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.

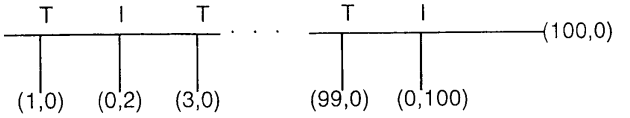
Innan vi fortsätter är det värt att poängtera att dessa extremt starka villkor på spelarna gör att teorins prediktiva styrka i verkliga situationer är låg.

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:

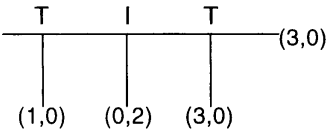
	IN	UT
kooperativ	3,3	5,2
aggressiv	1,0	5,2

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.