsvitenskap.*

HALLDÉN, S (1980): *Nyfikenhetens redskap*, Lund, Studentlitteratur.

MOLANDER, P (1994): *Akvedukten vid Zaghouran*, Stockholm, Atlantis.

PERSSON, T (1987): *Credibility of Macroeconomic Policy: An Introduction and a Broad Survey*, Stockholm.

PUTNAM, R D (1993): *Making Democracy Work*, Princeton University Press.

ROLF, B (1983): *Kunskapskritik och kunskapsgenetik,* Lund.

SAHLIN, N-E (1986): "How to be 100% certain 99,5% of the time", *The Journal of Philosophy.* Ursprungligen publicerad som kompendium 1983.

SCHELLING, T (1963): *The Strategy of Conflict*, Cambridge, Mass, Harvard University Press.

JAN ODELSTAD

Förväntad nytta i ljuset av alternativa numeriska representationer

1. Inledning

Principen om förväntad nytta spelar en viktig roll i beslutsteorin. Enligt denna princip är nyttan hos ett handlingsalternativ den förväntade nyttan hos handlingens utfall. Om $\{c_i | \ i \in I\}$ är de möjliga konsekvenserna av en handling a och för alla i, $i \in I$, c_i inträffar med sannolikheten p_i, så gäller enligt förväntade-nyttoprincipen att nyttan hos a är

$$\Sigma_{i \in I} \, p_i \cdot u(c_i)$$

där $u(c_i)$ är värdet eller nyttan hos utfallet c_i. Förväntade-nyttoprincipen kombineras ofta med tanken att man ska välja den handling som har den största [förväntade] nyttan.

Det tycks vara en vanlig uppfattning att principen om förväntad nytta kan ges ett formellt stöd. von Neumanns och Morgensterns klassiska teori om sannolikhetsblandningar från 1947 anses t ex ofta ge en axiomatisk grundval för denna princip. Per-Erik Malmnäs har dock nyligen (se Malmnäs (1991) och (1994)) argumenterat för att varken von Neumanns och Morgensterns eller andra existerande axiomatiska karakteriseringar av nyttobegreppet ger en formell grund för förväntade-nyttoprincipen. Jag ska i denna not diskutera några aspekter på frågan om den axiomatiska grundvalen för förväntade-nyttoprincipen. Som en bakgrund för denna diskussion ges i nästa sektion en kortfattad skiss av en variant av von Neumanns och Morgensterns teori om sannolikhetsblandningar.

2. Blandningsrum

$\langle A, \geq, \theta \rangle$ är ett *blandningsrum* (mixture space, se Roberts (1979) s 353) om $\geq$ är en binär relation på A och θ är en funktion sådan att

$$\theta : A \times [0,1] \times A \rightarrow A$$

där [0,1] är det slutna intervallet mellan 0 och 1. Som brukligt är använder vi "apb" för att beteckna "θ(a,p,b)". Vi kallar apb för en [sannolikhets]blandning (med utfallen a och b). Låt

$$\theta_p(a,b) = \theta(a,p,b) = apb.$$

Observera att θ_p är en binär operation på A. Vi kan alltså se θ som en familj av binära operationer på A.

Den intenderade tolkningen av en [sannolikhets]blandning apb är ett lotteri (spel) som med sannolikheten p ger a som utfall och med sannolikheten 1-p ger b som utfall. Observera att utfallen a och b själva kan vara lotterier (blandningar). Enligt den intenderade tolkningen förstås ≥ som en preferensrelation, dvs. a ≥b innebär att a är åtminstone lika bra som b.

Låt oss säga att ⟨A, ≥,θ⟩ är ett N-M-blandningsrum (uppkallat efter von Neumann och Morgenstern) om ⟨A, ≥,θ⟩ är ett blandningsrum och följande axiom är uppfyllda för alla a,b,c i A (se Roberts (1979) s 354).

(A1) ≥ är en svag ordning på A.

> är motsvarande strikta svaga ordning och = är motsvarande indifferensrelation.

(A2) (apb) = [b(1-p)a] för alla p sådana att 0≤p≤1.

(A3) [(apb)qb] = a(pq)b för alla p,q sådana att 0≤p,q≤1.

(A4) Om a > b, så (apc) > (bpc) för alla p sådana att 0<p<1.

(A5) Om a > b > c, så existerar p,q sådana att 0<p,q<1 och (apc) > b och b > (aqc).

I von Neumann och Morgenstern (1947) visades att följande representations- och entydighetsteorem gäller för N-M-blandningsrum (se Roberts (1979) s 356).

Representationsteoremet: Om ⟨A, ≥,θ⟩ är ett N-M-blandningsrum så finns en reellvärd funktion u definierad på A sådan att

$$a \geq b \text{ om och endast om } u(a) \geq u(b) \quad (1)$$

$$u(apb) = pu(a)+(1-p)u(b) \quad (2)$$

Entydighetsteoremet: Om u och v är funktioner som uppfyller (1) och (2) ovan så finns det reella tal r och s där r är större än 0, sådana att

v(a) = ru(a)+s för alla a i A. (u och v är således intervallskalor.)

Den intenderade tolkningen av funktionen u, vilken uppfyller villkoren (1) och (2) ovan, är att u är ett nyttomått för blandningsrummet $\langle A, \geq, \theta \rangle$.

Notera att (2) är en formulering av den förväntade nyttoprincipen; enligt (2) är nyttan hos blandningen apb den förväntade nyttan av utfallen a och b.

Vi ska nu se närmare på innebörden i representationsteoremet. Låt för den skull ω vara den funktion sådan att

$$\omega : \mathrm{Re} \times [0,1] \times \mathrm{Re} \rightarrow \mathrm{Re}$$

(där [0,1] är det slutna intervallet mellan 0 och 1) *och* det gäller att $\omega(x,p,y) = px+(1-p)y$. Låt $\omega_p(x,y) = \omega(x,p,y) = px+(1-p)y$. ω_p är en binär operation på Re. Vi kan alltså se ω som en familj av binära operationer på Re. Det kan vara bekvämt att använda $+_p$ som beteckning för ω_p och vi skriver $\omega_p(x,y) = x+_p y$. Således gäller att

$$x+_p y = px+(1-p)y.$$

Representationsteoremet innebär alltså att ett N-M-blandningsrum $\langle A, \geq, \theta \rangle$ kan homomorft representeras med $\langle \mathrm{Re}, \geq, \omega \rangle$ i följande mening: Det finns u : Re $\rightarrow$ Re sådan att (1) ovan gäller samt

$$u(\theta(a,p,b)) = \omega(u(a),p,u(b)) \quad (3)$$

Notera att (3) är ekvivalent med att

$$u(\theta_p(a,b)) = \omega_p(u(a),u(b)) \quad (4)$$

(3) är ju ett sätt att formulera den förväntade nyttoprincipen, vilken alltså utsäger att θ representeras med ω.

3. Alternativa numeriska representationer av blandningsrum

Det är lätt att inse att ett N-M-blandningsrum $\langle A, \geq, \theta \rangle$ kan homomorft representeras med andra numeriska strukturer än $\langle \mathrm{Re}, \geq, \omega \rangle$. I så fall kommer θ att representeras med en annan familj av binära operationer än ω, och således kommer den förväntade nyttoprincipen att ersättas med en annan princip. Låt oss se mer i detalj på detta. Först en definition.

Ett N-M-blandningsrum $\langle A, \geq, \theta \rangle$ kallas *icke-negativt* om det finns

en homomorfi u från $\langle A, \geq, \theta\rangle$ till $\langle Re, \geq, \omega\rangle$ sådan att u(a) är ett icke-negativt reellt tal för alla $a \in A$. $\langle A, \geq, \theta\rangle$ är alltså icke-negativt om det finns en homomorfi från $\langle A, \geq, \theta\rangle$ till $\langle Re^*, \geq, \omega\rangle$ där $Re^* = Re^+ \cup \{0\}$, dvs. Re^* är mängden av icke-negativa reella tal.

Vi ska nu se på en annan numerisk representation av ett icke-negativt blandningsrum $\langle A, \geq, \theta\rangle$ än $\langle Re, \geq, \omega\rangle$. Låt μ vara den funktion sådan att

$$\mu : Re^* \times [0,1] \times Re^* \rightarrow Re^*$$

och $\mu(x,p,y) = \sqrt{(px^2+(1-p)y^2)}$. Vi kan alltså se μ som en familj av binära operationer på Re^*. Låt $\mu_p(x,y) = \mu(x,p,y) = \sqrt{(px^2+(1-p)y^2)}$. Det kan vara bekvämt att använda $\otimes_p$ som beteckning för μ_p och vi skriver $\mu_p(x,y) = x \otimes_p y$. Således gäller att

$$x \otimes_p y = \sqrt{(px^2+(1-p)y^2)}.$$

Notera att om $f : Re^* \rightarrow Re^*$ sådan att $f(x) = \sqrt{x}$ så är f en isomorfi från $\langle Re^*, \geq, \omega\rangle$ till $\langle Re^*, \geq, \mu\rangle$. Ty f är omvändbar samt

$$x \geq y \text{ omm } f(x) \geq f(y)$$

och

$$f(\omega(xpy)) = \sqrt{(px+(1-p)y)} = \sqrt{(p(\sqrt{x})^2+(1-p)(\sqrt{y})^2)} =$$
$$= (\sqrt{x}) \otimes_p (\sqrt{y}) = \mu(f(x),p,f(y))$$

dvs för alla p, $1 \leq p \leq 1$,

$$f(x +_p y) = f(x) \otimes_p f(y).$$

Av detta följer att om $\langle A, \geq, \theta\rangle$ kan homomorft representeras med $\langle Re^*, \geq, \omega\rangle$ så kan $\langle A, \geq, \theta\rangle$ också homomorft representeras med $\langle Re^*, \geq, \mu\rangle$. Ty om u är en homomorfi från $\langle A, \geq, \theta\rangle$ till $\langle Re^*, \geq, \omega\rangle$, så är $f^{\circ}u$ (kompositionen av u och f) en homomorfi från $\langle A, \geq, \theta\rangle$ till $\langle Re^*, \geq, \omega\rangle$. Detta kan inses på följande sätt:

$$a \geq b \text{ omm } u(a) \geq u(b) \text{ omm } \sqrt{u(a)} \geq \sqrt{u(b)} \text{ omm } (f^{\circ}u)(a) \geq (f^{\circ}u)(b).$$

$$(f^{\circ}u)(apb) = f(u(a) +_p u(b)) = (f(u(a)) \otimes_p (f(u(b))$$
$$= \mu((f^{\circ}u)(a),p,(f^{\circ}u)(b)).$$

θ representeras alltså då med μ, vilket innebär att den princip som nu svarar mot principen om förväntad nytta är följande:

$$h(\theta(a,p,b)) = \mu(h(a),p,h(b)) = \sqrt{(ph(a)^2+(1-p)h(b)^2)}$$

där h är en homomorfi från $\langle A, \geq, \theta \rangle$ till $\langle Re^*, \geq, \mu \rangle$. Principen säger alltså att nyttan hos en blandning (ett lotteri) är kvadratroten ur det förväntade värdet av nyttan i kvadrat hos blandningens (lotteriets) utfall.

Av det som sagts följer att om $\langle A, \geq, \theta \rangle$ är ett icke-negativt N-M-blandningsrum, så kan $\langle A, \geq, \theta \rangle$ homomorft representeras med $\langle Re^*, \geq, \omega \rangle$. Vidare gäller att om h och h′ är homomorfier från $\langle A, \geq, \theta \rangle$ till $\langle Re^*, \geq, \omega \rangle$, så finns reella tal r och s sådana att

$$h' = rh \otimes s = \sqrt{(rh)^2 + s^2}.$$

En annan numerisk struktur som kan användas för att representera N-M-blandningsrum och som vi ska återkomma till är $\langle Re^*, \geq, \nu \rangle$, där

$$\nu : Re^* \times [0,1] \times Re^* \rightarrow Re^*$$

och

$$\nu(x,p,y) = p^2x + (1-p)^2y + 2p(1-p)\sqrt{(xy)}.$$

Vi kan, i analogi med vad vi gjort tidigare, se ν som en familj av binära operationer på Re*. Det är lätt att förvissa sig om att g: $Re^* \rightarrow Re^*$ sådan att $g(x)=x^2$ är en isomorfi från $\langle Re^*, \geq, \omega \rangle$ till $\langle Re^*, \geq, \nu \rangle$. Det innebär att om u är en homomorfi från $\langle A, \geq, \theta \rangle$ till $\langle Re^*, \geq, \omega \rangle$ så är u^2 en homomorfi från $\langle A, \geq, \theta \rangle$ till $\langle Re^*, \geq, \nu \rangle$. Notera att

$$\nu(x,p,y) = (p\sqrt{x} + (1-p)\sqrt{y})^2.$$

Vi kommer att använda oss av detta längre fram.

4. Förväntad nytta som ett specialfall av nyttoreduktion

Med ett *nyttomått* för ett blandningsrum $\langle A, \geq, \theta \rangle$ avser vi här en homomorfi h från $\langle A, \geq, \theta \rangle$ till en struktur $\langle B, \geq, \chi \rangle$, där $B \subseteq Re$ och χ en familj av binära operationer på B. Således gäller för h att

$$h(\theta(a,p,b)) = \chi(h(a),p,h(b)).$$

Ett annat sätt att uttrycka detta är att säga att nyttan hos en blandning är en funktion av utfallens nytta och sannolikhet. Detta är alltså en generell princip gällande nyttomått för blandningsrum. Man skulle kanske kunna kalla denna för en generell *nyttoreduktionsprincip;* nyttan hos blandningen reduceras till en funktion av nyttan och sannolikheten hos utfallen. I det specialfall då nyttomåttet representerar blandningsrummet med $\langle R, \geq, \omega \rangle$ där $R \subseteq Re$, så får nyttoreduktionsprin-

cipen den mer specifika formen av principen om förväntad nytta. Således gäller att principen om förväntad nytta är den form nytto-reduktionsprincipen antar då blandningsrummet representeras på det vanliga sättet.

Låt mig försöka uttrycka denna tanke något mer precist. För den skulle behövs några definitioner.

Antag att $\langle A,\geq,\theta\rangle$ är ett blandningsrum och att $f:A \rightarrow Re$ är en homomorfi från $\langle A, \geq\rangle$ till $\langle Re,\geq\rangle$. Vi säger då att:
(1) f uppfyller *förväntade-nyttovillkoret* om för alla $a,b\in A$ och $p\in[0,1]$

$$f(\theta(a,p,b)) = \omega(f(a),p,f(b)).$$

(2) f uppfyller *nyttoreduktionsvillkoret* om det finns en funktion

$$\chi : f[A] \times [0,1] \times f[A] \rightarrow Re$$

sådan att för alla $a,b\in A$ och $p\in[0,1]$

$$f(\theta(a,p,b)) = \chi(f(a),p,f(b)).$$

Vi kan nu uttrycka tankegången ovan på följande sätt: Varje nyttomått för ett N-M-blandningsrum uppfyller nyttoreduktionsvillkoret men inte förväntade-nyttovillkoret. Däremot uppfyller varje nyttomått som representerar ett N-M-blandningsrum med $\langle R,\geq,\omega\rangle$ där $R\subseteq Re$ förväntade-nyttovillkoret.

von Neumanns och Morgensterns axiomatisering av nyttoteorin görs i s k kvalitativa termer, dvs i termer av relationer och operationer på blandningarna. Axiomatiseringen görs således inte i termer av nyttomåttet, vilket i stället först kommer in i representationsteoremet för teorin. Det ligger i karaktären hos denna typ av axiomatisering att olika numeriska strukturen kan användas för representationen av blandningarna. Eftersom förväntad-nyttooperationen sammanhänger med en viss typ av representationen kommer alla nyttomått inte att uppfylla förväntade-nyttovillkoret. Detta är ingen defekt hos von Neumanns och Morgensterns teori utan en följd av den kvalitativa axiomatiseringens natur.

5. Axiomatisk grundval för förväntade-nyttoprincipen

Tanken bakom strävan att ge en axiomatisk grundval för förväntade-nyttoprincipen tycks vara följande. Ett axiomsystem för nyttoteorin utgör ett formellt stöd för förväntade-nyttoprincipen om axiomen

garanterar nyttomått som uppfyller förväntade-nyttovillkoret. Det är viktigt här att skilja på en stark och en svag variant av denna tanke. Låt oss säga att ett axiomsystem ger ett *starkt formellt stöd* för nyttoprincipen om varje nyttomått för ett godtyckligt blandningsrum som satisfierar axiomen uppfyller förväntade-nyttovillkoret. Däremot ger ett axiomsystem ett *svagt formellt stöd* för nyttoprincipen om det för varje blandningsrum som satisfierar axiomen finns ett nyttomått som uppfyller förväntade-nyttovillkoret.

Det som tycks intressera Malmnäs är om axiomatisering kan ge ett starkt formellt stöd för förväntade-nyttoprincipen. Enligt Malmnäs gäller nämligen att en agent uppfyller förväntade-nyttoprincipen om hans eller hennes nyttomått uppfyller förväntade-nyttovillkoret. (Malmnäs (1991) sid 21.) Vidare tycks han tänka sig att en axiomatisering av nyttoteorin ger ett formellt stöd för förväntade-nyttoprincipen om följande gäller: Om en agents preferenser uppfyller axiomen så uppfyller agenten förväntade-nyttoprincipen.

En av slutsatserna i Malmnäs (1991) och (1994) är att von Neumanns och Morgensterns axiomatisering lika lite som någon annan axiomatisering i kvalitativa termer kan ge ett formellt stöd för förväntade-nyttoprincipen. Av vad som tidigare har sagts i denna not framgår att jag helt instämmer i denna slutsats. Vi kan alltid välja att representera ett N-M-blandningsrum med en numerisk struktur sådan att homomorfierna från blandningsrummet till denna struktur inte uppfyller förväntade-nyttovillkoret.

Men av att von Neumanns och Morgensterns axiomatisering av nyttoteorin inte ger ett starkt stöd för förväntade-nyttoprincipen följer inte att den inte ger något stöd alls. Tvärt om ger von Neumanns och Morgensterns teori ett svagt formellt stöd i ovan angivna mening. Om blandningsrummet uppfyller villkoren för N-M-blandningsrum, kan vi representera det med en struktur, så att alla homomorfier från blandningsrummet till denna struktur uppfyller förväntade-nyttovillkoret.

Att en kvalitativ axiomatisering av nyttoteorin enbart kan ge ett svagt formellt stöd för förväntade-nyttoprincipen uppmärksammades tidigt av vissa beslutsteoretiker. Redan i Arrow (1951) påpekas att istället för att använda den nyttoskala som von Neumann och Morgenstern hade funnit, dvs ett nyttomått från ett N-M-blandningsrum till $\langle Re^*, \geq, \omega \rangle$, så kan man använda kvadraten på detta nyttomått. För, som Arrow framhåller, en individs beteende beskrivs då genom att säga att

individen strävar att maximera det förväntade värdet av kvadratroten ur hans nytta. Arrow tänker sig alltså att representera blandningsrummet med $\langle Re^*,\geq,v\rangle$. Om u är en homomorfi från $\langle A,\geq,\theta\rangle$ till $\langle Re^*, \geq,\omega\rangle$ så är ju u^2 en homomorfi från $\langle A,\geq,\theta\rangle$ till $\langle Re^*,\geq,v\rangle$ (se sektion 3). Låt h vara en homomorfi från $\langle A,\geq,\theta\rangle$ till $\langle Re^*,\geq,v\rangle$. I sektion 3 konstaterades att

$$v(x,p,y) = (p\sqrt{x}+(1-p)\sqrt{y})^2.$$

vilket innebär att

$$h(a,p,b) = (p\sqrt{h(a)}+(1-p)\sqrt{h(b)})^2.$$

Att söka de a och b för vilka h(a,p,b) antar maximalt värde är alltså detsamma som att söka de a och b för vilka $p\sqrt{h(a)}+(1-p)\sqrt{h(b)}$ antar maximalt värde, och detta är precis vad Arrow noterar. Arrow understryker att detta påpekande inte ifrågasätter användbarheten av von Neumanns och Morgensterns resultat. Vad det säger är i stället att bland alla olika nyttomått kan man plocka ut en delmängd bestående av nyttomått vilka är linjärtransformationer av varandra och "which has the property of stating the laws of rational behavior in a particularly convenient way" (Arrow (1963) s 10). (I detta sammanhang är också s 98 i Savage (1972) av intresse. Se också Weymark (1991) s 264.)

Det svaga formella stöd som t ex von Neumanns och Morgensterns axiomatisering av nyttoteorin ger förväntade-nyttoprincipen är således av stor betydelse; det gör det möjligt att för N-M-blandningsrum hitta ett nyttomått som uppfyller förväntade-nyttovillkoret. Ett dylikt nyttomått är i många sammanhang bekvämt att använda.

Vid sidan av en axiomatisk grundval för förväntade-nyttoprincipen åberopas ofta stora talens lag som stöd för principen. Skulle kanske stora talens lag kunna ge det starka formella stöd för principen som axiomsystem inte kan ge? Jag förmodar att svaret är nej, ty stora talens lag tycks övergår i andra principer då den numeriska representationen ändras. Men det är en annan historia.

Not

Jag vill tacka Sven Danielsson, Lars Lindahl, Sten Lindström, Per-Erik Malmnäs och Wlodek Rabinowicz för värdefulla synpunkter och

intressanta diskussioner. Som den uppmärksamme läsaren redan har förstått används symbolerna $\geq$, $>$ och $=$ i denna uppsats med två olika innebörder, beroende på om de gäller sannolikhetsblandningar (spel) eller reella tal.

Litteratur

ARROW, K *Social choice and individual values.* 2:a uppl, New York: Wiley, 1963. (1:a uppl 1951.)

MALMNÄS, P-E "Axiomatiska argument för utilitetsprincipen. En formell undersökning". I Rabinowicz, W (red): *Valets vedermödor. Sex beslutsteoretiska studier.* Thales, 1991.

MALMNÄS, P-E "Axiomatic justifications of the utility principle: a formal investigation". *Synthese* 99: 233–249, 1994.

VON NEUMANN, J & MORGENSTERN, O *Theory of games and economic behavior.* 2:a uppl, Princeton: Princeton University Press, 1947.

ROBERTS, F *Measurement theory with applications to decisionmaking, utility and the social sciences.* Reading: Addison-Wesley, 1979.

SAVAGE, L *The foundations of statistics.* 2:a uppl, New York: Dover, 1972. (1:a uppl 1954).

WEYMARK, J "A reconsideration of the Harsanyi-Sen debate on utilitarianism". I Elster, J & Roemer, J (red): *Interpersonal comparisons of well-being.* Cambridge: Cambridge University Press, 1991.

WLODEK RABINOWICZ

Om Seidenfelds kritik av sofistikerade brott mot oberoendeaxiomet

Sammanfattning

Värdet hos ett lotteri är enligt oberoendeaxiomet positivt beroende av värden hos de möjliga lotteriutfallen. Om en agents preferenser bryter mot axiomet kan han förväntas vara *dynamiskt inkonsistent* – initiera handlingsplaner som han inte kommer att fullfölja. Då dynamisk inkonsistens ofta är ofördelaktig för agenten, tycks vi därför ha ett pragmatiskt argument mot kränkningar av detta centrala axiom i teori om förväntad nytta. (Se nedan, sektion 1.) Men är argumentet verkligen övertygande? En agent vars preferenser strider mot axiomet kan fortfarande undgå dynamisk inkonsistens genom att i planeringen ta hänsyn till sitt framtida beteende *(sofistikerat beslutsfattande*, se sektion 2 nedan) eller genom att i efterhand anpassa sina preferenser till den valda handlingsplanen *(resolut beslutsfattande,* jfr McClennen 1990, och Machina 1991). Han kan i själva verket kombinera förutseende med en större eller mindre grad av anpassningsförmåga *(klokt beslutsfattande,* jfr Rabinowicz 1995).

Det utmärkande för den sofistikerade ansatsen, som vi här skall koncentrera oss på, är användningen av *baklängesinduktion:* agenten löser sitt dynamiska beslutsproblem bakifrån. Han identifierar de bästa dragen i slutnoderna av sitt beslutsträd och sedan, i förlitan på sin framtida rationalitet (som skall få honom att göra det bästa), använder han denna informationsinput för att identifiera de bästa dragen i de närmast föregående slutnoderna. På detta sätt fortsätter han tills han når trädets början. Endast de handlingsplaner som, efter att ha påbörjats, vid alla senare beslutsnoder legitimeras av baklängesinduktion betraktas som utförbara. Sådana planer leder inte till dynamisk inkonsistens: blev de valda, skulle de slutföras. Bland utförbara planer väljer den sofisti-

kerade agenten en som leder till det bästa förväntade utfallet. Endast sådana planer är acceptabla.

Är den sofistikerade behandlingen av brott mot oberoende en framkomlig väg? Denna fråga behandlas i sektion 3 nedan. Teddy Seidenfeld (1988a) har visat att en sofistikerad agent som bryter mot oberoende kommer att bryta mot *dynamisk substitution:* en princip enligt vilken en handlingsplan förblir acceptabel när vi vid en av beslutsnoderna byter ut ett alternativ som ingår i planen mot ett likvärdigt sådant. Då Seidenfeld betraktar dynamisk substitution som ett koherensvillkor på dynamiskt beslutsfattande drar han slutsatsen att sofistikering inte räcker för att rädda en agent med "beroende" preferenser från inkoherens.

Men varför skall vi betrakta dynamisk substitution som ett koherensvillkor? Att brott mot oberoende är förenade med brott mot dynamisk substitution kan bero på att dessa principer är nära besläktade. Därför kan de anses vara lika problematiska. (Jfr McClennen 1988 och 1990.) I sitt svar på denna invändning försöker Seidenfeld visa att dynamisk substitution inte kan ges upp av en sofistikerad agent. Under antagandet att agentens preferenser förblir oförändrade under beslutsprocessens gång kan det bevisas, hävdar han, att dynamisk substitution logiskt följer ur två villkor som en sådan agent kan antas respektera: (i) "Ordning", enligt vilket acceptabla val från olika mängder av utförbara planer baseras på en och samma underliggande preferensordning (vilken kan antas motsvara värderingen av de olika planernas förväntade utfall); samt (ii) "Dynamisk Utförbarhet", vilket är Seidenfelds beteckning på baklängesinduktion. (Jfr Seidenfeld 1988b.)

En central roll i beviset spelar beslutsproblem i vilka agenten får välja mellan likvärdiga alternativ vid en av beslutnoderna. För att bedöma beviset tvingas vi därför närmare analysera hur baklängesinduktion handskas med sådana situationer. Vad kan agenten anta om sitt framtida beteende i en beslutsnod i vilken flera drag är lika attraktiva? Är en specificerad handlingsplan verkligen utförbar om den går genom en sådan beslutsnod? Nedan skisseras tre olika sätt att handskas med den sortens problem: *uppspaltning, randomisering,* och *liberaliserad utförbarhet.* Inget av dessa visar sig hjälpa Seidenfeld. De första två gör hans bevis ogiltigt, medan det tredje innebär att en agent som bryter mot oberoende saknar anledning att respektera ordningsvillkoret. Seidenfelds inkoherensinvändning tycks därför kunna undvikas.

1. Dynamisk inkonsistens och besvärliga preferenser

Anta att jag gör upp en handlingsplan, bestämmer mig för en sekvens av handlingar, påbörjar sekvensen men sedan avviker från den valda planen. Under sådana omständigheter kan jag sägas vara dynamiskt inkonsistent i mitt beslutsfattande. Dynamisk inkonsistens kan ibland vara välmotiverad. Oförutsedda händelser som inträffar under planens implementering kan beröva mig förmågan att fullfölja handlingsplanen eller ge mig goda skäl att inte fullfölja den. Här vill jag emellertid bortse från sådana fall och enbart betrakta situationer i vilka agenten inte hotas av det oförutsedda. Visserligen behöver han inte vara säker på vad som kommer att hända men han vet vad som *kan* hända och förbereder sig på de olika eventualiteterna i sin planering. Ponera att en av dessa eventualiteter inträffar. Agenten är dynamiskt inkonsistent om han då inte följer sin ursprungliga plan för eventualiteten ifråga. Dynamisk inkonsistens beror ibland på att agenten "fördärvas" under implementeringsprocessen. Han upphör att handla förnuftigt; antingen upphör han att handla i enlighet med sina preferenser, eller också blir själva hans preferenser oförnuftiga. Här vill jag emellertid bortse även från sådana fall. Vi antar att agenten hela tiden handlar i enlighet med sina preferenser och att dessa inte kommer i olag under processens gång. Kan en sådan person råka ut för dynamisk inkonsistens?

Man har ibland hävdat att sådant är möjligt om *strukturen* hos agentens preferenser är tillräckligt "oregerlig" – om hans preferenser inte kan representeras av en förväntadnyttofunktion. Går de i cirkel eller bryter de mot det s k oberoendeaxiomet, kan han konfronteras med beslutsproblem i vilka hans beteende blir dynamiskt inkonsistent. Dynamisk inkonsistens skapar problem för agenten: den blir ibland till hans garanterade nackdel. Det utfall han realiserar genom ett inkonsistent beteende blir sämre, vad som än händer, än ett utfall som han annars kunde ha tillförsäkrat sig. Men i så fall – har man ibland resonerat – tycks vi här ha ett gott pragmatiskt argument mot den preferensstruktur som ger upphov till ett inkonsistent, och därigenom garanterat ofördelaktigt beteende. Preferensstrukturen ifråga är *irrationell* genom att den är "självvederläggande": den får agenten att realisera utfall som är sämre i termer av hans egna preferenser. Genom sådana pragmatiska argument tycks vi kunna etablera olika rationalitetskrav på preferenser.

För att ta ställning till det ovan skisserade resonemanget, låt oss titta

på ett exempel. En av grundstenarna i teorin om förväntad nytta är oberoendeaxiomet. Detta axiom måste vara uppfyllt om agentens preferenser skall kunna representeras genom en förväntadnyttofunktion. För att formulera axiomet, låt mig introducera ett par förkortningar. X, Y,... står för möjliga utfall. "XpY" är ett lotteri som antingen ger X, med sannolikhet p, eller Y, med sannolikhet 1-p. Observera att lotterier på utfall också räknas till möjliga utfall. Det är alltså inget som utesluter att ett utfall X självt är ett lotteri på vissa andra utfall. "X > Y" läses som "X föredras framför Y".

OBEROENDE: XpZ > YpZ om och endast om X > Y. (p > 0)

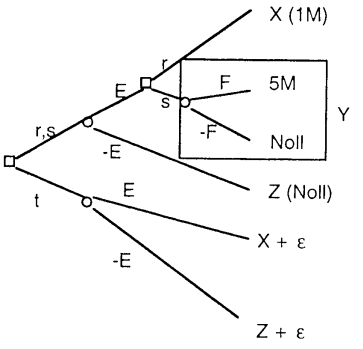
Intuition: Värdet hos ett lotteri är positivt beroende av värden hos de möjliga lotteriutfallen tagna var för sig. I så måtto är värdebidraget från varje utfall oberoende av värdebidraget från det alternativa utfallet. Notera att axiomet är tänkt att gälla alla lotterier, inte bara lotterier på "enkla" utfall, som själva saknar lotterikaraktär. Tillämpades det bara på lotterier på enkla utfall, skulle det vara mycket mindre kontroversiellt, men dess styrka skulle därigenom minska radikalt.

Ett välbekant exempel på preferenser som bryter mot oberoendeaxiomet kommer från den franske ekonomen och nobelpristagaren Maurice Allais. Låt X stå för en miljon dollar "for sure" (1M), och Y för ett lotteri med 10/11 chans till fem miljoner: (5M 10/11 Noll). Många agenter föredrar X framför Y: en riskaversiv agent vill ta det säkra före det osäkra. Låt nu Z vara nollutfallet och låt p vara lika med 0,11. Observera att XpZ = 11% chans till 1M, medan YpZ = 11% chans på 10/11 chans till 5M = 10% chans till 5M. I valet mellan en relativt liten chans till en stor vinst och en obetydligt mindre chans till en betydligt större vinst kan det mycket väl hända att agenten satsar på den större vinsten. Vi har alltså YpZ > XpZ samtidigt som X > Y – ett brott mot oberoendeaxiomet.

Låt oss nu se efter hur en agent med sådana "beroende" preferenser kan fås att handla på ett dynamiskt inkonsistent sätt. Låt E vara en slumphändelse sådan att P(E) = 0,11. Vi tänker oss att E:s sannolikhet är välkänd för agenten och helt oberoende av vad agenten gör. Vi antar dessutom att E är en händelse som i sig är helt likgiltig – en "neutral" händelse, som "krona" eller "klave". Låt F vara en annan sådan neutral slumphändelse, oberoende av E, och anta att P(F) =

10/11. Anta nu att agentens dynamiska beslutsproblem kan beskrivas mha följande beslutsträd (jfr McClennen (1990), figur 1.1):

FIGUR 1



P(E) = p = 0.11

P(F) = 10/11

Förklaring: Kvadraterna står för beslutsnoder – tillfällen då agenten gör sitt drag. Cirklarna representerar chansnoder, där det är "naturen" som gör sitt val. Till exempel, efter agentens drag i första noden, väljer naturen om E eller icke-E blir fallet. Naturens val är, som vi har påpekat, helt oberoende av vad agenten gör. Olika kombinationer av agentens och naturens drag – olika grenar i trädet – leder till olika utfall för agenten, angivna i slutet av varje gren. ε står för en liten extra belöning, säg 10 dollar, som utbetalas till agenten om han går ner i första draget.

En *plan* talar om hur agenten skall agera vid olika beslutstillfällen. Närmare sagt: den specificerar agentens drag vid alla de beslutsnoder som agenten kan nå, eventuellt med naturens hjälp, genom att utföra de tidigare dragen i planen. Exempel: plan r säger att agenten skall först gå upp, och sedan, om E inträffar, gå upp igen. s föreskriver att först gå upp, och sedan, om E inträffar, gå ner. t föreskriver slutligen att agenten går ner i första draget.

Med varje plan kan vi associera det utfall som planens utförande förväntas leda till. När planen passerar chansnoder blir dess förväntade utfall ett lotteri. I vårt exempel har vi följande förväntade utfall för de olika planerna: r ger XpZ, s ger YpZ, och t ger (XpZ)+ε (dvs (X+ε)p (Z+ε)). Vi vet att agenten föredrar YpZ framför XpZ. Om den extra belöning ε som läggs till XpZ inte är för stor, kommer den inte att påverka riktningen hos agentens preferens. Vi kan alltså anta att YpZ > (XpZ)+ε > XpZ. Vi tänker oss nu att agenten väljer den plan som

leder till det bästa förväntade utfallet: s. Han går därför upp i den första beslutsnoden, med avsikt att gå ner om E inträffar. Men, om och när E inträffar, ställs agenten inför ett val mellan X och Y, och bland dessa alternativ föredrar han ju X! I stället för att fullfölja den ursprungliga planen och gå ner väljer han därför att gå upp. Denna dynamiska inkonsistens innebär att agenten får X om E, och Z om icke-E. Vad han får är alltså garanterat sämre än det han kunde tillförsäkrat sig om han gick ner i första draget. Då skulle han ha fått precis lika mycket, X om E och Z om icke-E, plus den extra belöningen! Vi tycks alltså ha konstruerat ett pragmatiskt argument för oberoendeaxiomet. Preferenser som strider mot detta axiom ger upphov till ett dynamiskt inkonsistent beteende som ibland är till garanterad nackdel för agenten.

2. *Den sofistikerade ansatsen*

Men är detta pragmatiska argument korrekt? Många har påpekat att argumentet ifråga lätt kan tillbakavisas. En förutseende agent vars preferenser bryter mot oberoendeaxiomet skulle inte bete sig på ett dynamiskt inkonsistent sätt: Han skulle *inse* att han kommer att gå upp i andra draget om han går upp i första draget och naturen väljer E. Han skulle med andra ord inse att planen s inte kommer att slutföras. Därför skulle han istället gå ner i första draget och lägga beslag på den extra belöningen. Detta sätt att resonera i dynamiska sammanhang brukar kallas för den *sofistikerade* ansatsen. (Beteckningen myntades av Hammond 1976. Jfr också Hammond 1988.) Det utmärkande för ansatsen ifråga är användning av *baklängesinduktion*. Den sofistikerade agenten försöker lösa sitt beslutsproblem "bakifrån": inte från början till slutet, utan tvärtom. Först identifierar han det bästa draget (eller dragen) i den sista beslutsnoden på varje gren i trädet. Han hyser tilltro till sin framtida rationalitet och därför utgår han ifrån att han kommer att göra sitt bästa. Han utgår dessutom från att han kommer att bevara denna tilltro även i fortsättningen och därför tror han sig veta vad som blir hans förväntan om det sista draget när han hunnit till den näst sista beslutsnoden (på varje gren). Sedan använder han sig av detta informationsinput för att hitta det bästa draget (de bästa dragen) i den näst sista beslutsnoden, och så vidare. En handlingsplan är *utförbar* ur agentens synpunkt om han kan förvänta sig att den skulle slutföras om den sattes igång. Följaktligen är den utförbar om dess

första drag kan utföras och om varje efterföljande drag kan visas vara bäst mha baklängesinduktion. I vårt exempel är det därför endast r och t som utgör utförbara handlingsplaner, medan s inte är utförbar. En utförbar plan är *acceptabel* om dess förväntade utfall är bäst, i jämförelse med de andra utförbara planerna. Plan t är därför den enda acceptabla i vårt exempel.

För framtida bruk bör vi säga något om hur fortsatt baklängesinduktion går till när flera drag visar sig vara lika bra i en beslutsnod n. Vad som då kan göras är att man "*spaltar upp*" beslutsproblemet i ett antal versioner, en för varje sådant drag, där agenten i varje version förutsätter att han kommer att utföra det motsvarande draget om och när han når noden n. Sedan kan han använda denna förutsättning som input för att hitta det bästa draget (eller dragen) för denna version i den omedelbart föregående beslutsnoden. En handlingsplan är då utförbar om dess första drag kan utföras och om varje efterföljande drag som ingår i planen kan visas vara bäst i varje version i vilken agenten följer planen så länge det är tillåtet genom baklängesinduktion. Som förut är en utförbar plan acceptabel om den utlovar ett optimalt förväntat utfall i jämförelse med de andra utförbara planerna.

Tanken bakom ovanstående definition av utförbarhet är att agenten kan antas följa den igångsatta planen så länge han saknar anledning att avvika från den – så länge de fortsatta drag som planen föreskriver inte är sämre än de alternativa drag som står till hans förfogande. Anser man däremot att en avvikelse från planen inte kan uteslutas när det avvikande draget är lika attraktivt som det av planen föreskrivna, blir uppspaltning olämplig. Istället kan man då överväga en *randomiserande* tolkning av baklängesinduktion. Enligt denna bör agenten i ett baklängesresonemang behandla sitt beteende i en framtida beslutsnod med likvärda drag som en slumpvariabel, i så måtto att han ex ante betraktar de olika likvärda dragen som lika sannolika. Vad denna randomiserande lösning innebär för planernas utförbarhet kommer att diskuteras senare.

De som förfäktar den sofistikerade ansatsen beskyller den dynamiskt inkonsistente agenten för *kortsynthet*. En sådan agent tar inte hänsyn till sitt framtida beteende och därför betraktar alla teoretiskt möjliga planer som utförbara. Det är detta misstag som får honom att handla inkonsistent: satsa på den bästa bland teoretiskt möjliga planer, men sedan så småningom avvika. Ibland uttrycker man denna tanke genom

att säga att den kortsiktiga ansatsens grundläggande misstag består i att den behandlar *extensiva* beslutsproblem (dvs beslutsproblem i vilka agenten väljer vid mer än ett tillfälle) som om de vore beslutsproblem i s k *normalform*, i vilka agenten bara gör ett val: han väljer handlingsplanen och sedan rent mekaniskt utför planen ifråga (eller kanske överlämnar utförandet till en trogen medhjälpare eller en robot). Vore vårt exempel av den senare typen, så skulle ju planen s vara utförbar. Det skulle inte finnas något tillfälle till dynamisk inkonsistens i och med att agenten bara träffar ett val – vid själva utgångspunkten. Den sofistikerade ansatsen förkastar med andra ord följande antagande som spökar i den kortsiktiga ansatsen:

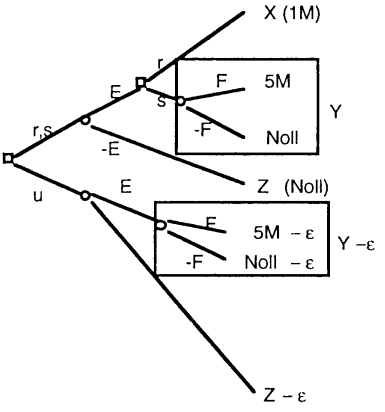
REDUKTION TILL NORMALFORM: En handlingsplan är acceptabel i ett extensivt beslutsproblem om och endast den är acceptabel i ett motsvarande beslutsproblem i normalform.

3. Sofistikerade svårigheter

Den sofistikerade ansatsen är emellertid inte oproblematisk. En svårighet med denna approach som här bara kan kort omnämnas har att göra med frågan om baklängesinduktionens rimlighet. Denna är som bekant ett kontroversiellt förfarande, på grund av sin tillsynes självupphävande karaktär. Samtidigt som den gör anspråk på att vara en rationell beslutsprocedur, förutsätter den att agenten inte bara förlitar sig på sin framtida rationalitet utan dessutom förväntar sig att bevara denna förlitan överallt i beslutsträdet, alltså även vid de framtida beslutsnoder som endast kan nås genom en serie drag som inte legitimeras av baklängesinduktionen. En sådan envishet i självtillit tycks inte vara särskilt förnuftig. (För en diskussion av denna fråga, se t ex Pettit och Sugden (1989) eller Sliwinski (1995).) Denna svårighet aktualiseras emellertid först i samband med beslutsträd vars grenar är tillräckligt långa, något som inte är fallet i de exempel vi här är intresserade av. Ett i sammanhanget allvarligare problem är däremot att sofistikering inte tycks skydda agenter med "besvärliga" preferenser mot all pragmatisk kritik. I synnerhet kan man visa att sofistikerade agenter med preferenser som strider mot oberoendeaxiomet ibland väljer att (1) satsa på strikt dominerade handlingsplaner (dvs planer för vilka det finns alternativ som leder till garanterat sämre utfall än något av alternativen), (2) slippa information, (3) slippa framtida valfrihet. Här skall jag bara visa att (1) och (3) gäller. (Vad beträffar (2), se

Wakker (1988), Machina (1991), sektion 3.3, och Rabinowicz 1995. Att (1) föreligger har bl a påpekats av McClennen (1990), sektion 11.3.) Betrakta följande beslutsproblem:

FIGUR 2



P(E) = p = 0.11

P(F) = 10/11

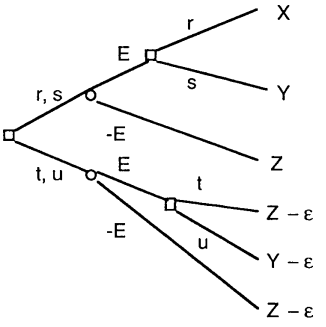
Detta beslutsproblem liknar det förra, med ett undantag. Genom att gå ner i första draget (plan u) tillförsäkrar sig agenten, till en liten kostnad ε, precis samma utsikter som de som utlovas av den attraktiva planen s. Vi antar att det förväntade utfallet av u är bättre än det förväntade utfallet av r: (YpZ)-ε > XpZ. Det är lätt att se att den sofistikerade agenten kommer att gå ner – av samma skäl av vilka han gick ner i det föregående problemet. Men skillnaden mellan att gå ner och gå upp består ju bara i att man berövar sig framtida valfrihet (att välja X snarare än Y om E skulle inträffa) och dessutom betalar (ε) för denna frihetsbegränsning! (3) gäller alltså och (1) gäller likaså ty plan u är strikt dominerad av s: om agenten följde s skulle han, hur det än går, få precis samma slutliga utfall (fem miljoner eller noll, beroende på om E&F inträffar) utan den extra kostnaden ε.

Hur skall nu den sofistikerade agenten förhålla sig till dessa observationer? Vad beträffar (2) och (3) måste hans försvarslinje vara att det är en fördom att tro att valfrihet och information aldrig kan vara skadliga för en förnuftig person. Att (1) gäller kommer han förmodligen att ta med ro. Det är riktigt, kan han påpeka, att s strikt dominerar den valda planen u. Vad framtiden än bär i sitt sköte, kommer att ge honom mindre än om han hade valt s i stället. Men – och detta är ett stort men – den attraktiva planen s är helt enkelt inte utförbar! Dominans är av intresse endast när den dominerande planen

inte blott är teoretiskt möjlig utan dessutom kan utföras. Det finns ingen anledning att ångra att man har underlåtit att satsa på en plan som inte var utförbar.

Det finns emellertid andra, potentiellt mera bekymmersamma fall, i vilka den sofistikerade agenten med "beroende" preferenser tycks handla på ett sätt som han oundvikligen kommer att ångra. Ett sådant exempel har konstruerats av Teddy Seidenfeld (1988a). Nedanstående beslutsproblem skiljer sig från Seidenfelds exempel men kan användas för att illustrera samma problem:

FIGUR 3



P(E) = p = 0.11

X, Y och Z tolkas som förut, med X som föredras framför Y och YpZ som föredras framför XpZ. ε är igen en liten summa som inte påverkar riktningen hos agentens preferenser. Detta innebär att YpZ > (YpZ) – ε > XpZ > Z – ε. Givet dessa antaganden, finns det bara två utförbara planer för den sofistikerade agenten: r och u. Den attraktiva planen s som utlovar det bästa förväntade utfallet (YpZ) är inte utförbar, som i våra tidigare exempel. Detsamma gäller plan t, med det sämsta förväntade utfallet. I jämförelsen mellan r och u har plan u har ett bättre förväntat utfall och därför rekommenderas den av den sofistikerade ansatsen: (YpZ)– ε > XpZ.

Den rekommenderade planen u återigen domineras av s, men s är ju inte utförbar, som vi vet. Allvarligare är det kanske att en agent som väljer u och därför går ner i första draget oundvikligen kommer på ett visst stadium att ångra att han inte hade gått upp istället. Om E inte inträffar blir agentens lott Z – ε, medan han skulle ha fått Z om han hade gått upp. Och om E skulle inträffa, blir agentens *bägge* alternativ, Y – ε och Z – ε, sämre än alternativen Y och X som han skulle ha fått

om han hade gått upp istället. Detta är minst sagt besvärande (jfr Seidenfeld (1988a), s 278), men kanske ändå något man kan leva med. Försvararen av den sofistikerade ansatsen till brott mot oberoende skulle kunna påpeka att agentens oundvikliga ånger inte behöver vara varaktig. (Jfr McClennen (1990), s 290.) Om han går ner och E inträffar, kommer ha ju att satsa på Y – ε. Han kan då sedan lyckas inhösta fem miljoner (minus e) i vinst på detta lotteri. Men om han hade gått upp, då skulle han – som han väl vet – aldrig ha fått lika mycket: han skulle då ha satsat på en säker miljon (X).

Seidenfelds kritik av sofistikerade brott mot oberoende är i själva verket inte alls baserad på dylika invändningar. Felet med den sofistikerade agenten som bryter mot oberoendeaxiomet ligger annanstans. En sådan agent kommer nämligen att bryta mot följande koherensvillkor på dynamiskt beslutsfattande:

DYNAMISK SUBSTITUTION: En plans acceptabilitet påverkas inte genom att de drag som ingår i planen vid några av beslutsnoderna byts ut mot likvärdiga drag.

Anta att vi i figur 3 byter ut Y-ε i beslutsträdets nedre del mot dess "penningekvivalent" $Y-ε. $Y är en penningssumma som väljs på ett sådant sätt att agenten är likgiltig mellan den säkra utsikten att få $Y-ε och lotteriet Y-ε. Det är klart att för vår agent, som föredrar X framför Y, blir $Y en betydligt lägre summa än en miljon. Den kan t ex ligga i närheten av sex eller sju hundra tusen dollar. Efter detta likvärdiga utbyte är det fortfarande samma två planer som är utförbara: r och u. Men plan u i vilket utbytet har gjorts är inte längre acceptabel: dess nya förväntade utfall, ($YpZ)-ε, är ju klart sämre än XpZ, vilket är det förväntade utfallet för r. Seidenfeld skyller inte detta brott mot Dynamisk Substitution på den sofistikerade agentens användning av baklängesinduktion. Skulden läggs istället på den bristande respekten för oberoendeaxiomet. Den sofistikerade agenten blir inkoherent i sitt dynamiska beslutsfattande *om* han bryter mot Oberoende.

Men vad finns det egentligen för skäl att betrakta Dynamisk Substitution som ett koherensvillkor på beslutsfattande? Man skulle t o m kunna hävda att ett sådant villkor blir "question-begging" när oberoendeaxiomets vara eller inte vara står på spel. Dynamisk Substitution är mycket lik den gängse substitutionsprincipen (SUB) enligt vilken en plans acceptabilitet inte påverkas genom likvärdiga utbyten vid några

av de *chans*noder som planen passerar igenom. Men denna princip skulle ju förkastas av en person som bryter mot Oberoende. En sådan person skulle ju t ex välja ett lotteri YpZ framför $YpZ, även om det senare lotteriet fås ur det förra genom att ett av lotteriutfallen, Y, byts ut mot ett likvärdigt utfall, $Y. Men om en plans acceptabilitet inte bevaras under likvärdiga utbyten vid chansnoderna, varför skulle den då behöva bevaras under likvärdiga utbyten vid beslutsnoderna? (För denna invändning, se McClennen 1988, och 1990, sektion 10.5.)

Som svar på denna invändning har Seidenfeld (1988b) argumenterat att Dynamisk Substitution inte är något som en sofistikerad agent kan behandla lättvindigt. Om en sådan agent har "stabila värderingar", dvs om hans preferenser för olika alternativ inte ändras under processens gång, kan man visa, menar Seidenfeld, att Dynamisk Substitution logiskt följer ur två villkor som en sofistikerad agent antas uppfylla: (i) "Ordning", enligt vilket agentens val av acceptabla planer ur olika tillgängliga planmängder baserar sig på en och samma underliggande preferensordning över planerna (en ordning som motsvarar preferensordningen över deras förväntade utfall); samt (ii) "Dynamisk Utförbarhet", vilket är Seidenfelds beteckning på den princip som kopplar en plans utförbarhet till dess legitimering genom baklängesinduktion. Att dessa villkor räcker för Dynamisk Substitution försöker Seidenfeld bevisa som följer:

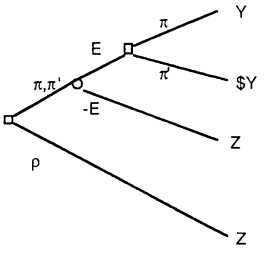
> Let o and o′ be indifferent options. (Stability of values insures that this relation is well defined over choice nodes.) Consider two sequential plans π and π' that differ solely in that, at one designated choice node, π selects o and π' selects o′. The two plans are identical otherwise. First, we establish that π and π' are indifferent, as follows.
>
> Imagine a simple sequential decision that has an initial choice between π/π' plans and another option ρ which is dispreferred to each of π and π'. Since o and o' are indifferent, both are admissible at the designated node where there is a choice between them. Since π and π' are identical otherwise, by Dynamic Feasibility, both plans π and π' are admissible at the initial node, and ρ is inadmissible there. Then, by ordering, π and π' are indifferent; both are admissible in this decision.
>
> Second, since ordering requires that indifference is transitive, it is preserved under iterated substitutions of indifferents at (other) choice

nodes. [Detta innebär att mängden av de planer som fås från varandra genom substitution av indifferenta alternativ vid deras beslutsnoder är inkluderad i en ekvivalensklass m a p indifferensrelationen mellan planer.] ...

Last, by ordering, admissibility is invariant over elements of an equivalence class of indifferent options. Hence, [Dynamic Substitution] obtains." (Seidenfeld (1988b), s 314. Jag har gjort små ändringar i notationen.)

Vad skall vi säga om detta argument? Såvitt jag förstår kan kritik riktas mot dess första steg, i vilket Seidenfeld försöker visa att planerna π och π' kommer bägge att vara acceptabla i ett beslutsproblem där de bägge förekommer, med ett ytterligare plan ρ som det enda, undermåliga alternativet. Att detta steg är ett *non sequitur* kan inses om man betraktar följande exempel:

FIGUR 4



P(E) = p = 0.11

Här är såväl π som π' klart bättre än ρ och de skiljer sig från varandra bara vid en beslutsnod: π föreskriver där valet av lotteriet Y, medan $\pi\rho$ istället innebär att agenten väljer den likvärdiga penningekvivalenten $Y. Det är emellertid lätt att inse att π är den enda planen som är acceptabel i detta problem: dess förväntade utfall, YpZ, föredras ju av agenten framför π':s förväntade utfall: $YpZ. Seidenfelds bevis tycks alltså stranda innan det har börjat.

Observera emellertid att jag här har antagit att agenten följer en plan så länge han inte har någon positiv anledning att avvika från den – ett antagande som ligger bakom uppspaltningsapproachen till likvärdiga drag i baklängesinduktion. (En liknande ansats till baklängesinduktion i dynamiska beslutsproblem med likvärda drag har förresten nyligen utvecklats av Geir Asheim. Jfr hans definition av en "revideringS-

okänslig beslutsregel" –"revision-proof decision rule" – i Asheim (1995).) Det är därför som vi överhuvudtaget kan förutsätta att planerna π och π' är utförbara, dvs att var och en av dem skulle utföras om den valdes av agenten. Om vi ger upp detta antagande om agentens beredskap att följa en påbörjad plan så länge han inte förlorar något på att göra det, måste vi ge en annan analys av exemplet. (Det förefaller som om Seidenfeld själv inte är beredd att acceptera antagandet ifråga. Jfr hans (1988a), s 279, och (1988b), sektion 1.) Vi kan då välja den randomiserande tolkningen som innebär att agenten, innan han nått den andra beslutsnoden, får betrakta som lika sannolikt att han där kommer att gå upp eller ner. Detta innebär att ingen av planerna π och π' kan längre anses vara utförbar. För ingen av dem gäller att agenten kan förvänta sig att den skulle slutföras om den sattes igång. Vad agenten kan bestämma sig för *ex ante* är blott den "disjunktiva" (ofullständiga) planen: att gå upp i första beslutsnoden och sedan, om och när E inträffar, gå upp *eller* ner. Detta innebär emellertid att Seidenfelds bevis igen strandar i första steget. Nu är *ingen* av planerna π och π' acceptabel, eftersom ingen av dem är utförbar (jfr McClennen ss 179f).

Finns det då ingen möjlighet att rädda det första steget i beviset? Hittills har vi utgått ifrån att en plan är utförbar endast om den *skulle* följas om den initierades. Anta att vi försvagar detta krav och nöjer oss med ett svagare villkor *(liberaliserad utförbarhet):* För att en plan skall vara "svagt utförbar" räcker det att den *kanske* skulle följas om den initierades. (Denna möjlighet bör då föreligga för alla drag som naturen kan tänkas göra.) I denna liberaliserade mening är nu såväl π som π' igen (svagt) utförbara, även om vi inte längre kan utgå ifrån att de med säkerhet skulle slutföras om de valdes av agenten. Kommer de emellertid bägge att vara acceptabla? Hittills har vi antagit att en utförbar plan evalueras på basis av sitt förväntade utfall. Men när det gäller en plan som endast är svagt utförbar, för vilken det inte finns någon garanti att den verkligen kommer att slutföras, förefaller det mycket rimligare att evaluera den i termer av något som kan kallas för planens *förväntade valutfall.* Med detta menar jag det utfall som kan förväntas om agenten initierade planen ifråga. I bestämningen av en plans förväntade valutfall tar man hänsyn till möjliga planavvikelser. Om en plan som π initieras och E inträffar, är sannolikheten för att agenten fullföljer den eller avviker lika stor. Följaktligen har π som

sitt förväntade valutfall ett komplext lotteri (Y1/2$Y)pZ (dvs 5% chans till 5M, 5,5% chans till $Y, där $Y är lika med ungefär 0,6 eller 0,7 M, och annars noll). Den alternativa planen π' har givetvis exakt samma förväntade valutfall, och plan ρ som inte ger något alls i utbyte är klart undermålig. Detta innebär att såväl π som π' blir nu acceptabla och det första steget i Seidenfelds bevis blir därigenom giltigt.

Men nu blir *ordningsvillkoret*, som behövs för att resten av beviset skall gå igenom, inte längre plausibelt! Ty det förväntade valutfallet av en svagt utförbar plan beror på vilka andra planer som är tillgängliga i ett givet beslutsproblem. I detta avseende finns det alltså en väsentlig skillnad mellan en plans förväntade utfall och dess förväntade valutfall. Det senare, men inte det förra, är klart kontextberoende. Tänk på vad som händer om vi i vårt exempel tar bort alternativet Y i den andra beslutsnoden, så att plan π inte längre är tillgänglig. Det förväntade valutfallet hos den kvarvarande planen π' blir nu plötsligt mycket sämre: det faller ner till $YpZ (dvs 11% chans till $Y). Detta kontextberoende gör att en acceptabel plan som π' kan upphöra att vara acceptabel om vi tar bort en annan plan, π, från den tillgängliga planmängden. Detta skulle faktiskt inträffa i vårt exempel om vi bara modifierade den lite grann. Det räcker att vi höjer ρ:s utfall från Z till ($YpZ) + ε. Men om vi i detta modifierade exempel istället för π tog bort π', då skulle ju π fortfarande vara acceptabel: dess förväntade valutfall skulle ju till och med höjas till YpZ. Detta visar att substitutionen av en sådan plan för den andra, i ett beslutsproblem där blott en av dem är tillgänglig, kan mycket väl inverka på acceptabiliteten. Dynamisk Substitution bör därför ges upp.

Seidenfelds bevis förefaller alltså inte kunna räddas. För vissa tolkningar av utförbarhet och baklängesinduktion blir beviset ogiltigt i sitt första steg, medan en liberaliserad tolkning av utförbarhet gör att en av bevisets premisser (Ordning) inte längre framstår som rimlig.

Kan vi i så fall dra slutsatsen att den sofistikerade ansatsen är den rätta medicinen mot dynamisk inkonsistens för folk som bryter mot Oberoende eller lider av andra likartade besvär (preferenscykler, etc)? Inte nödvändigtvis. McClennen (1990) har argumenterat att dynamisk inkonsistens bäst botas inte genom förutseende, som utesluter vissa planer som outförbara, utan genom den resoluta ståndaktigheten i ens självpåtagna åtaganden (*den resoluta ansatsen)* – genom en beredskap att i efterhand ändra ens preferenser och anpassa dem till den plan

man har valt. Liknade idéer framförs av Machina (1991). McClennen och Machina kritiserar den sofistikerade ansatsen för dess uteslutande framåtriktade karaktär; den sofistikerade agentens preferenser vid varje beslutsnod är helt opåverkade av det förflutna – helt "separerbara" från det som har varit. Genom att uppge denna separerbarhet kan den resoluta ansatsen hålla kvar vid reduktion till normalform och ändå undvika dynamisk inkonsistens. I Rabinowicz (1995) argumenterar jag för en mera generell approach som ger utrymme åt såväl förutseende som preferensanpassning till valda planer (*den kloka ansatsen*). I denna approach antas varken reduktion till normalform eller separerbarhet som generella villkor på dynamiskt beslutsfattande. Men det är en annan historia.

Not

Denna artikel är baserad på Rabinowicz (1995), sektioner 1 till 3, samt 6, och har presenterats vid Filosofidagarna 1995 i Umeå. Jag vill tacka Sten Lindström och de andra organisatörerna för hjälp och synpunkter.

Litteratur

ASHEIM, GEIR (1995), "Individual and Collective Time-Consistency", Department of Economics, University of Oslo, opublicerat manuskript.

HAMMOND, PETER J (1976), "Changing Tastes and Coherent Dynamic Choice", *Review of Economic Studies 43,* ss 159-73.

HAMMOND, PETER J (1988), "Consequentialist foundations for expected utility", *Theory and Decision 25*, ss 25–78.

MACHINA, MARK (1991), "Dynamic Consistency and Non-expected Utility", in Michael Bacharach and Susan Hurley (eds), *Foundations of Decision Theory – Issues and Advances*, Blackwell, Oxford, ss 39–91.

MCCLENNEN, EDWARD (1988), "Ordering and Independence. A Com ment on Professor Seidenfeld", *Economics and Philosophy 4,* ss 298–308.

MCCLENNEN, EDWARD (1990), *Rationality and Dynamic Choice*, Cambridge University Press, Cambridge.

PETTIT, PHILIP, and ROBERT SUGDEN (1989), "The backward induction paradox", *The Journal of Philosophy 86*, ss 169–82.

RABINOWICZ, WLODZIMIERZ (1995), "To Have One's Cake and Eat It, Too: Sequential Choice and Expected Utility Axioms", *The Journal of Philosophy, 92*, ss 586–620.

SEIDENFELD, TEDDY (1988a), "Decision Theory Without 'Independence' Or Without 'Ordering'", *Economics and Philosophy 4,* ss 267–90.

SEIDENFELD, TEDDY (1988b), "Rejoinder [to Hammond and McClennen]", *Economics and Philosophy 4,* ss 309–15.

SLIWINSKI, RYSIEK (1995), "Balängesinduktion i spelteori", föredrag vid Filosofidagarna i Umeå (denna volym).

RYSIEK SLIWINSKI

Baklänges induktion i spelteori

Man brukar särskilja tre analysnivåer av ett spel: extensiv form, normal form, och karakteristisk form, där ordningen mellan dessa former återspeglar deras växande abstraktionsnivå. Själva spelen brukar man dela i kooperativa och icke-kooperativa.

I centrum av spelteorin ligger ett studium av icke kooperativa spel i extensiv form. Enklaste typen av sådana spel utgörs av spel med fullständig information. Baklänges induktionsteorin är tänkt att utgöra ett universalverktyg för att bestämma hur rationella spelare bör spela dessa spel. I detta korta föredrag kommer jag att diskutera riktigheten hos denna teori.

I spel i extensiv form antas varje spelare i varje skede av spelet dels känna till vilka möjliga drag han kan utföra och vilka möjliga motdrag hans motspelare kan utföra och dels alltid veta i vilket skede av spelet hon befinner sig. Sådana spel kan beskrivas med hjälp av träddiagram där varje väg från trädets rot till en av trädets grentoppar representerar ett möjligt spelförlopp och där ändpunkterna av toppgrenarna är försedda med matriser som anger vilka utbetalningar spelarna erhåller givet det spelförloppet som leder till dessa grenar.

Baklänges induktionsteorin innebär att man arbetar baklänges från trädets toppar till roten av trädet genom att med hjälp av rationalitetsantaganden om spelarna, successivt "klippa av" trädets grenar, dvs eliminera de drag vilka en rationell spelare inte skulle välja.

Teorin löser inte alla spel i extensiv form men den antas ge oss lösningar till alla spel med *perfekt information*. Denna klass av extensiva spel definieras av följande villkor: 1) spelet måste vara ändligt, det måste finnas ändligt antal noder i trädet, 2) spelarna måste utföra sina drag ett i taget, dvs inga samtidiga drag är tillåtna, 3) vid varje skede av spelet måste varje spelare känna till alla tidigare drag som

har utförts dvs han måste veta var någonstans i trädet han befinner sig. Givet dessa villkor kan vi tillämpa baklänges induktionsalgoritmen på följande sätt:

Steg 1
för varje möjligt spelförlopp
a) identifiera den siste spelaren. Att detta är möjligt garanteras av villkoren 1 och 2.
b) identifiera den siste spelarens optimala drag dvs drag som maximerar hans utbetalning givet att han befinner sig vid denna beslutspunkt. (Det är lite problem med indifferenser men det lämnar vi åt sidan.)

Steg 2
a) för varje möjligt spelförlopp, givet den siste spelarens optimala drag, identifiera den näst siste spelarens optimala drag på samma sätt.

Stegen 3 - n
b) fortsätt denna process tills roten av trädet är nådd.

När processen är avslutad har vi erhållit en mängd av strategier, en för varje spelare, som består av alla spelares optimala val vid varje möjligt skede av spelet. Den erhållna mängden uppfyller ett formellt villkor som brukar anses vara ett nödvändigt villkor för att en mängd av strategier (en för varje spelare) skall utgöra en lösning. Detta villkor kallas *Nash equilibrium* och kan formuleras (för tvåpersoners spel) på följande sätt: varje strategi i mängden utgör ett bästa svar på den andra strategin. Dessutom uppfyller mängden en ännu starkare villkor, något som kallas delspelsequilibrium. Men hur argumenterar man för att påvisa att den erhållna mängden utgör en lösning av spelet, dvs att rationella spelare bör välja att utföra de drag som algoritmen föreskriver?

Argumentet är mycket enkelt: Den siste spelaren i varje möjlig spelförlopp kommer p g a sin rationalitet att välja det alternativ som ger henne ett optimalt utfall. Men detta betyder att hennes sista val är fixerad. I så fall kan vi nu istället betrakta något kortare spel där de sista spelarnas sista val är borttagna och ersatta med de fixerade utbetalningsmatriserna. På det sättet kommer de näst sista spelarna i det ursprungliga spelet att vara de sista spelarna i det nya spelet och vi kan upprepa denna elimineringsprocedur tills vi kommer till spelträdets rot. Den beskrivna proceduren kan utföras av varje rationell

spelare som på detta sätt kan resonera sig fram till vilka val hon bör göra.

Observera att argumentet förutsätter att spelarna är rationella och vet en hel del om spelet och om sina motspelare. Förutom de ovan nämnda villkoren på spelet måste följande villkor på spelarna vara uppfyllda:

(4) varje spelare måste ha fullständig kunskap om spelets struktur, dvs veta hur spelets träddiagram ser ut.
(5) varje spelare måste känna till vilka utbetalningar han och de övriga spelare erhåller för varje möjligt spelförlopp.
(6) varje spelare måste vara övertygad om både sin egen och de andra spelarnas rationalitet.
(7) Villkor (3)–(6) ovan måste utgöra vad som ofta kallas "publik kunskap" ("common knowledge") bland spelarna dvs varje spelare måste veta att varje spelare vet att varje spelare vet och så vidare. (Detta sista villkor kan försvagas något. Bicchieri (1989) har visat att antal kunskapsnivåer som krävs för att tillämpa baklängesinduktionsteorin är ändligt).
(8) Villkor (3)–(7) ovan måste vara uppfyllda inte bara i början av spelet utan också vid varje möjligt skede av spelet.

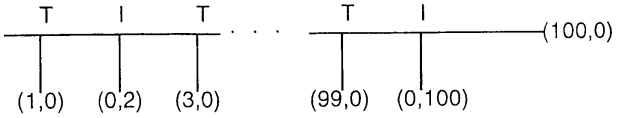
Innan vi fortsätter är det värt att poängtera att dessa extremt starka villkor på spelarna gör att teorins prediktiva styrka i verkliga situationer är låg.

Såsom teorin är nu presenterad kan den tillämpas endast på spel med fullständig information. Men den kan lätt utvidgas till så kallade multi-skede-spel med observerbara val, där simultana drag är tillåtna under förutsättning att det finns dominerande val vid varje skede av spelet. I själva verket kan teorin utvidgas ytterligare så att den omfattar större klass av spel. Vi kommer då till spel som uppfyller så kallad "subgame perfection". Men det finns inte plats att gå in på dessa utvecklingar. I fortsättningen kommer vi att begränsa oss till spel med fullständig information.

Låt oss titta på ett exempel för att se hur teorin fungerar.

Take it or leave it – spelet (Reny 1988). Det finns två spelare, Tristan och Isolde. Dessutom finns det en tävlingsledare som är utrustad med 100 enkronor. Spelet går till på följande sätt. 1: Tävlingsledaren lägger fram en krona på bordet och erbjuder Tristan att ta den.

Om Tristan tar den så avslutas spelet. Om Tristan inte tar den så lägger tävlingsledaren en krona till på bordet och erbjuder bägge mynten till Isolde. Om Isolde tar dem så avslutas spelet. Om inte så läggs det en krona till på bordet och Tristan får chansen att ta tre mynt och avsluta spelet. Om han inte tar den så upprepas proceduren. Spelet avslutas då någon av spelarna tar den summa pengar som erbjuds honom/henne eller om Isolde vägrar att i sista skedet av spelet ta 100 kronor. I så fall tillfaller pengarna Tristan. Spelträdet ser ut på följande sätt:



Låt oss tillämpa baklängesinduktionsteorin på spelet. Vid det sista möjliga valet för Isolde, har hon att välja mellan att ta 100 kronor eller att lämna dem till Tristan. Då det är publik kunskap bland spelarna att Isolde är rationell, kan man med säkerhet förutsäga hennes val, dvs att hon kommer att ta 100 kronor. Låt oss nu titta på Tristan vid näst sista skedet av spelet. Han har att välja mellan att ta 99 kronor eller att låta Isolde bestämma över 100 kronor. Men han vet redan att Isolde givet chansen kommer att ta 100 kronor. Alltså står hans val mellan att ta 99 kronor eller att låta Isolde ta 100 kronor. Vi vet att Tristan är rationell alltså vet vi att han kommer att ta 99 kronor. Vi upprepar proceduren tills vi kommer till det första valet och kommer fram till en unik lösning som består i att Tristan tar första kronan och därmed avslutar spelet.

Tillämpning av baklänges induktionsteorin på vissa spel leder till åtminstone prima facie paradoxala resultat. Jag tänker i första hand på *Itererad fångens dilemma* och på *Kedjebutiksspelet.* Då det första spelet inte är ett spel med fullständig information, tittar vi i stället på det andra spelet.

Kedjebutiksspelet (Selten 1978). Det finns en butikskedja som är representerad i tjugo små städer. I varje stad finns det en potentiell rival som funderar på att öppna en konkurrerande butik (IN-alternativet). Vi antar att de potentiella rivalerna kan öppna sina butiker med en månads intervall dvs

Rivalen i den 1:a staden kan öppna butik i januari 1996.
Rivalen i den 2:a staden kan öppna butik i februari 1996.

.

.

.

Rivalen i den 20:e staden kan öppna butik i augusti 1997.

Vi antar också att varje rival har endast en chans att öppna sin butik. Väljer han att inte öppna (UT-alternativ) så är han ute ur spelet.

När butikskedjan konfronteras med en rival som har valt att öppna sin butik har den två alternativ att välja mellan: den aggressiva (man dumpar priserna i den aktuella staden för att för att få rivalen bort från marknaden) och den kooperativa (man köper ut rivalen). Utbetalningsmatrisen för varje rond av spelet ser ut på följande sätt:

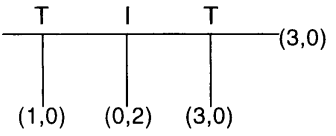
	IN	UT
kooperativ	3,3	5,2
aggressiv	1,0	5,2

Hur ska nu butikskedjan och de potentiella rivalerna handla? Enligt baklänges induktionsteorin ska butikskedjan samarbeta med varje aktuell rival och varje rival ska öppna en butik. Är denna lösning tilltalande? Det förefaller vara rimligt att anta att åtminstone i början av spelet är det bättre att spela aggressivt för att skrämma kommande rivaler från att öppna sina butiker. Vi återkommer till detta spel senare.

Trots sådana paradoxala resultat har baklänges induktionsteorin sedan den formulerats av Kuhn 1953 blivit allmänt accepterad av spelteoretiker. Detta är speciellt anmärkningsvärt då denna eliminationsmetod visar slående likheter med studentens sätt att eliminera möjliga dagar för examination i den välkända så kallade *Oväntade examinations-paradoxen* – en paradox som formulerades på 1940-talet och som sedan dess har livligt diskuterats i den filosofiska litteraturen. De flesta som har diskuterat paradoxen har konstaterat att det är något fel på studentens resonemang fast de har haft olika teorier om vari felet består. Däremot har baklänges induktionsteorin levt en bekymmerslös tillvaro. Men under 1980-talet har antalet kritiker växt och deras gemensamma ansträngningar har lett till att teorin har förlorat mycket av

sin trovärdighet. Vi kommer nu att titta närmare på några av deras argument.

Låt oss återvända till *Take it or leave it - spelet*, men för enkelhetens skull betrakta den i en något kortare version, där vi har endast tre möjliga faser av spelet.



Vi har sett att baklänges induktionsteorin föreskriver att i varje skedet av spelet, spelaren som är vid draget skall ta erbjudna pengar. Men är det en rimlig lösning? Anta att Tristan väljer att inte ta en krona. Hur skall då Isolde handla? Teorin föreskriver att hon skall ta två kronor men enligt samma teori borde Tristan ha tagit en krona och eftersom han inte gjort det måste hennes beslut bero på hur hon förklarar hans avvikande drag. Det förefaller finnas åtminstone fyra möjliga förklaringar:

(1) Tristan har gjort ett oavsiktligt misstag,
(2) Utbetalningar till Tristan är annorlunda än vad Isolde tror de är.
(3) Tristan försöker lura Isolde, han vill få henne att tro att han kommer att vägra att ta 3 kronor, fast i verkligheten kommer han säkert att ta dem.
(4) Tristan är irrationell.

Problemet är att Isoldes förväntningar på hur Tristan kommer att handla i nästa drag beror på vilket av de fyra alternativen hon betraktar som den korrekta förklaringen. Låt oss titta på närmare på de olika alternativen.

Om vi antar att varje avvikelse från det av teorin föreskrivna draget beror på ett "litet" misstag och att sannolikheterna för sådana misstag vid skilda skeden av spelet är oberoende av varandra, kan vi fortsätta att använda baklänges induktion för att förutsäga hur spelet kommer att fortgå. Detta utgör kärnan i Seltens berömda "trembling hand" modell från 1975, som förutsätter att varje gång en spelare vill utföra draget A finns det en mycket liten men positiv sannolikhet att han av misstag utför ett annat oavsiktligt drag B. Detta kan förklaras av spelarnas

skakande händer, så att även om de beslutar sig för att trycka på knappen för handling A, så på grund av dessa handskakningar trycker de på den felaktiga knappen för handling B. Observera att denna teori förutsätter dels att rationalitet och misstag ar förenliga med varandra och att misstag sker helt slumpvis och oberoende av varandra.

Men är sådan "trembling hands" hypotes alltid en rimlig förklaring för varje avvikelse? Om vi är i ett skede av ett "stort" spel, ett skede som man kan komma till endast genom att en viss spelare utför 50 på varandra följande felaktiga drag förefaller det vara fullständigt absurt att betrakta dem som slumpvisa misstag. Borde man inte leta efter någon närmare till hands liggande förklaring?

Låt oss betrakta det andra alternativet. Således antar vi att varje avvikelse kan förklaras genom att hävda att de aktuella utbetalningarna ser annorlunda ut än de vi ursprungligen har tänkt oss. (Se Fudenberg, Kreps och Levine 1988.) Detta alternativs största fördel är att det undviker problemet med att behöva besvara frågan vilka övertygelser en spelare skall ha i ett skede av spelet som han tillskriver sannolikhet 0 att han kommer att hamna i. Frågan om hur en avvikelse skall förklaras formuleras nu i stället som en fråga om vilka alternativa utbetalningar är mest sannolika givet en viss tidigare del av spelförloppet. Observera dock att denna osäkerhet i fråga om utbetalningar bryter mot villkoret att kunskap om utbetalningarna skulle utgöra publik kunskap bland spelarna. Och ifall detta villkor inte är uppfyllt så kan vi inte använda baklänges induktion.

Nu till det tredje alternativet. En avvikelse förklaras som ett avsiktligt försök att lura motspelaren. Att skapa hos henne förhoppningar som inte kommer att infrias. (Se Sorenson 1988.) Med är det möjligt att lura en rationell spelare? Kommer hon inte att genomskåda varje sådant försök? Kommer hon inte att inse att sådana förhoppningar är tomma dvs att de aldrig kommer att infrias? Här kan vi dra en parallell till kedjebutiksspelet. Där var det frågan om hot istället för förhoppningar. Rivalen som överväger om han skall öppna butiken eller ej måste ställa sig frågan om butikskedjans möjlighet att välja ett aggressivt alternativ utgör ett verkligt hot mot honom. Men om baklänges induktionsteorin är korrekt så kan han inte ta dessa hot på allvar. De utgör endast "tomma hot" ty han vet ju (tack vare teorin) att butikskedjan aldrig kommer att verkställa dem. Så vad skall en potentiell rival tro om han får veta att butikskedjan har valt ett

aggressivt spel mot en tidigare rival? Han kan inte gärna tro att butikskedjan genom att spela aggressivt hotar honom ty han vet att alla sådana hot är tomma och att en rationell spelare aldrig ger tomma hot (avslöjade tomma hot utgör ju en kostnad för hotaren). I stället för att ta denna aggressiva handling som hot tar han det som ett bevis för att butikskedjan handlar irrationellt.

Detta leder oss rakt till den fjärde möjliga förklaringen av en avvikelse. Ett drag som avviker från den av teorin föreskrivna visar att spelaren som har utfört draget är irrationell eller inte längre har den absoluta graden av tro på alla spelares rationalitet som teorin kräver. (Se Binmore 1987 och Reny 1985.) Men då bryter man igen mot en av teorins förutsättningar.

Sammanfattningsvis kan man hävda att när en avvikelse har ägt rum så kan inte längre antagandet att rationalitet utgör publik kunskap bland spelarna försvaras. Man kan argumentera på följande sätt: Om spelare A har utfört ett drag som enligt teorin är irrationellt så kan inte nästa spelare fortfarande tro att A både är rationell och har den absoluta graden av tro på alla spelares rationalitet som teorin kräver. Någon av dessa utsagor måste vara falsk. Och kan inte en spelare tro på konjunktionen så kan hon inte heller veta att den är sann då brist på tro implicerar brist på kunskap. Så även om det rådde publik kunskap i början av spelet så kan den inte överleva efter att drag som strider mot teorin har utförts.

Om kritikerna har rätt så måste slutsatsen vara att det finns spel med fullständig information där publik kunskap om rationalitet vid varje möjlig punkt i spelet inte kan upprätthållas. Men publik kunskap var ju ett nödvändigt villkor för att kunna tillämpa teorin. Alltså finns det spel med fullständig information som inte kan lösas med hjälp av teorin. I sådana spel kan en spelare genom att utföra ett drag som enligt teorin är irrationellt göra sina motspelare osäkra på om alla förutsättningar som måste vara uppfyllda för att teorin skall fungera verkligen är uppfyllda. Men om de är inte uppfyllda så kan inte teorin tillämpas och då behöver inte den enligt teorin irrationella draget verkligen vara irrationellt. Om vi inte kan tillämpa baklänges induktionsteorin så lämnas vi i sticket och kaos uppstår tills vi hittar en teori som är mer generell och bättre motiverad.

Hur kan vi då undvika detta hotande kaos? En möjlig utväg tycks vara att ta fasta på att spelarnas föreställningar om varandras rationa-

litet inte för alltid är givna utan att de förändras beroende på hur spelet fortgår. Således kan det vara på sin plats att utvidga teorin med en teori om hur dessa övertygelser förändras. Men är en sådan utvidgning överhuvudtaget möjlig? Teorin ska ju användas för att identifiera rationella strategier. Men för att kunna använda den utvidgade teorin måste vi veta när en övertygelseförändring måste göras, det vill säga, vi måste veta när en avvikelse har ägt rum. Men det betyder ju att vi i förväg måste känna till vilka drag som är rationella. Vi har således hamnat i en *Moment 22*-situation.

En annan möjlig utväg är att förstärka villkoren på spelarnas rationalitet och på deras publika kunskap om varandras rationalitet på sådant sätt att man kräver att samtliga spelare i varje rond av spelet skall ha en *orubblig* tilltro till allas rationalitet och att de dessutom skulle bevara denna tilltro oberoende av vad som tidigare har hänt. Här kommer vi in på Howard Sobels försvar av baklänges induktionsteorin (Sobel 1993). Tyvärr så saknas det utrymme för att redogöra för de mycket sofistikerade argument Sobel presenterar, men det är värt att poängtera (såsom Sobel själv gör) att även om han har rätt så innebär denna förstärkning av de redan mycket starka villkor att teorin har mycket lite att göra med hur spelare i det verkliga livet bör handla. Dessutom kan man hävda att Sobel-spelarnas dogmatiska tilltro till varandras rationalitet kanske i själva verket är ett tecken på deras bristande rationalitet.

Betyder allt detta att baklänges induktionsteorin överhuvudtaget inte kan tillämpas? Nej, det finns två klasser av spel som teorin klarar av. Den första klassen kan beskrivas som klassen av spel för vilka teorin som lösning ger oss en mängd av strategier (en för varje spelare) som passerar genom alla beslutsnoder i spelträdet. Den andra klassen har följande egenskap (jag beskriver endast tvåpersonersspel, men det kan lätt utvidgas till flerpersonersspel): efter att en spelare har utfört ett enligt teorin felaktigt drag, så finns det en strategi för hans motspelare som garanterar henne det optimala utfallet oberoende av hur den felande spelaren beter sig i fortsättningen av spelet. Exempel på sådana spel är olika brädspel där endast tre möjliga utfall finns (vinst, remi, förlust). Det mest kända spelet av den typen är förstås schack. Som en historisk kuriositet kan tilläggas att schack är det första spelet som baklänges induktionsteorin har tillämpats på för att påvisa existensen av en lösning. Detta gjordes redan 1913 av Ernst Zermelo.

Han har visat att det finns en strategi för åtminstone en av spelarna som garanterar henne åtminstone oavgjort resultat. Det kan dessutom vara så att strategin garanterar henne vinst men vi kan inte veta vilka av dessa fall som gäller på grund av spelets enorma komplexitet vad det gäller antalet möjliga drag. Men resultatet visar att om vi kunde rita spelets träddiagram så skulle vi kunna identifiera denna strategi genom att tillämpa baklänges induktionsalgoritmen.

Litteratur

BICCHIERI, C, 1988. "Common Knowledge and Backward Induction: A Solution to the Paradox", i *Proceedings of the Second Conference on Theoretical Aspects of Reasoning About Knowledge*, red av Moshe Vardi, Kaufman, Los Altos, California, ss 381–393.

BICCHIERI, C, 1989. "Self-Refuting Theories of Strategic Interaction: A Paradox of Common Knowledge", *Erkenntnis* 30, 69–85.

BICCHIERI, C, 1992. "Knowledge-dependent Games: Backward Induction" i *Knowledge, Belief, and Strategic Interaction*, red av C. Bicchieri och Dalla Chiara, Cambridge University Press, Cambridge.

BINMORE, K, 1987. "Modeling Rational Players, Part 1", *Economics and Philosophy* 3, 179–214.

KUHN, H, 1953. "Extensive Games and the Problem of Information", I *Contributions to the Theory of Games*, Vol. 2, Princeton, Princeton University Press, 193–216.

NASH, J, 1951. "Non-cooperative games"Extensive Games", *Annals of Mathematics* 54, 286–295.

PETTIT, P och SUDGEN, R, 1989, "The Backward Induction Paradox", *Journal of Philosophy* 86, 169–182.

RENY, P, 1988. "Rationality, Common Knowledge and The Theory of Games", opublicerat manuskript.

SORENSEN, R A, 1988. *Blindspots*, Clarendon Press, Oxford.

SELTEN, R, 1975. "Re-examination of the Perfectness Concept for Equilibrium in Extensive Games", *International Journal of Game-Theory* 4, 22–25.

SELTEN, R, 1978. "Chain Store Paradox", *Theory and Decision* 9, 127–159.
SOBEL, H, 1978. "Backward Induction Arguments in Finitely Iterated Prisoners' Dilemmas: a Paradox Regained", *Philosophy of Science*, 60, 114–133.
ZERMELO, E, 1913. "Über eine Anwendung der Mengenlehre auf die Theorie des Schachspiels", *i Proceedings of the 5th International Congress of Matematicians*, ss 501–504.

Notis

Medarbetare i detta ovanligt tjocka nummer av *Filosofisk tidskrift*: *Thomas Wetterström* är docent i praktisk filosofi i Göteborg, *Paul Needham* är docent i teoretisk filosofi i Stockholm, *Martin Gustafsson* är doktorand i teoretisk filosofi i Uppsala, *Lars Bergström* är professor i praktisk filosofi i Stockholm, *Krister Bykvist* är doktorand i praktisk filosofi i Uppsala, *Erik Carlson* är forskarassistent i praktisk filosofi i Uppsala, *Ingmar Persson* är docent i praktisk filosofi i Lund, *Jan Österberg* är docent i praktisk filosofi i Uppsala, *Peter Pagin* är docent i teoretisk filosofi i Stockholm, *Ingar Brinck* är doktorand i teoretisk filosofi i Lund, *John Cantwell* är doktorand i teoretisk filosofi i Uppsala, *Bertil Rolf* är docent i teoretisk filosofi i Lund, *Jan Odelstad* är docent och forskarassistent i teoretisk filosofi i Stockholm, *Wlodek Rabinowicz* är professor i praktisk filosofi i Lund och *Rysiek Sliwinski* är doktorand i teoretisk filosofi i Uppsala.

Thales *ger även ut en* filosofisk tidskrift *på engelska*

THEORIA
A Swedish journal of Philosophy

Redaktörer: Bengt Hansson, Ingmar Persson och Krister Segerberg
Utkommer med 3 nummer/år
Prenumerationspris: 150:-/år
Prenumeration sker genom insättning av 150:-
på pg 15 01 27-9 Theoria – ange årgång/årgångar
Adress till Theoria: Filosofiska institutionen, Lunds universitet,
Kungshuset Lundagård, 222 22 Lund

1993 parts 1-3 The Philosophy of Sören Halldén

Med bidrag av: SÖREN HALLDÉN, THOMAS ANDERBERG, STEFAN BJÖRKLUND, MATS FURBERG, GÖRAN HERMERÉN, INGMAR PERSSON, ISAAC LEVI, MARTIN EDMAN, JAAKKO HINTIKKA, NILS-ERIC SAHLIN, GEORGE SCHUMM, VIKTOR FINN, REVAZ GRIGOLIA

1994 part 1

Med bidrag av: JORDAN HOWARD SOBEL, KAZIMIERZ SWIRYDOWICZ, ISTIYAQUE HAJI, JAN BERG, OLAV GJELSVIK

1994 part 2

Med bidrag av: DOUGLAS ODEGARD, KAI NIELSEN, MARIA CONCETTA DE MAIO, DAN EGONSSON, FREDRIK STJERNBERG, JOHANNES PERSSON

1994 part 3 The Rolf Schock Prize in Logic and Philosophy 1993
A Special Issue on the Philosophy of W.V. Quine

Med bidrag och intervjuer av: W.V. QUINE och DONALD DAVIDSON

Båda årgångarna är under distribution

Metafysik & mening

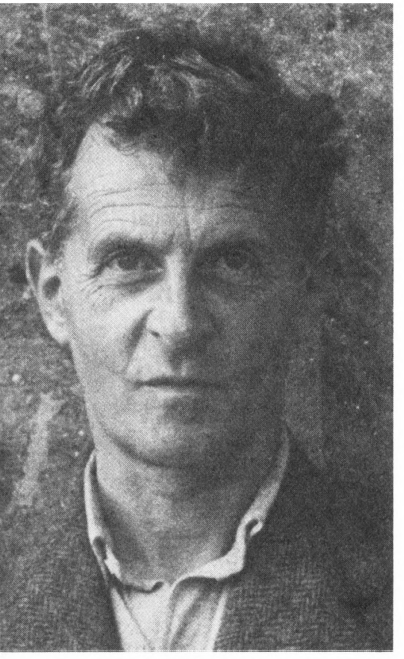
Essäer av Michael Dummett
i översättning av Fredrik Stjernberg

Metafysik och mening är ett representativt urval av Michael Dummetts mycket inflytelserika essäer. Dummett, född 1925, var mellan 1979 och 1992 professor i filosofi vid universitetet i Oxford. I Freges och Wittgensteins efterföljd har han givit talrika bidrag till språkfilosofi och metafysik, vilka hör till de mest diskuterade i samtida filosofi. Hans projekt har framförallt varit att närma sig traditionella metafysiska frågor via meningsteorin.

Michael Dummett har av Kungl. Vetenskapsakademien tilldelats 1995 års Rolf Schockpris i logik och filosofi.

INNEHÅLL: Förord av Dag Prawitz • 1. Metafysiska tvister om realism• 2. Kan analytisk filosofi vara systematisk, och bör den vara det? • 3. Vad är en meningsteori? (II) • 4. Vad har jag för kunskap, när jag kan ett språk? • 5. Frege och Wittgenstein • 6. Ett försvar för McTaggarts bevis för tidens overklighet

Thales 1995. ISBN 91-87172-70-4. 208 sidor. Cirkapris inbunden 217 kr

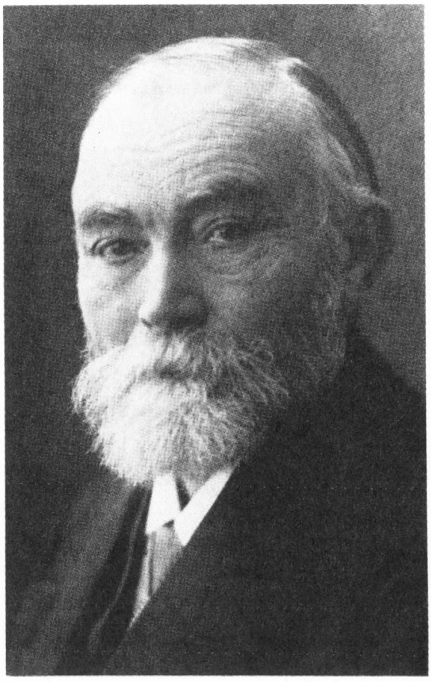


Ludwig Wittgenstein
Zettel

UTGIVNA AV
GEORG HENRIK VON WRIGHT
& ELIZABETH ANSCOMBE
I ÖVERSÄTTNING AV
BETTINA JOBIN

Gottlob Frege
Skrifter i urval

I ÖVERSÄTTNING AV
DANIEL BIRNBAUM OCH
SVEN-OLOV WALLENSTEIN



I detta fylliga urval möter läsaren för första gången på svenska Freges centrala skrifter, kompletterade med några postumt utgivna texter av utomordentlig kvalitet och delar av hans brevväxling med Russell och Wittgenstein.

INNEHÅLL: 17 kärnsatser om logiken • Funktion och begrepp • Om mening och betydelse • Om begrepp och föremål • Kommentarer till mening och betydelse • Logik • Tanken • Negationen • Tankefogning • Brev Frege & Russell • Brev Frege & Jourdain • Brev Frege & Hilbert • Brev Frege & Wittgenstein

Thales 1995. ISBN 91-87172-43-7. 204 sidor. Cirkapris inbunden 217 kr

Mer än hälften av texterna är skrivna under åren 1946-48 sedan den första delen av *Filosofiska undersökningar* blivit klar och innan dess andra del utformades. Samlingen kompletterar därmed Wittgensteins huvudarbete.

Originalupplagan av denna bok utgavs parallellt på tyska och engelska 1967. Den nu gjorda översättningen följer den andra reviderade upplagan som utgavs 1979.

Thales 1995. ISBN 91-87172-65-8. 143 sidor. Cirkapris häftad 147 kr

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

- sändes till redaktören, Lars Bergström, Reimersholmsg. 39, 117 40 Stockholm
- skall vara försedda med namn och adress
- skall som regel vara skrivna på svenska, och citat från andra språk bör översättas till svenska
- noter och litteraturhänvisningar bör inarbetas i texten
- särskild litteraturförteckning upprättas i alfabetisk ordning och placeras sist i manus
- manus skall vara maskinskrivna på A4-papper med skrift endast på arkets ena sida och med bred vänstermarginal
- manus bör dessutom bifogas på diskett för Macintosh (disketten kommer naturligtvis att returneras)
- manus kan också skickas per datorpost till adress Lars.Bergstrom@philosophy.su.se
- för icke beställt material ansvaras ej
- korrektur läses i regel endast av redaktören
- ändringar mot manuskript bekostas icke av förlaget; författaren debiteras kostnad för sådana ändringar
- införda bidrag honoreras ej för närvarande
- i stället för särtryck erhåller författaren gratis 10 exemplar av det nummer av tidskriften i vilket bidraget varit infört

Tidigare årgångar av Filosofisk tidskrift kan beställas från Bokförlaget Thales, Redaktionen, Djupdalskvior 5, 620 12 Hemse.

För 1980 — 1991 är priset 50 kr/årg, enstaka ex 15 kr
För 1992 — 1995 är priset 120 kr/årg, enstaka ex 35 kr. Nr 4 1995 75 kr
Porto ingår