Wlodzimierz Rabinowicz

Om ratifikationismen. Kritik av Jeffreys nya "beslutslogik"

1. Inledning

Denna uppsats behandlar en beslutsteoretisk fråga. När man talar om beslutsteori, är det i själva verket ingen bestämd teori som man normalt åsyftar. Termen ifråga fungerar i stället som en samlande benämning på en hel flora av teorier om beslutsfattande. Av dessa har en del mer eller mindre beskrivande karaktär: deras syfte är att modellera beslutsprocesser och därigenom möjliggöra beslutsförklaringar och beslutsförutsägelser. Andra har i stället en normativ funktion: syftet blir då att tala om hur vi bör fatta våra beslut eller att ange generella villkor som ett beslut måste uppfylla för att vara korrekt eller "rationellt". (Gränsen mellan normer och idealiserade beskrivningar kan emellertid vara ganska flytande.)

Med "ratifikationismen" eller "ratifierbarhetsmaximen" avses här den normativa beslutsprincip som har formulerats av Richard Jeffrey i den senaste utgåvan av hans *Logic of Decision* (1983). Min avsikt är att förklara Jeffreys nya princip och ifrågasätta den.[1]

Först en kort bakgrundshistoria. I första upplagan av sin bok, som kom 1965, försvarade Jeffrey en beslutsprincip enligt vilken agentens handlingsbeslut är korrekt (rationellt, om man så vill) om, och endast om, den förväntade "önskvärdheten" hos den beslutade handlingen är maximal. D v s maximal i jämförelse med andra handlingsalternativ som står agenten till buds. Den förväntade önskvärdheten, eller nyttan, hos en handling definieras som en viktad summa av handlingens önskvärdhetsvärden (nyttovärden) under olika möjliga sakernas tillstånd. Som vikter använde Jeffrey betingade sannolikheter för de olika tillstånden givet handlingen ifråga. De jeffreyska vikterna hade alltså formen

Filosofisk tidskrift 1985 nr 4, 16-33

$P(S/A)$,

där S är (beskrivningen av) ett möjligt tillstånd, A är (beskrivningen av) den handling som är föremål för bedömningen, och där P (S/A) - den betingade sannolikheten för S givet A - definieras, som vanligt, som en proportion mellan sannolikheten för S och A och sannolikheten för A:

$$\frac{P(S\&A)}{P(A)}.$$

Observera att $P(S/A)$ är definierad endast i de fall då $P(A)$ är större än noll. Det är också nödvändigt att påpeka att de relevanta sannolikheterna här ges en 'subjektiv' tolkning: de är mått på styrkan i agentens övertygelser. Dessa sannolikheter varierar mellan ett - absolut säkerhet - och noll - absolut misstro (absolut säkerhet på motsatsen). Funktionen P antas representera agentens övertygelser omedelbart före valögonblicket.

Jeffreys ansats från 1965 bör lämpligen kontrasteras med den klassiska approach som utvecklades av Leonard Savage i *The Foundations of Statistics* (1954). Den för oss mest intressanta skillnaden mellan Jeffrey och Savage ligger i att den senare som sina vikter väljer 'obetingade' sannolikheter för tillstånden:

$P(S)$

I fortsättningen skall jag tala om handlingens "Savagevärde" för att beteckna dess förväntade nytta beräknad enligt Savages recept: i termer av de obetingade sannolikheterna.

Jeffrey försvarade sin ansats genom att påpeka att de för utfallet relevanta tillstånden ibland kan vara mer eller mindre beroende av de handlingar som är föremål för bedömning. Och agenten kan vara medveten om detta beroendeförhållande. I sådana fall skulle valet av obetingade sannolikheter leda agenten på avvägar. Betingade sannolikheter däremot tycks utgöra ett adekvat mått på handlingens förväntade kausala inflytande på tillstånden.

Emellertid finns det alla skäl att tro att Jeffrey 1965 skulle ha betraktat den av Savage föreslagna beslutsprincipen som helt oantastlig *om* bara denna princip fick begränsas till de fall i vilka agenten, före valet, är säker på att de olika tillstånden är kausalt oberoende av de

tillgängliga handlingarna. Jeffrey skulle ha uppfattat denna begränsade Savageprincip som ett specialfall av sin egen beslutsprincip. Vid denna tid tycktes han nämligen anta att säkert kausalt oberoende automatiskt implicerar probabilistiskt oberoende:

Oberoendeantagandet. Om agenten, före valet, är säker på att tillståndet S är kausalt oberoende av handlingen A, så

$$P(S/A)=P(S).$$

På senare år har detta oberoendeantagande förkastats av flera filosofer, bl a av Jeffrey själv.[2] Antagandet håller inte i de fall i vilka agenten, före valet, är övertygad om att tillstånden befinner sig utanför hans inflytande men samtidigt betraktar sitt kommande handlingsval som ett mer eller mindre tillförlitligt *tecken* eller symptom på det föreliggande tillståndet. Agenten tillskriver alltså sina handlingar rent evidentiell relevans med avseende på tillstånden – antingen därför att han uppfattar sina handlingar som kausalt beroende av tillstånden (i stället för tvärtom) eller också därför att såväl hans handlingar som tillstånden enligt honom kan hänföras till en gemensam orsak. Jag skall kalla sådana fall för *Newcomblika* eftersom det mest omtalade fallet av denna typ har blivit känt under namnet "Newcombs problem". I Newcomblika fall kan den betingade sannolikheten för ett tillstånd givet en handling – $P(S/A)$ – avvika från den obetingade sannolikheten för samma tillstånd – $P(S)$ – även om agenten är övertygad att tillståndet ifråga är kausalt oberoende av handlingen.

Newcombs problem

"Professor L konfronteras med två lådor, en genomskinlig och en ogenomskinlig. Den genomskinliga lådan kan ses innehålla 1 000 dollar. Han får ta antingen den genomskinliga lådan med allt dess innehåll eller bägge lådorna med allt deras innehåll. En förutsägare, som är mycket duktig på att förutsäga val som människor träffar, har lagt 1 000 000 dollar i den ogenomskinliga lådan om han har förutsagt att professor L skall ta endast den ogenomskinliga lådan, och inget om han har förutsagt att professor L skall ta bägge. Förutsägaren är inte bara mycket duktig i största allmänhet; han är också mycket duktig vad gäller dem som tar endast den ogenomskinliga lådan, d v s för dem som tar enbart den ogenomskinliga lådan är hans andel av kor-

rekta förutsägelser mycket hög. Likaledes vad gäller dem som tar bägge lådorna. [Allt detta känner professor L till.] Under dessa omständigheter tar professor L enbart den ogenomskinliga lådan. I den hittar han 1 000 000 dollar. 'Jag är rik', utropar han. 'Du hade varit 1 000 dollar rikare om du hade tagit bägge lådorna', anmärker professor G" (Skyrms 1981, s 262).[3]

Låt *1M* respektive $\overline{1M}$ representera de alternativa tillstånden: *Det ligger en million dollar i den ogenomskinliga lådan* respektive *Den ogenomskinliga lådan är tom.* A_1 respektive A_2 skall stå för handlingsalternativen: *Professor L tar endast den ogenomskinliga lådan* respektive *Professor L tar bägge lådorna.*

Den tillgängliga informationen om förutsägarens skicklighet gör att professor L, före valet, uppfattar sin kommande handling som en tillförlitlig indikation på det föreliggande tillståndet. Hans betingade sannolikheter för *1M* givet A_1 och för *1M* givet A_2 är bägge mycket höga. De måste alltså bägge vara större än ½. Följaktligen blir deras summa större än ett. Därför måste åtminstone en av dem (och antagligen bägge) vara större än professor L's obetingade sannolikhet för motsvarande tillstånd. Ty hans obetingade sannolikheter för *1M* och $\overline{1M}$ summeras till ett. Samtidigt är professor L säker på att de handlingar som just nu står honom till buds inte kan kausalt påverka innehållet i den ogenomskinliga lådan. Newcombs problem är således Newcomblikt.

I Newcomblika fall kan Jeffreys beslutsprincip från 1965 mycket väl komma i konflikt med den begränsade Savageprincipen. Följer man Jeffreys princip kan det inträffa att man väljer en handling endast på grund av dess höga 'evidentiella värde', d v s endast därför att handlingen ifråga tyder på förekomsten av ett visst fördelaktigt tillstånd. Detta kan vara fallet även om handlingens förväntade *kausala* effekter är sämre än de som agenten kan förväntas uppnå genom att utföra en annan handling i stället. Så, till exempel, i Newcombs problem föreskriver Jeffreys princip från 1965 att professor L bör ta enbart den ogenomskinliga lådan - och därigenom förlora tusen dollar. Ty denna handling tyder på att den ogenomskinliga lådan innehåller en million. Vi skulle kunna säga att Jeffreys förväntade önskvärdhet, genom att den definieras i termer av betingade sannolikheter för tillstånden, inte så mycket mäter handlingens förväntade

kausala värde som dess värde *som nyhet.*[4] När agenten är övertygad att han saknar möjligheter att påverka tillstånden tycks det förväntade kausala värdet hos hans handlingar i stället sammanfalla med deras Savagevärde. Tycker man nu att det är handlingens förväntade kausala värde som bör vägleda valet så följer det att, i Newcomblika fall, de rätta föreskrifterna genereras av den begränsade Savageprincipen och inte av Jeffreys princip. Så till exempel, i Newcombs problem, föreskriver Savageprincipen att professor L utför A_1 - att han tar bägge lådorna. Det är nämligen lätt att inse att A_2 *dominerar* A_1: A_2 ger mer än A_1 under varje tillstånd - såväl under IM som under $\overline{IM}$. Och man kan visa att Savagevärdet hos en dominerande handling med nödvändighet överstiger Savagevärdet hos den handling som är dominerad.

Konfronterad med Newcomblika fall bestämde sig Jeffrey (1983) att revidera sin ursprungliga teori och han framkastade nu en ny beslutsprincip: "ratifierbarhetsmaximen". Därför faller det sig naturligt att undersöka hur denna nya maxim förhåller sig till den begränsade Savageprincipen. Är dessa två förenliga med varandra eller leder de ibland till motstridiga föreskrifter?

2. Ratifikationismen

I den nya utgåvan av *The Logic of Decision* påpekar Jeffrey att handlings*beslut* brukar vara lika tillförlitliga tecken på tillstånd som handlingarna själva. Så till exempel skulle man kunna hävda att redan professor L's beslut att ta enbart den ogenomskinliga lådan ger honom ett lika bra tips om innehållet i denna låda som den handling som åtföljer beslutet. Handlingens evidentiella värde uttöms, med andra ord, i förväg av handlingsbeslutet. Detta förhållande kan vi utnyttja, tänker sig Jeffrey. När vi ställs inför ett val bör vi först göra ett rent hypotetiskt antagande om vårt kommande handlingsbeslut och endast därefter - på grundval av detta antagande - bör vi jämföra de olika handlingar som står oss till buds med avseende på deras förväntade önskvärdhet. Tanken är att genom en sådan 'konditionalisering' av förväntade önskvärdhetsvärden på ett hypotetiskt antagande om vårt handlingsbeslut avskärmar vi, så att säga, de skillnader mellan hand-

lingar som uteslutande beror på deras olika evidentiella relevans. Och en dylik avskärmning är givetvis nödvändig eftersom de rent evidentiella skillnaderna mellan handlingarna inte får tillåtas att vägleda valet.

Låt mig nu beskriva Jeffreys ansats litet mer i detalj. Vi låter '*bA*' stå för påståendet att agentens slutgiltiga beslut blir att utföra handlingen *A*. Observera att för Jeffrey är ett beslut slutgiltigt i och med att det inte kommer att ändras av agenten. Emellertid är misslyckanden alltid möjliga. Det är alltid möjligt att agenten misslyckas att genomföra sitt slutgiltiga beslut och att han utför en annan handling i stället. Därför förutsätter Jeffrey, för varje två handlingsalternativ *A* och *A'*, att den villkorliga sannolikheten för A givet *bA'* alltid är större än noll, även om den kan vara ytterst liten.[5]

Om *A* och *A'* ingår i agentens alternativmängd, kan vi tala om *A:s förväntade önskvärdhet på villkor att bA'*. D v s om det förväntade önskvärdhetsvärde som tilldelas handlingen *A* på basen av ett hypotetiskt antagande att agentens slutgiltiga beslut blir att utföra handlingen *A'*.[6] Detta värde beräknas på samma sätt som Jeffreys ursprungliga förväntade önskvärdhet hos *A* men vikterna är nu annorlunda: alla vikter av formen

P(*S*/*A*)

konditionaliseras nu på det hypotetiska antagandet *bA'*. P(*S*/*A*) ersätts därför med

P(*S*/*A*&*bA'*).[7]

Nu kan vi förklara vad Jeffrey menar med ratifierbarhet. Den intuitiva idén är enkel: ett handlingsbeslut är ratifierbart om det inte berövar den beslutade handlingen dess förväntade värde. Med en mera precis formulering:

Beslutet att utföra *A* är *ratifierbart* om, och endast om, den förväntade önskvärdheten hos *A* på villkor att *bA* är åtminstone lika stor som den förväntade önskvärdheten hos varje alternativ handling bedömd *på samma villkor* (d v s *bA*).

För enkelhets skull kommer jag i fortsättningen att ibland tala om ratifierbara handlingar (och inte bara om ratifierbara handlingsbeslut).

En handling antas vara ratifierbar om beslutet att utföra den är ratifierbart.

Jeffreys ratifierbarhetsmaxim föreskriver nu agenten att fatta ratifierbara beslut. Ratifierbarhet framstår som ett både nödvändigt och tillräckligt villkor för korrekt handlande (jfr Jeffrey 1983, kap 1).[8]

Hur kan denna maxim ta hand om Newcomblika fall? Redan tidigare har jag skisserat Jeffreys svar på denna fråga. Genom att konditionalisera på ett givet beslut kan vi normalt helt och hållet "avskärma" de rent evidentiella skillnaderna mellan olika handlingar. Beslutet att handla utgör med andra ord ett lika pålitligt tecken på det föreliggande tillståndet som handlingen själv. *Om* nu detta "avskärmningsantagande" håller så möjliggör en konditionalisering på ett bestämt handlingsbeslut en rätt sorts jämförelse mellan den motsvarande handlingen och dess alternativ - en jämförelse som uteslutande sker i termer av det förväntade kausala värdet hos de olika handlingarna. Och det är precis en sådan konditionalisering som utgör grunden för vår bedömning om en given handling är ratifierbar eller ej.

Så till exempel, givet avskärmningsantagandet, kommer professor L's betingade sannolikheter för IM givet $A_1\&bA_1$ respektive givet $A_2\&bA_1$ att sammanfalla: bägge kommer att vara mycket höga. Och analogt blir det med hans sannolikheter för $\overline{IM}$ givet $A_1\&bA_2$ respektive givet $A_2\&bA_2$. Att ta bägge lådorna kommer därför att framstå som ett bättre alternativ på basen av varje hypotetiskt antagande om professor L's handlingsbeslut. Att göra så minskar ju inte sannolikheten för en million (ty denna sannolikhet, givet avskärmningsantagandet, bestäms helt och hållet av handlingsbeslutet och påverkas alltså ej längre av den efterföljande handlingen) och samtidigt ger det alltid tusen dollar mer. Ratifierbarhetsmaximen kommer därför att föreskriva A_2 - precis som den begränsade Savageprincipen.

3. Jeffrey möter Savage

Jag skall här inte diskutera den självklara invändningen som kan göras mot Jeffreys avskärmningsantagande: ibland kan agenten uppfatta sitt eventuella beslut att handla som en något sämre indikation på det

föreliggande tillståndet än den som skulle utgöras av handlingen själv. I sådana fall förmår han inte att fullständigt avskärma handlingens rena nyhetsvärde genom att konditionalisera på handlingsbeslutet. Ratifierbarhetsmaximen kommer därför att leda honom på avvägar. Jeffrey själv erkänner styrkan hos denna invändning[9] och han begränsar därför explicit sin ratifierbarhetsansats endast till de fall i vilka avskärmningsantagandet är satisfierat. Jag skall här följa honom i detta avseende.

Följande bör observeras: om vi accepterar avskärmningsantagandet och om vi håller oss till de fall i vilka agenten är säker på att tillstånden står utanför hans inflytande, kan Jeffreys vikter av formen

$P(S/A\&bA')$

förenklas till

$P(S/bA')$.

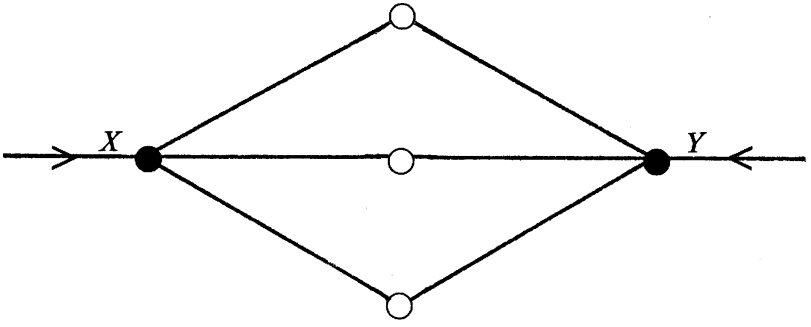
Vi kan alltså undvika referensen till A. Ty i sådana fall är agenten säker på att utförandet av A inte kan ha några kausala effekter på S, och samtidigt (avskärmningsantagandet!) tillmäter han inte utförandet av A någon evidentiell relevans med avseende på S utöver den som redan har tagits om hand genom konditionsaliseringen på bA'.

Vad händer om inget av handlingsalternativen är ratifierbart? Jeffreys maxim kommer då inte att ge oss någon vägledning. Jeffrey gissar att ett sådant problem endast kan uppkomma för en irrationell agent. När inget av handlingsalternativen är ratifierbart finns det enligt Jeffrey fog för misstanke att agentens övertygelser och/eller preferenser vid tiden för valet måste vara patologiska på något sätt. Varför Jeffrey tror att så måste vara fallet är för mig en gåta. I en annan uppsats beskriver jag ett icke-patolgiskt Newcomblikt fall i vilken agenten saknar ett ratifierbart handlingsalternativ (se Rabinowicz 1984; ett liknande fall beskrivs för övrigt så pass tidigt som 1978 av Allan Gibbard och William Harper). Existensen av sådana fall visar redan på en viktig skillnad mellan ratifierbarhetsmaximen och den begränsade Savageprincipen: den senare ger agenten vägledning även i avsaknad av ratifierbara alternativ. Här vill jag emellertid i stället

koncentrera mig på ett Newcomblikt fall i vilket Jeffreys maxim *ger* ett bestämt utslag men där Jeffreys föreskrift kommer i konflikt med den begränsade Savageprincipen.

Vägvalet

Vi tänker oss två agenter, X och Y, som är ute för att mötas. De har startat från olika platser och nu har de kommit fram till var sitt vägskäl. För var och en av dem står valet mellan att gå rakt fram, till vänster eller till höger. Vägkartan ser ut så här:



Punkter anger X och Y:s nuvarande positioner och cirklar visar på möjliga mötesplatser. Som vi ser kommer de att mötas endast om bägge går rakt fram eller också om den ene går till vänster och den andre till höger. Om bägge fortsätter rakt fram blir vägen visserligen något kortre men de blir tvungna att betala höga tullavgifter. (Om vi så vill kan vi tänka oss att den raka sträckan är en fransk motorväg eller att X och Y är två medeltida köpmän.) Vi antar dessutom att vägen till vänster är bekvämare än vägen till höger, något som varje agent känner till. Alla dessa faktorer tillsammans bestämmer deras önskvärdhetsvärden för de olika utfallen. Här presenterar jag endast X:s värdematris. Y:s matris antas vara analog.

		Y		
		vänster	rakt fram	höger
X	vänster	0,1	0,1	1,1
	rakt fram	—0,5	0,5	—0,5
	höger	1	0	0

Det bästa (1,1) är att möta den andre efter att man har tagit den bekväma vänstra vägen. Men även om mötet kommer till stånd efter vissa strapatser är det också mycket bra, sålänge man inte behöver betala tull (1). Att behöva betala tull för att mötas är inte särskilt lockande (0,5) men det är i alla fall bättre än att inte mötas alls – oberoende av om vägen annars är bekväm (0,1) eller strapatsrik (0). Och det sämsta (-0,5) är givetvis om man betalar tull och ändå missar mötet.

Ingen kommunikation och inget samarbete mellan agenterna är möjligt: avståndet är för stort och de saknar tillgång till telefon eller radio. För enkelhets skull antar jag också att ingen av agenterna får invänta den andre (de har ont om tid) och att ingen av dem får träffa sitt val genom att kasta mynt eller tärning (kanske blir de hårt bestraffade om de försöker; eller också förbjuder deras religion att viktiga val överlämnas åt slumpen).

X och *Y* antas vara mycket lika varandra psykologiskt sett: de är mer eller mindre psykologiska tvillingar och de känner till att så är fallet. De har dessutom råkat ut för liknande koordineringsproblem flera gånger tidigare och då har de alltid eller nästan alltid handlat på samma sätt: vänster–vänster, höger–höger, eller rakt fram–rakt fram. Emellertid, i de flesta tidigare fall, har de bägge undvikit att gå rakt fram (för att slippa betala tull) och valt att gå till vänster i stället. Att de hade föredragit vänster framför höger beror på att den vänstra vägen varit så bekväm. Bägge agenterna har alltså varit benägna att ta

det säkra före det osäkra: det sämsta möjliga utfallet när man går till vänster (0,1) är bättre än de sämsta utfallen för de två andra handlingsalternativen (-0,5 resp. 0).

Läsaren skulle här kanske vilja påpeka att *X* och *Y* borde för länge sedan ha koordinerat sitt beteende i dylika situationer genom en lämplig överenskommelse. T ex: "I framtiden skall jag, *X*, gå till vänster medan du, *Y*, skall ta den mindre bekväma högra vägen. Kostnaden för dina större strapatser skall vi bära tillsammans". Ja, det borde de ha gjort. Det var dumt av dem att inte komma överens, men nu är det för sent. Nu står de där de står och var och en av dem måste göra det bästa han kan.

Varje agent känner till allt det här och denna kunskap bestämmer hans sannolikhetsbedömningar. I själva verket är agenternas sannolikhetsbedömningar helt analoga. Om vi därför koncentrerar oss på *X* och låter P vara *X*:s sannolikhetsfunktion vid tiden just före vägvalet, så gäller följande:

(1) P(*Y* går till vänster) är hög, säg, högre än 0,5;
(2) P(*Y* går i riktningen *r*/*b*(*X* går i riktningen *r*)) är mycket hög, säg, högre än 0,8.

(1) bestäms av *X*:s kunskap om *Y*:s tidigare vänstervridna beteende och om den benägenhet att ta det säkra före det osäkra som ligger bakom beteendet ifråga. (2) är uppfyllt oberoende av vilken riktning *r* står för: vänster, höger eller rakt fram. "b(*X* går i riktningen *r*)" representerar påståendet att *X*:s slutgiltiga beslut blir att gå i riktningen *r*. Att (2) gäller beror på att *X* uppfattar sitt kommande handlingsbeslut, vilket det än vara månde, som ett pålitiligt tecken på hur *Y*, hans psykologiska tvilling, kommer att handla. Låt oss också, för argumentets skull, förutsätta att avskärmningsantagandet är satisfierat. Här slutar min beskrivning av vägvalsexemplet.

X uppfattar sina handlingar och handlingsbeslut som tecken på tillstånden (*Y*:s handlingar). Samtidigt är *X* säker på att tillstånden är kausalt oberoende av de handlingar som just nu står honom till buds: *X* och *Y* handlar oberoende av varandra. Vägvalsexemplet är alltså Newcomblikt.

Den begränsade Savageprincipen föreskriver att X skall gå *till höger:* den 'obetingade' sannolikheten för att Y kommer att gå till vänster är ju hög (jfr (1) ovan), och om Y går till vänster kommer de bägge agenterna att mötas endast om X går till höger.

Ratifierbarhetsmaximen föreskriver däremot att X skall gå *rakt fram* eftersom att gå rakt fram är det enda ratifierbara handlingsalternativet. Låt mig förklara. Ponera först att X beslutar att gå rakt fram. På basen av detta hypotetiska antagande är det mycket sannolikt (i X:s ögon) att Y kommer att gå den raka vägen. (2) implicerar att denna sannolikhet överstiger 0,8. Och om Y går den raka vägen så kommer de att mötas endast om X gör likadant. I själva verket kan man lätt beräkna att, på villkor att X beslutar att gå rakt fram, blir den förväntade önskvärdheten hos alternativet 'rakt fram' större än 0,3 ($0{,}8\times\frac{1}{2}+0{,}2\times(-\frac{1}{2})$). På samma villkor blir däremot den förväntade önskvärdheten hos 'vänster' mindre än 0,3 ($0{,}8\times0{,}1+0{,}2\times1{,}1$), medan det motsvarande värdet för 'höger' blir mindre än 0,2 ($0{,}8\times0+0{,}2\times1$). 'Rakt fram' framstår alltså som ett ratifierbart alternativ. Ponera nu i stället att X beslutar att gå till höger. På basen av *detta* hypotetiska antagande blir det högst sannolikt att Y kommer att gå till höger (jfr (2)). Och då kommer X, genom att gå till höger, att missa sitt möte med Y. Mötet skulle komma till stånd endast om X, i strid med sitt beslut, gick till vänster i stället. Det är därför klart att alternativet 'höger' inte är ratifierbart. Ett exakt analogt resonemang visar att samma sak gäller alternativet 'vänster'.

Att det föreligger en sådan konflikt mellan de två principerna är lätt att förstå. När agenten är säker på att tillstånden står utanför hans kontroll, föreskriver ratifierbarhetsmaximen att agenten bör träffa sitt val i termer av sina *ex post* sannolikheter. D v s sannolikheter av formen $P(S/bA)$. Valet bör träffas utifrån de (hypotetiska) sannolikheter som agenten skulle tilldela tillstånden efter valet. Den begränsade Savageprincipen rekommenderar däremot agenten att använda tillståndssannolikheter *ex ante.* Valet bör träffas utifrån agentens sannolikhetstilldelningar *före* valet. Och det är uppenbart att i Newcomblika fall kommer dessa bägge typer av sannolikheter att vara olika. Ty i sådana fall antas agentens val vara evidentiellt relevant med avseende på tillstånden. Det är därför inte förvånande att de två beslutsprinciperna ibland kommer att leda till motsatta föreskrifter.

4. Savage på defensiven

Det vore bra att kunna avsluta denna uppsats med ett konklusivt argument till förmån för en av de två beslutsprinciperna eller, eventuellt, till nackdel för bägge. Tyvärr har jag inget sådant argument att komma med. Det enda som jag kan prestera är ett försvar för Savage i det konkreta fall som jag har beskrivit - ett försvar mot två tänkbara invändningar som skulle kunna anföras av en Jeffreyanhängare.

Invändning 1

Enligt Savage bör X gå till höger. Men samma föreskrift måste i så fall gälla även Y. X och Y befinner sig ju i exakt likadan situation gentemot varandra: vägvalsexemplet är helt symmetriskt. Emellertid, om bägge agenterna följer Savages föreskrift, så kommer de inte att mötas. Det vore bättre för var och en av dem om de bägge gick rakt fram i stället. De skulle då visserligen få betala tull men de skulle också träffa varandra. Detta är bättre än att inte mötas alls. Och att gå rakt fram är just det som varje agent bör göra enligt ratifierbarhetsmaximen. Att en beslutsprincip rekommenderar alla agenter ett sätt att handla vars sammanlagda resultat med nödvändighet är sämre för varje agent än det sammanlagda resultatet av ett alternativt handlingssätt - är inte detta något som visar på principens ohållbarhet?

Savage kan bemöta denna invänding genom att påpeka att samma svårighet mycket väl kan uppkomma även för Jeffrey. Låt mig ge ett exempel. Givet avskärmningsantagandet kan det visas att ratifierbarhetsmaximen, i likhet med Savage-principen, alltid rekommenderar agenten att utföra den dominerande handlingen, om en sådan finns att tillgå. (En handling är dominerande, som vi kommer ihåg, om den under varje tillstånd ger agenten mera än de andra handlingsalternativen.) Detta gäller sålänge agenten är säker på att tillstånden står utanför hans inflytande.

Ponera nu att en viss situation involverar flera agenter som får handla oberoende av varandra, att utfallet bestäms av vad de olika agenterna gör (så att ur varje agents perspektiv de andra agenternas alternativa handlingsmönster kan ses som olika möjliga tillstånd som

agenten ifråga måste räkna med), och att det finns ett sätt att handla som är dominerande för varje agent men vars sammanlagda utfall är sämre för var och en än det sammanlagda utfallet hos något alternativt handlingssätt. (Det som man vinner genom att utföra den dominerande handlingen kompenserar en inte för det man förlorar genom att andra handlar likadant.) En situation av detta slag brukar som bekant kallas för "fångens dilemma" och den har varit föremål för omfattande besluts- och spelteoretiska diskussioner. I fångens dilemma kommer ratifierbarhetsmaximen att rekommendera varje agent det dominerande handlingssättet. [10] Den kommer alltså att leda till samma oroande effekt som Savageprincipen leder till i vägvalsexemplet: om alla agenter följer principen blir det sämre för var och en av dem än om de alla hade valt ett annat sätt att handla. [11]

Att hans maxim slår ut på detta sätt i fångens dilemma är givetvis något som Jeffrey är väl medveten om och som han inte alls anser vara någon nackdel hos maximen. Tvärtom (jfr Jeffrey 1983, kap 1). Han själv skulle därför aldrig komma på tanken att kritisera Savage på detta sätt. Desto större vikt skulle han däremot, om jag inte tar fel, lägga vid den invändning som nu följer.

Invändning 2

Om X bestämde sig att gå till höger, såsom Savage-principen föreskriver, så skulle han, efter att ha tagit detta beslut, vara ganska övertygad om att Y, hans psykologiska tvilling, också kommer att gå till höger. Denna övertygelse från X:s sida skulle ju också vara fullständigt befogad under sådana omständigheter. Och om Y går till höger, kommer X genom att gå till höger att gå miste om mötet. Visar detta inte att beslutet att gå till höger måste vara irrationellt? Själva detta beslut skulle ju leda X till en övertygelse i vars ljus den beslutade handlingen måste framstå som klart ofördelaktig. Bör inte X gardera sig mot sådant genom att i stället träffa sitt val i termer av sina hypotetiska *eftervals*sannolikheter (*ex post* sannolikheter)?

Jag gissar att Savages svar skulle lyda så här: Det är riktigt att beslutet att gå till höger skulle leda X till en under dessa omständigheter fullständigt befogad övertygelse om att Y också kommer att gå till höger. *Men,* före valet tror X att Y kommer att gå till *vänster.* Vad

mera är, X är säker på att hans eget val, vilket det än skulle bli, omöjligtvis kan påverka Y:s beteende. Detta innebär att X, före valet, betraktar sin hypotetiska framtida övertygelse om att Y kommer att gå till höger som *falsk*. Befogad, men falsk. Vad finns det då för skäl för X att träffa sitt val i termer av hypotetiska eftervalsövertygelser som han nu, före valet, anser vara falska?

Jag tror att denna fråga är på sin plats. Åtminstone i vägvalsexemplet tycks Jeffreys ratifierbarhetsmaxim leda oss på avvägar.

Noter

1. En längre version av denna uppsats kommer att publiceras i ett annat sammanhang. Se Rabinowicz (1984). Många personer har hjälpt mig med synpunkter och kommentarer. Särskilt vill jag tacka Lars Bergström, Sven Danielsson, Allan Gibbard, Sten Lindström och, inte minst, Howard Sobel.

2. Jfr t ex Nozick (1969), Gibbard och Harper (1978), Sobel (1979) och (1984), Skyrms (1980) och (1982), Lewis (1981), Jeffrey (1983). Oberoendeantagandet försvaras däremot av Eells (1982).

3. Skyrms beskriver också ett antal andra Newcomblika fall. Newcombs problem har först presenterats i tryck av Robert Nozick (1969). Jeffreys favoritexempel på ett Newcomblikt fall är en version av det s k "fångens dilemma", i vilken fångarna antas vara (och känna till att de är) psykologiska tvillingar. Följden blir att varje fånge uppfattar sitt eget handlingsbeslut som en indikation på att den andre fången kommer att handla likadant. Se Jeffrey (1983), kap 1. Jfr Lewis (1979) och Sobel (1983).

4. 1965 ansåg Jeffey att dessa bägge värden sammanfaller med varandra. Jfr Jeffrey (1965), ss 73 f. Konstigt nog återkommer samma påstående i Jeffrey (1983), s 84, fast hans åsikter på denna punkt har under tiden undergått väsentliga förändringar. Jag skulle gissa att det är fråga om ett förbiseende från Jeffreys sida.

5. Den nämnda förutsättningen är oundgänglig för Jeffreys teori (se nedan, not 8) men den är samtidigt inte alls okontroversiell. Det är ju en sak om agenten erkänner att han alltid kan misslyckas i genomförandet av sitt slutgiltiga beslut. Därigenom behöver han inte erkänna att ett sådant misslyckande i princip kan leda till vilken annan alternativ handling som helst. Kan jag verkligen misslyckas att genomföra mitt slutgilitga beslut att stanna hemma ikväll på ett sådant sätt att jag går på bio i stället? Och detta, märk väl, utan att jag 'ändrar' mig, ty om jag ändrade mig, så hade mitt slutgiltiga beslut att stanna hemma inte varit slutgiltigt.

6. Ett specialfall föreligger när $A=A'$. Normalt kommer emmellertid A och A' att vara olika handlingar.

7. Egentligen borde vi också tillämpa samma sorts konditionalisering vad gäller de viktade önskvärdhetsvärdena. Men för enkelhets skull antar vi här att denna komplikation inte behövs. Vi antar alltså, för varje S och A, att A:s värde under S är detsamma oberoende av vilket som var agentens slutgiltiga beslut. Med andra ord: utfallet bestäms uteslutande av vad agenten faktiskt gör och inte av vad han beslutar sig att göra.

8. Som nämndes ovan förutsätter Jeffrey att $P(A/bA')>0$, för alla handlingsalternativ A och A'. Detta är detsamma som att anta att $P(A\&bA')$ alltid är större än noll. Att denna förutsättning är väsentlig för Jeffrey kan nu lätt inses. Om $P(A\&bA')=0$, blir alla vikter av formen $P(S/A\&bA')$ odefinierade. Därför kan den förväntade önskvärdheten hos A på villkor att bA' inte heller definieras, och det blir följaktligen omöjligt att tillämpa ratifierbarhetsmaximen på handlingen A'. A' kommer inte att kunna jämföras – enligt Jeffreys recept – med den altenativa handlingen A och beslutet att utföra A' kommer att framstå som varken ratifierbart eller oratifierbart.

9. Han hänvisar i detta sammanhang till Bas van Fraassen (se Jeffrey 1983, kap 1). Samma argument framförs också av Sobel (1984).

10. Följande värdematris kan exemplifiera fångens dilemma:

		Y: gör si	Y: gör så
X	gör si	1 / 1	0 / 3
	gör så	3 / 0	2 / 2

(I varje cell anger den nedre siffran X:s värdering av utfallet; den övre siffran anger utfallets värde för Y.) Att göra si dominerar att göra så för varje agent. Men om bägge gör si får var och en mindre än om de bägge gör så i stället. Givet avskärmningsantagandet föreskriver ratifierbarhetsmaximen, liksom Savageprincipen, varje agent att göra si.

11. Denna oroande effekt uppkommer för övrigt även i andra typer av situationer, bland annat sådana som, till skillnad från fångens dilemma, saknar dominerande handlingsätt och i vilka, återigen till skillnad från fångens dilemma, råder det en fullständig överensstämmelse mellan agenternas värdering

av de olika utfallen. Ett exempel: Anta att X och Y ställs inför likadana val (mellan tre handlingar: h_1, h_2 och h_3), att de får handla oberoende av varandra och att utfallet bestäms av vad de bägge gör, att det vad den ene väljer saknar evidentiell relevans för vad den andre kommer att göra (situationen är alltså *inte* Newcomblik), och att X och Y har exakt samma värdematris:

		Y		
		h_1	h_2	h_3
X	h_1	10	0	0
	h_2	0	9	0
	h_3	0	0	10

Förutsatt att varje agent anser det vara mest sannolikt att den andra utför h_2 (något som väl är att vänta om agenterna saknar ytterligare informationer), kommer h_2 att föreskrivas av såväl Jeffrey som Savage. Men om bägge agenterna följer föreskriften blir det sämre för var och en av dem än om de bägge hade valt ett annat sätt att handla (d v s om de bägge utförde h_1 eller om de utförde h_3).

Litteratur

Ellery Eells, 1982, *Rational Decision and Causality* Cambridge University Press, Cambridge.

Allan Gibbard och William L. Harper, 1978, 'Counterfactuals and Two Kinds of Expected Utility', i A. Hooker, J.J. Leach and E.F. McClennen (utg.), *Foundations and Applications of Decision Theory,* vol. 1, Reidel, Dordrecht, Holland; omtryckt i W.L. Harper, R. Stalnaker, G. Pearce (utg.), *Ifs,* Reidel, Dordrecht, Holland, ss. 153-190.

Richard C. Jeffrey, 1965, *The Logic of Decision,* Mac Graw-Hill, New York.

Richard C. Jeffrey, 1983, *The Logic of Decision,* andra utgåvan, University of Chicago Press, Chicago and London.

David Lewis, 1979, 'Prisoner's Dilemma is a Newcomb Problem', *Philosophy and Public Affairs* 8, ss. 235-240.

David Lewis, 1981, 'Causal Decision Theory', *Australian Journal of Philosophy* 59, ss. 5-30.

Robert Nozick, 1969, 'Newcomb's Problem and Two Principles of Choice', i Nicholas Rescher (utg.), *Essays in Honour of Carl G. Hempel,* Reidel, Dordrecht, Holland.

Wlodzimierz Rabinowicz, 1984, 'Ratificationism without Ratification: Jeffrey meets Savage', antagen till publicering i *Theory and Decision.*

Leonard J. Savage, 1954, *The Foundations of Statistics,* Wiley, New York; andra omarb. uppl., Dover, New York, 1972.

Brian Skyrms, 1980, *Causal Necessity,* Yale Univ. Press, New Haven.

Brian Skyrms, 1981, 'The Prior Propensity Account of Subjunctive Conditionals', W.L. Harper, R. Stalnaker, G. Pearce (utg.), *Ifs,* Reidel, Dordrecht, Holland, ss. 259-265.

Brian Skyrms, 1982, 'Causal Decision Theory', *The Journal of Philosophy* 79, ss. 695-711.

J. Howard Sobel, 1979, *Probability, Chance and Choice: A Theory of Rational Agency,* opublicerad.

J. Howard Sobel, 1983, 'Not Every Prisoner's Dilemma is a Newcomb Problem', opublicerad.

J. Howard Sobel, 1984, 'Adequate Partitions of Circumstances and Dominance Arguments in a Causal Decision Theory', antagen till publicering i *Synthese.*