~~[start kap]~~

Kapittel 3

Handling og rasjonalitet

Edmund Henden

Innledning

Et viktig mål i samfunns- og helsevitenskapelig forskning er å forklare ulike typer sosiale fenomener. Sosiale fenomener oppstår som resultat av samhandling mellom mennesker. I denne betydningen er stater, bedrifter, skoler og sykehus alle eksempler på sosiale fenomener. Andre eksempler er kollektive praksiser som å vaksinere seg, gå til helsekontroll, betale skatt, stemme og så videre. Sosiale fenomener kan bare eksistere dersom mange mennesker handler på lignende måter og reagerer på hverandres handlinger. Et vanlig syn er at det å forklare slike fenomener innebærer å besvare en form for «hvorfor-spørsmål» om dem. Hvorfor øker sykefraværet? Hvorfor velger noen pasienter private helseløsninger? Hvorfor fortsetter folk å røyke til tross for at de er godt kjent med skadevirkningene? En viktig problemstilling i vitenskapsteori dreier seg om hvilken form svarene på slike spørsmål bør ha for å være *vitenskapelige* forklaringer.

Ifølge et vitenskapsteoretisk syn som kalles «metodologisk individualisme» er forklaringer på sosiale fenomener bare fullstendige dersom de er basert på enkeltindividene som gir opphav til disse fenomenene. Enkeltindivider er motiverte til å handle og reagere på bestemte måter. En vitenskapelig forklaring på et sosialt fenomen påviser en sammenheng mellom motivasjonen og handlingene til mange enkeltindivider og fenomenet som skal forklares. Metodologiske individualister oppgir ulike grunner for dette synet. Én grunn som går tilbake til sosiologen Max Weber, er at sosiale fenomener er iboende «meningsfulle». Det betyr at vi bare har tilgang til dem gjennom fortolkning eller forståelse («*verstehen*»). Fordi det er umulig å forstå et sosialt fenomen uten å forstå handlingene det oppstår som resultat av, og fordi det er umulig å forstå handlinger uten å forstå hva det er som motiverer dem, så må forklaringer av sosiale fenomener være forankret i handlingene og motivasjonen til enkeltindivider. En annen grunn som ofte oppgis er at forklaringer identifiserer «årsaker». Fordi årsakene til sosiale fenomener må antas å virke gjennom handlingene til enkeltindivider, så må også forklaringene av dem ta utgangspunkt i handlingene til enkeltindivider. Kort sagt: Forklaringer på sosiale fenomener bør legges på et «handlingsteoretisk nivå».

Men hva betyr det å legge forklaringer på et «handlingsteoretisk nivå», og hvordan er det mulig å gi forklaringer på komplekse sosiale fenomener på dette nivået? Enkeltindivider kan vel ha alle mulige motiver for å handle som de gjør. Må samfunnsforskere undersøke motivasjonen til hver enkelt av disse individene for å finne forklaringen på fenomenet? Svaret er *nei*. Det skyldes at det er felles mønstre i folks handlinger som gjenfinnes på tvers av enkeltindivider. Et utbredt syn er at disse mønstrene kan forklares ut fra en antagelse om at folk flest stort sett er «rasjonelle». Denne antagelsen danner grunnlag for det mange mener er det mest forklaringskraftige paradigme i samfunnsvitenskapelig forskning, nemlig teorien om rasjonelle valg. I dette kapittelet skal vi se nærmere på denne teorien og dens filosofiske bakgrunn. Til tross for at antagelsen om rasjonalitet spiller en viktig rolle i mye samfunns og helsevitenskapelig forklaring, er den langt fra ukontroversiell. Mens noen mener den bygger på et alt for snevert bilde av hva som motiverer handling, mener andre at den er empirisk usann. Vi skal diskutere begge typer kritikk. Men før vi kommer så langt, må vi reise et mer grunnleggende spørsmål. Hva betyr det egentlig *å handle*?

Handling og grunner

Du sykler nedover veien. Når du kommer til rundkjøringa rekker du ut venstre arm. Det å rekke ut venstre arm er noe *du gjør*. Det er det vi kaller en «handling». Hvorfor gjorde du det? Det var selvfølgelig mange ulike ting som skjedde med kroppen din da du rakk ut armen. En rekke fysiske forandringer fant sted. Nerveimpulser raste nedover armen fra hjernen. Muskler trakk seg sammen. Armen beveget seg til venstre. Men slike fysiske forandringer kan jo i teorien skje uten at du gjør noe. Man kan jo tenke seg (selv om det er søkt!) at en plutselig krampe i armen fikk den til å bevege seg til venstre. Da ville vi ikke si at det å rekke ut armen var noe *du gjorde*. Det var ingen «handling». En «handling» er ikke er det samme som en kroppslig bevegelse (selv om den ofte vil involvere en kroppslig bevegelse). Så hva er det egentlig som skiller ting *du gjør* fra ting som *skjer deg* og kroppen din?

En naturlig første tanke er selvfølgelig at det som gjorde armbevegelsen din til en «handling», var at den fant sted som resultat av at du *ville* bevege den. Armbevegelsen var «viljestyrt». Den var det filosofer kaller «intensjonal». Men hva betyr egentlig det? Et innflytelsesrikt svar på dette spørsmålet ble foreslått av den britiske filosofen Elizabeth Anscombe (1919–2001). Hun argumenterte for at et definerende trekk ved intensjonal handling er at det gir mening å spørre aktøren *hvorfor* han utfører den. Det er fordi han vil ha en grunn til å utføre den. *Hvorfor* beveget armen din seg til venstre da du kom til rundkjøringa? Selvfølgelig fordi du hadde en grunn til å bevege den til venstre: Du ville signalisere til andre trafikanter at nå svinger jeg til venstre. Sett at du ved et uhell velter kaffekoppen. *Hvorfor* veltet du kaffekoppen? Du hadde ingen grunn. Det

var noe som bare skjedde. Et «uhell». Det gir derfor lite mening å spørre «hvorfor veltet du kaffekoppen?». Anscombe mente at en viktig forskjell på handlinger og kroppslige bevegelser er at handlinger alltid finner sted i lys av *grunner* aktøren har til å utføre dem (Anscombe, 1957).

Men hva er egentlig «grunner»? På spørsmål om hvorfor du rakk ut armen i rundkjøringa, oppgir du grunnen din. Du sier at du rakk ut armen «for å signalisere til andre trafikanter at du svinger til venstre». Denne beskrivelsen ser ut til å implisere to ting om deg. For det første hadde du et ønske om å signalisere til andre trafikanter at du svinger til venstre. For det andre hadde du en oppfatning om at du ved å rekke ut venstre arm oppnår å signalisere til andre trafikanter at du svinger til venstre. Et standard syn på «grunner» til handling er at de er kombinasjoner av slike ønsker og oppfatninger (på engelsk: «beliefs and desires»). Når vi oppgir grunnen vår til noe vi gjør, uttrykker vi altså noe om hva vi ønsker, hvilket mål vi har med det vi gjør, og noe om hva vi tror, at det vi gjør er et middel for å oppnå dette målet. I «hverdagspsykologiske» forklaringer bruker vi typisk informasjon om folks ønsker og oppfatninger for å trekke slutninger om hva de gjør. («Ole ønsker å betale regningen og tror at han kan gjøre det ved å legge en hundrelapp på bordet. Det Ole gjør når han legger hundrelappen på bordet, er derfor å betale regningen.») I tillegg bruker vi informasjon om hva folk gjør for å trekke slutninger om hvilke ønsker og oppfatninger de har. («Det regner og Kari slår opp paraplyen. Derfor ønsker Kari å holde seg tørr, og tror at hun kan holde seg tørr ved å slå opp paraplyen.») Vi benytter typisk begreper som «ønsket», «trodde», «ville», «mente», «prøvde» og så videre. De fleste handlingsforklaringer i samfunnsvitenskap baserer seg på slike hverdagspsykologiske begreper. Når for eksempel Max Weber skulle forklare fremveksten av moderne kapitalisme, viste han til kalvinistenes sterke ønske om å vite om de var blant de utvalgte som oppnår frelse, samt oppfatning om at verdslig suksess er et tegn på at man er blant de utvalgte. Webers hypotese var at denne typen ønske og oppfatning var sterkt medvirkende til den strenge arbeidsmoralen han mente var en forutsetning for fremveksten av moderne kapitalisme (Weber, 1995).

Men hva skal til for å kunne bruke informasjon om folks ønsker og oppfatninger for å trekke slutninger om hva de gjør (eventuelt informasjon om hva de gjør for å trekke slutninger om hvilke ønsker og oppfatninger de har)? La oss gå tilbake til eksempelet. Du kommer altså syklende nedover veien, og når du kommer til rundkjøringa rekker du ut venstre arm. De andre trafikantene observerer denne bevegelsen og trekker basert på dette en slutning om at du signaliserer at du svinger til venstre. Det de faktisk observerer er imidlertid bare at du rekker ut venstre arm. Denne observasjonen er selvfølgelig forenlig med at det du gjør kan være mye annet enn å signalisere at du svinger til venstre. Kanskje rekker du ut venstre arm for å si noe om den politiske orienteringen din eller hvilket stjernetegn du er født i. Mulighetene er mange. Men de er selvfølgelige absurde. Hvem rekker ut armen i rundkjøringer for å si noe om politisk orientering eller hvilket stjernetegn de er født i? Hvis målet er å informere folk om slike ting, er åpenbart

ikke det å rekke ut armer i rundkjøringer et særlig egnet middel. Slike «absurde» muligheter slår oss derfor ikke i det hele tatt. Det henger sammen med at vi tar for gitt at det stort sett er en rimelig sammenheng mellom det folk ønsker og det de gjør for å oppnå det de ønsker, eller mellom målene de setter seg og midlene de velger for å oppnå disse målene. Sagt på en annen måte: Vi antar at folk stort sett er *instrumentelt rasjonelle*. Denne antagelsen er nødvendig for at vi skal kunne bruke informasjon om folks ønsker og oppfatninger for å trekke slutninger om hva de gjør. Det betyr ikke at vi trenger å anta at *det* de ønsker er spesielt fornuftig, eller at *det* de gjør for å oppnå det er særlig lurt. Antagelsen om at en person er «instrumentelt rasjonell», utelukker ikke at vi kan betrakte både personens mål og handlinger som nokså «dumme». Poenget er bare at for at atferden deres overhodet skal kunne fremstå som *forståelig* for oss, som «handlinger» snarere enn tilfeldige kroppslige bevegelser, så må vi gjøre noen antagelser om de mentale tilstandene deres. Vi må anta at det de gjør i det minste for dem selv fremstår som et fornuftig middel for å oppnå noe de ønsker. Det trenger ikke å utelukke at det de ønsker og det de gjør kan fremstå som veldig ufornuftig *for oss*. Instrumentell rasjonalitet refererer til hva som er subjektivt rasjonelt, hva som fremstår som rasjonelt *sett i aktørens eget perspektiv*. Når vi skal forklare andres handlinger må vi derfor sette oss inn i *deres* perspektiv på situasjonene de befinner seg i. Vi må finne ut hvilke grunner *de* hadde for det de gjorde. Hva *de* ønsket og trodde.

Men hvilken status har egentlig prinsippet om instrumentell rasjonalitet? Dette er et dypt og vanskelig spørsmål som har vært mye diskutert i handlings- og vitenskapsteori. Noen har ment at det er en slags «empirisk lov» om menneskelig psykologi på linje med naturlovene. Akkurat som naturlovene beskriver empiriske regulariteter mellom egenskaper ved fysiske fenomener, så beskriver prinsippet om instrumentell rasjonalitet empiriske regulariteter mellom egenskaper ved psykologiske fenomener, nærmere bestemt mellom ulike typer ønsker og oppfatninger (eller «grunner») og bestemte typer handlinger. Abstrakt beskrevet kan denne psykologiske «loven» uttrykkes på følgende måte: hvis en person, *S*, ønsker å oppnå mål *m*, og tror at den beste måten å oppnå *m* på er å gjøre *x*, så vil *S* gjøre *x*. Basert på denne «loven» og en antagelse om at du for eksempel ønsker å signalisere til andre trafikanter at du svinger til venstre i rundkjøringa og tror at du kan gjøre det ved å rekke ut venstre arm, trekker vi en logisk slutning om at det *du gjør* er å signalisere at du vil svinge til venstre. Noen av dem som har forsvart dette «naturalistiske» synet på rasjonalitet, har ment at det viser at handlingsforklaringer har samme form som kausalforklaringer ellers i vitenskapen (Churchland, 1970). Ifølge et klassisk syn på vitenskapelige forklaringer er de nemlig enten «deduktivt-nomologiske» eller «induktivt-statistiske» (Hempel, 1942; Ruben, 2004). Det betyr at de alltid er basert på logisk slutning (enten deduktiv eller induktiv) fra én eller flere empiriske lover.

Det naturalistiske synet på rasjonalitet har en veldig viktig konsekvens. Dersom prinsippet om rasjonalitet bør betraktes som en empirisk hypotese (en hypotese om at vi mennesker er underlagt

en empirisk «rasjonalitetslov»), så er det mulig at dette prinsippet er usant! Uansett hvor godt bekreftet en empirisk hypotese er, så kan man nemlig aldri være 100 prosent sikker på at den er sann. Man kan for eksempel ikke være like sikker på at den er sann som man kan være sikker på at en påstand i matematikk er sann dersom den kan bevises matematisk. Empiriske hypoteser (inkludert velbekreftede hypoteser om naturlige lovmessigheter) kan i prinsippet alltid undermineres av nye observasjoner uansett hvor lite sannsynlig det er at det vil skje. Det betyr at det er teoretisk mulig at fremtidig vitenskap vil kunne empirisk falsifisere prinsippet om rasjonalitet, altså vise at rasjonalitet *ikke* spiller noen rolle i forklaring av menneskelig atferd. Konsekvensen blir i så fall at mye av grunnlaget for vår «commonsense» måte å forstå og forklare hverandre på, vil bortfalle. Istedenfor å benytte begreper vi alle behersker, som «ønsket», «trodde», «ville», «mente» og så videre, ville vi antagelig måtte gå over til å benytte begreper hentet fra fremtidig nevro- eller kognisjonsvitenskap (se f.eks. Stitch, 1983; Churchland, 1989; Bickle, 1998). Det er liten tvil om at sistnevnte vitenskaper har gjort store fremskritt når det gjelder å utvide forståelsen vår av det mennesklige sinnet. Det er derfor kanskje ikke overraskende at det naturalistiske synet på rasjonalitet har fått økt vind i seilene de siste tiårene.

Mange filosofer og vitenskapsteoretikere er imidlertid skeptiske til det naturalistiske synet. Det er fordi de mener at prinsippet om rasjonalitet ikke bør betraktes som en empirisk hypotese. Hva mener vi egentlig når vi sier ting som: «Det er *rasjonelt* av Ole å gjøre x», eller «Det er *rasjonelt* av Kari å tro at *p*»? Vi mener at Ole har «gode grunner» til å gjøre *x*, og at Kari har «gode grunner» til å tro at *p*. Dette er ikke det samme som å si at det er en «empirisk lovmessig» sammenheng mellom Ole og Karis grunner og en bestemt type handling eller oppfatning. Tvertimot bruker vi begrepet om «rasjonalitet» for å si noe om hva vi mener Ole og Kari *bør* gjøre eller *bør* tro, gitt at de har de grunnene de har. Uttrykket «bør» indikerer at sammenhengen mellom grunn og handling (eller oppfatning) er *normativ*, ikke kausal. Det betyr at prinsippet om rasjonalitet, snarere enn å beskrive en empirisk lov, beskriver en «norm» eller «regel» som mennesker stort sett overholder i tanke og handling. Til støtte for dette synet har det blitt fremholdt at i motsetning til empiriske lover tillater normer eller regler *brudd*, og rasjonalitet er åpenbart noe som kan brytes! Sett for eksempel at jeg har bestemt meg for å slutte å røyke, og vet at jeg må kaste sigarettpakken min for å klare det. Da har jeg det jeg selv mener er en veldig god grunn til å kaste sigarettpakken, og ingen god grunn til å beholde den. Likevel får jeg meg ikke til å kaste sigarettpakken. Dette er et brudd på rasjonalitet (jeg oppfører meg irrasjonelt), men så veldig uvanlig er vel denne typen brudd neppe. Og det er her forskjellen med empiriske lover ligger, mener mange: Empiriske lover kan ikke brytes. Hvis de er sanne, er de *nødvendig* sanne. Sett at objektene rundt oss plutselig begynte å sveve rett opp. Det ville ikke vært noe «brudd» på loven om gravitasjon. Det ville vist at objektene rundt oss ikke er underlagt en slik lov, altså at «loven» om gravitasjon ikke er en lov likevel. Kritikere av det naturalistiske synet mener det

faktum at brudd på rasjonalitet ikke bare er mulig men til og med nokså vanlig, tyder på at prinsippet om rasjonalitet ikke kan være en empirisk lov. Mest sannsynlig uttrykker det noe om vår menneskelige natur, nemlig at vi er en form for norm- eller regelstyrte skapninger.

Idéen om at rasjonalitet er normativt, har spilt en viktig rolle i synet på hva samfunnsvitenskap er og bør være. Noen har for eksempel ment at det viser at samfunnsvitenskap ikke kan være en kausalvitenskap på samme måte som naturvitenskapene (Winch, 1958; Dray, 1963; Taylor, 1971). Sentralt i begrunnelsen for dette synet har vært en tese om at handlingsforklaringer er «rasjonaliserende» snarere enn kausale. I motsetning til naturvitenskapelige forklaringer som typisk identifiserer årsaker, lover eller mekanismer med sikte på å forklare hvorfor fenomener skjer, så er ikke målet med rasjonaliserende forklaringer å forklare hvorfor fenomener skjer. La oss si at vi skal forklare hvorfor et maleri er vakkert, hvorfor et bestemt trekk i sjakk er forbudt, eller hvorfor kvadratroten av 2 ikke er et rasjonelt tall. Det gjør vi ikke ved å finne årsaker, lover eller mekanismer. Tvert imot finner vi *grunner* til at maleriet er vakkert, eller til at sjakktrekket er ulovlig, eller til at kvadratroten av 2 ikke er et rasjonelt tall. Handlingsforklaringer er av samme type. Når vi skal forklare hvorfor du rekker ut venstre arm i rundkjøringa, peker vi på *grunnen din*: Du ønsker å signalisere at du svinger til venstre og tror at den beste måten å gjøre det på er å rekke ut venstre arm. Denne grunnen beskriver ingen «årsak» til det du gjør. Den gir en mer fullstendig beskrivelse av det. Ikke bare rekker du ut venstre arm: Du rekker ut venstre arm *for å signalisere at du svinger til venstre.* (Det betyr selvfølgelig ikke at det ikke var fysiske årsaker til at armen din beveget seg. Poenget er at disse årsakene ikke sier noe om hvorfor du rakk ut armen din.) Når vi innser hva som var grunnen din, kan vi sette det du gjør inn i en bestemt normativ sammenheng. Dermed fremstår det som forståelig for oss. Grunner «rasjonaliserer» i denne betydningen handlingene de forklarer. Det innebærer at det konseptuelt sett egentlig ikke er noe skarpt skille mellom «grunner» og «handlinger». Det lar seg simpelthen ikke gjøre å beskrive det ene uten å beskrive det andre. Beskrivelsene av handlingen og grunnen din er logisk forbundet. Konsekvensen er at handlingsforklaringer ikke kan være kausale (Hampshire, 1959; Melden, 1961. For en god innføring i handlingsteori, se f.eks. Moya, 1990).

Mange filosofer og vitenskapsteoretikere har avvist dette resonnementet. Selv om de aksepterer at handlingsforklaringer er rasjonaliserende, mener de at dette er helt forenlig med at de likevel kan være en form for kausalforklaringer. Et berømt argument til støtte for dette synet ble foreslått av den amerikanske filosofen Donald Davidson (1917–2004) på begynnelsen av 1960-tallet. Davidson la til grunn at «årsaker» og «virkninger» er hendelser som utspiller seg på bestemte steder på bestemte tidspunkter. Slike konkrete hendelser, mente han, kan alltid lingvistisk beskrives på forskjellige måter. Men det at disse *beskrivelsene* er «logisk forbundet», utelukker ikke at det kan være en kausal sammenheng mellom *hendelsene* de beskriver. Som illustrasjon ta følgende sanne kausalsetning: «Å sette kaffekanna på kokeplata var det som

forårsaket at kaffen begynte å koke.» Beskrivelsen: «å sette kaffekanna på kokeplata» refererer til samme hendelse som beskrivelsen: «årsaken til at kaffen begynte å koke». Det betyr at vi kan erstatte den første beskrivelsen med den andre i setningen ovenfor uten at sannheten til setningen endrer seg. Da får vi: «Årsaken til at kaffen begynte å koke var det som forårsaket at kaffen begynte å koke.» Denne siste setningen er imidlertid en tautologi (triviell sann). Det er fordi beskrivelsene av årsak og virkning er logisk forbundet. (Hvis det er sant at det var en «årsak til at kaffen begynte å koke», er det nødvendigvis også sant «at kaffen begynte å koke».) Men dette betyr jo ikke at det ikke er en *kausal* sammenheng mellom de to hendelsene denne setningen beskriver! Faktisk er det rimelig, mente Davidson, å betrakte grunnen som forklarer en handling nettopp som «årsak» til denne handlingen. Når en person handler kan hun ha mange forskjellige grunner. Selv om flere av disse grunnene kan tenkes å «rasjonalisere» handlingen hennes (sette den inn i forskjellige normative sammenhenger som gjør den forståelig), er det likevel kanskje bare én av dem som faktisk *forklarer* hvorfor hun utførte den der og da. Stort sett vil det være grunnen hun selv ville ha oppgitt dersom hun hadde blitt spurt hvorfor hun utførte handlingen: Jeg gjorde det *fordi* […].[1] Men hva ligger egentlig i uttrykket «fordi» som gjorde *akkurat denne* grunnen utslagsgivende? Det er vanskelig å se at det kan være noe annet enn noe «kausalt», mente Davidson. Nærmere bestemt må det ha vært denne grunnen som *forårsaket* handlingen hennes. Til tross for at handlingsforklaringer er rasjonaliserende og normative, er det derfor gode grunner til å anta at de likevel er en form for kausalforklaringer, konkluderte han (Davidson, 1980).

I neste avsnitt skal vi se nærmere på hvordan «commonsense»-begrepet vårt om instrumentell rasjonalitet har blitt utviklet til en abstrakt, matematisk teori om rasjonelle valg.

Nyttemaksimeringsteorien om rasjonalitet

Kjernen i hverdagspsykologiske handlingsforklaringer er prinsippet om instrumentell rasjonalitet. Skjematisk uttrykker dette prinsippet at hvis en person, *S*, ønsker å oppnå mål *m*, og tror at den beste måten å oppnå *m* på er å gjøre *x*, så vil *S* gjøre *x*. For mer vitenskapelige formål er imidlertid ikke dette skjema til noe særlig hjelp. I virkelighetens verden har jo folk sjeldent bare *ett* mål. Tvert imot har de gjerne mange mål de ønsker å oppnå samtidig, noen mere enn andre. Folk

[1] Selv om det er rimelig å anta at vi vanligvis vet hvilke grunner vi har for å gjøre det vi gjør, så innebærer ikke dette synet at vi *alltid* må vite det. Noen ganger trenger vi andres hjelp for å innse hva som var de *egentlige* grunnene våre. Da vil det være disse «ubevisste» grunnene som rasjonaliserer og forårsaker handlingene våre.

rangerer derfor målene sine basert på hvor mye de «verdsetter» dem. I tillegg er det alltid mange alternative handlinger de kan utføre for å oppnå et mål, noen mer effektive enn andre. Folk rangerer derfor også alternative handlinger de kan utføre for å oppnå målene sine. I denne rangeringen er det rimelig å anta at en sentral faktor er «sannsynlighetene» de tror det er for at de ulike handlingene faktisk bidrar til at de oppnår målene de har satt seg. For å ta prinsippet om instrumentell rasjonalitet i bruk i vitenskapelig sammenheng må det derfor presiseres på en måte som ivaretar at folk rangerer både mål og midler basert på «verdier» og «sannsynligheter» (Risjord, 2014). Men dette er komplisert. Verdiene folk tillegger målene sine er jo ikke uavhengig av sannsynlighetene de tror det er for at de kan oppnå dem. Tvert imot påvirkes de av dem. Hvis du innser at det er veldig lite sannsynlig at du kan oppnå noe du ønsker, så vil det gjerne svekke styrken på ønsket ditt. Da vil du gjerne nedjustere verdien av det. I tillegg observerer vi jo vanligvis bare hva folk *gjør*. Vi observerer ikke hvilke «verdier» og «sannsynligheter» de tillegger mål og midler.

Tidlig i forrige århundre kom noen økonomer på en enkel idé for å løse disse vanskelighetene. Tenk deg at du står i en kantinekø. Du ser på hyllene med fristelser og kjenner at du ønsker sjokoladekake mer enn du ønsker kyllingsalat. Hva betyr egentlig det? Det betyr i hvert fall at dersom du får valget mellom sjokoladekake og kyllingsalat, så *velger* du sjokoladekake. I det øyeblikket du velger, er det ikke lenger et åpent spørsmål hvilke ønsker, oppfatninger, verdier og sannsynligheter du har. Tvert imot har du uttrykt en *preferanse*. Preferansen din (slik disse økonomene tenkte seg det) er simpelthen identisk med *valget* ditt. Det å ha en «preferanse» for sjokoladekake fremfor kyllingsalat betyr bare at du velger sjokoladekake snarere enn kyllingsalat dersom du får valget mellom dem. Men alle vanskelighetene er ikke over med dette. Preferansen din for sjokoladekake fremfor kyllingsalat er jo ikke den eneste preferansen du har. I tillegg har du kanskje en preferanse for eplekake fremfor sjokoladekake, for karbonadesmørbrød fremfor kyllingsalat, for iskrem fremfor eplekake og så videre. Kort sagt, du har *en mengde* preferanser. Det virker rimelig å anta at det må være noen slags «føringer» på hvordan alle disse preferansene henger sammen. Preferansene dine må være «konsistente» med hverandre, eller «ordnet». Hvis ikke blir det vanskelig å se hvordan du kan *velge* noe som helst (snarere ville du vel bare blitt helt handlingslammet!). Men hva betyr det at en mengde preferanser er «ordnet»? Et svar ble foreslått på 1940-tallet av matematikeren John Von Neumann og økonomen Oskar Morgenstern i boka *Theory of Games and Economic Behavior* (1944). I denne boka definerte de hva det betyr at en mengde preferanser er «ordnet». Slik de så det, må preferansene oppfylle noen helt bestemte betingelser. For eksempel må de være det de kalte «komplette», eller vel-definerte. Det betyr at aktøren enten foretrekker alternativ *a1* over alternativ *a2*, eller er indifferent mellom *a1* og *a2*, eller foretrekker *a2* over *a1*. I tillegg må preferansene være det de kalte «transitive». Det betyr at hvis aktøren foretrekker *a1* over *a2*, og *a2* over *a3*, så foretrekker hun *a1* over *a3* (hvis man

foretrekker epler fremfor bananer og bananer fremfor pærer, så bør man foretrekke epler fremfor pærer!). Von Neumann og Morgenstern definerte ytterligere fire (litt mer kompliserte) betingelser de mente preferanser må oppfylle for å være «ordnet». Teorien deres i denne boka kalles «aksiomatisk nytteteori». I utgangspunktet er dette bare en rent formell eller matematisk teori om en abstrakt, matematisk relasjon $\mathbb{R}$ mellom et sett med variabler {*a1, a2, ... an*}. Det er imidlertid vanlig å tolke $\mathbb{R}$ som «… *foretrekkes mer enn (eller minst like mye som)* …», altså som preferanserelasjonen. Teorien gir matematiske definisjoner av de ulike egenskapene («kompletthet», «transitivitet» osv.) som denne relasjonen må ha for å være «ordnet». Disse definisjonene utgjør «aksiomene» (grunnsatsene) i aksiomatisk nytteteori (for en klassisk fremstilling av teorien, se Luce & Raiffa, 1989).

Det viktigste elementet i Von Neumann og Morgensterns teori er imidlertid «representasjonsteoremet», som er utledet fra disse aksiomene. Representasjonsteoremet er et matematisk bevis som viser at hvis en aktør har en «ordnet mengde» preferanser definert over de tilgjengelige alternativene {*a1, a2, ... an*}, så kan det defineres en matematisk funksjon U som tilordner reelle tall til disse alternativene. Denne funksjonen kalles aktørens «nyttefunksjon». U tar handlingsalternativer {*a1, a2, ... an*} som objekter og setter tallverdier på dem. Disse tallverdiene rangerer handlingsalternativene fra høyere til lavere og kalles handlingsalternativenes «nytte» (på engelsk: «utility»). I kantine-eksempelet kan vi for eksempel tenke oss at U tilordner nytteverdi 8 til sjokoladekake, 5 til kyllingsalat, 3 til eplekake og så videre.[2] Det betyr at du foretrekker sjokoladekake over kyllingsalat, og kyllingsalat over eplekake. Det er viktig å merke seg at i denne teorien betyr ikke «nytte» det samme som det vi i dagligtalen kaller «nytte» eller «nyttig». «Nytte» er en verdi som er rent operasjonelt definert: Det at *a1* har «større nytte» enn *a2,* betyr bare at «*a1* foretrekkes over *a2*». Det impliserer ingenting substansielt om hva det er som *gjør* at «*a1* foretrekkes over *a2*», for eksempel at *a1* er «nyttigere» enn *a2*, eller at *a1* er «mer tilfredsstillende» enn *a2*, eller at *a1* «gir mer økonomisk vinning» enn *a2*, og så videre. «Nytte» er simpelthen bare en helt abstrakt verdi som beskriver aktørens preferanser. I prinsippet kan denne verdien være hva som helst – alt fra penger til godhet for miljøet (Broom, 1999).

[2] Hvordan sette tall på aktørers «nytte»? Den metoden Von Neumann og Morgenstern foreslo var å intervjue dem for å få dem til å velge mellom to forskjellige alternativer hvor det første har sikkert utfall, mens det andre har usikkert utfall. Ved å presentere dem gjentatte ganger for dette valget men systematisk variere sannsynlighetene for det usikre utfallet inntil aktørene er indifferente mellom de to alternativene, kan man sette et tall på nytten av det første alternativet. Dette tallet er simpelthen sannsynligheten som kommer ut av slike gjentatte valg. Ved hjelp av denne metoden mente de man kan konstruere en nytteskala for aktørene.

Generelt er det vanskelig å forstå hvordan man skulle kunne velge rasjonelt *mellom* alternativer, hvis det ikke hadde eksistert en slik «felles verdi» man kunne sammenligne dem ved hjelp av.

Men som vi allerede har vært inne på, så er ikke handlingsvalg bare basert på verdiene som aktørene tillegger målene sine. Rasjonelle aktører er også opptatt av sannsynlighetene for å oppnå disse målene, gitt at de utfører bestemte handlinger. Når de skal treffe valg mellom handlingsalternativer prøver de å se fremover. De vurderer hva som vil bli utfallene dersom alternativene velges. Basert på dette foretrekker de det alternativet som gir størst nytte. Teorien gir en abstrakt matematisk beskrivelse av hva som ligger i dette: Først multipliserer aktøren nytten av hvert mulige utfall av et handlingsalternativ med sannsynligheten for akkurat dette utfallet. Når det er gjort, legges summene for alle de mulige utfallene av handlingsalternativet sammen. Den summen aktøren da sitter igjen med, er handlingsalternativets «forventningsnytte» (på engelsk: «expected utility»).[3] Teorien sier at det handlingsalternativet som gir størst forventningsnytte, er det alternativet aktøren rasjonelt sett bør velge. Rasjonelle aktører «maksimerer» derfor alltid forventningsnytte. Hvorfor er det *rasjonelt* å maksimere forventningsnytte? Teorien sier at det er rasjonelt fordi hvis man *ikke* maksimerer forventningsnytte, så vil man ha «uordnede» eller inkonsistente preferanser. Da klarer man ikke å velge de beste midlene til målene man har. Ifølge aksiomatisk nytteteori er derfor en rasjonell aktør strengt tatt ikke noe annet enn en aktør med ordnede eller konsistente preferanser (altså preferanser som tilfredsstiller aksiomene i aksiomatisk nytteteori). Faktisk kan dette betraktes som en normativ begrunnelse for prinsippet om nyttemaksimering: Teorien sier at normativt sett er det å være «rasjonell» det samme som å ha ordnede eller konsistente preferanser.

Et godt eksempel på hvordan aksiomatisk nytteteori har kommet til anvendelse i samfunnsvitenskap, er det som kalles «spillteori». Spillteori er aksiomatisk nytteteori (utvidet med noen tilleggsantagelser) anvendt på en type situasjoner hvor to eller flere aktører må treffe valg mellom ulike handlingsalternativer (eller «strategier», som det kalles i spillteori), hvor forventningsnytten til en strategi for en aktør avhenger av hvilken strategi de andre aktørene velger, og hvor alle aktørene antas å være rasjonelle (nyttemaksimerende). «Aktørene» i spillteori kan være enkeltindivider, bedrifter, kommuner eller stater. Spillteori anvendes for å analysere og forutsi hvordan slike aktører vil oppføre seg under nærmere angitte betingelser (Hovi, 2008). Et av de mest berømte og omdiskuterte «spillene» går under navnet «fangens dilemma». La oss

[3] Dette forutsetter et «kardinalt» nyttebegrep, hvor nyttefunksjonen er unik opptil positiv lineær transformasjon. Med et kardinalt nyttebegrep kan aktørens nytte tallfestes på mange forskjellige måter, mens den tallmessige differansen mellom to alternativer sier noe om hvor mye mere eller mindre nytte de har i forhold til hverandre (f.eks. dobbelt så mye, tre ganger så mye osv.).

tenke oss en situasjon hvor to forbrytere blir arrestert av politiet. De blir deretter plassert i separate avhørsrom uten mulighet til å kommunisere med hverandre. Her må de bestemme seg for om de vil tilstå forbrytelsen de har begått eller holde tett. De opplyses om at hvis begge tilstår, får de 5 års fengsel hver; hvis derimot begge holder tett, sitter politiet likevel på nok bevis til at de får 1 års fengsel hver; mens hvis én tilstår og den andre holder tett, vil den som tilstår frigis, mens den som holder tett får 10 års fengsel.

Fordi de to forbryterne er rasjonelle, predikerer spillteori at begge vil ende opp med å tilstå. Det er fordi de vet at hvis de holder tett og den andre tilstår, så får de 10 års fengsel, noe som ville være det aller verste utfallet (minst forventningsnytte). Hvis de derimot tilstår, vil én av to ting skje: Enten går de fri (hvis den andre holder tett), noe som ville være det beste utfallet (mest forventningsnytte), eller så får de 5 års fengsel (hvis den andre tilstår), noe som ville være det nest beste utfallet (nest mest forventningsnytte). De vil derfor begge resonnere at det å tilstå er den beste strategien *uansett* hva den andre velger å gjøre (den strategien som gir høyest forventningsnytte). Det å tilstå sies derfor å være den «dominante» strategien i dette spillet. (Men sett i ett annet perspektiv kan jo dette virke nokså rart! Et utfall som åpenbart ville vært bedre for begge, er jo at begge holdt tett og dermed bare fikk 1 års fengsel hver. Hvordan kan det være *rasjonelt* å velge noe som gir *et dårligere* utfall enn alternativet?)

Det generelle dilemma som denne typen situasjon skaper, er mellom å samarbeide og dermed høste gevinstene av et slikt samarbeid, og *ikke* å samarbeide og dermed unngå kostnadene et slikt samarbeid medfører for en selv. (I historien ovenfor er disse kostnadene risikoen for 10 års fengsel.) Grunnen til at mange samfunnsforskere har vært spesielt interessert i «fangens dilemma», er nettopp at mange ulike typer sosiale situasjoner kan se ut til å skape akkurat denne typen dilemma. Aktører som opptrer rasjonelt i disse situasjonene, kan derfor se ut til å ha et insentiv til å være såkalte «free riders». Det betyr at de har insentiv til å velge *ikke* å samarbeide, fordi samarbeid medfører kostnader for dem selv, og fordi de kan høste gevinstene av andres samarbeid uten selv å samarbeide. Hvis alle andre vaksinerer seg, hvorfor trenger *jeg* gjøre det? Hvis alle andre stemmer, hvorfor skal *jeg* gidde å stemme? og så videre. Dersom alle opptrer «rasjonelt», kan utfallet se ut til å bli verre for alle: Ingen vaksinerer seg, ingen stemmer og så videre.

I praksis er selvfølgelig ikke «free riding» et så stort problem som spillteori kanskje skulle lede en til å tro. Faktisk er det mye som tyder på at folk flest oftere velger å samarbeide enn *ikke* å samarbeide i fangens dilemma situasjoner. Er dette et problem for teorien? Her er det ulike syn. Noen mener at det ikke er det. Problemet skyldes at aktørenes nyttefunksjoner i fangens dilemma er feilspesifiserte. Folk kan være motiverte av lojalitet eller empati, noe som er fullt forenlig med at de er nyttemaksimerende. Andre mener problemet stikker dypere. Det skyldes at spillteori ikke tar hensyn til at folk ofte er påvirket av sosiale normer som får dem til å velge og samarbeide til

tross for at samarbeid ikke maksimerer nytten deres. Vi skal komme tilbake til noen ulike syn på denne problemstillingen i siste avsnitt, men først skal vi se nærmere på hvordan aksiomatisk nytteteori danner grunnlag for det mange samfunnsforskere mener er det mest forklaringskraftige paradigme i samfunnsvitenskap, nemlig «teorien om rasjonelle valg».

Teorien om rasjonelle valg

Von Neumann og Morgenstern betraktet aksiomatisk nytteteori primært som en normativ og prediktiv teori. Teorien forteller oss hvordan aktører *bør* velge for å være «rasjonelle», og den gjør oss i stand til å utlede prediksjoner om handlingene deres, gitt at forutsetningen om rasjonalitet er oppfylt. Hensikten med teorien, slik de så det, var ikke å gi psykologiske forklaringer på disse handlingene. Teorien påstår således ingenting om hva som faktisk motiverer folk. Von Neumann og Morgenstern hadde heller ingen formening om at folk flest *er* rasjonelle i den betydningen teorien definerer. Teorien sier bare at *dersom* vi antar at aktørene vi studerer er «rasjonelle» og vi vet hvilke preferanser de har, så kan vi forutsi hva de vil gjøre. De hadde derfor et «instrumentalistisk» syn på aksiomatisk nytteteori. Det er et syn som også har vært vanlig blant en del økonomer. (Ofte knyttes det til den nobelprisvinnende økonomen Milton Friedman (1953), selv om dette er omdiskutert.)

Gir aksiomatisk nytteteori en tilfredsstillende definisjon av hva det betyr å være en rasjonell aktør? Det er gode grunner til å være skeptisk til det, i hvert fall hvis vi anser det som et rimelig krav til en slik definisjon at den skal være noenlunde psykologisk realistisk. Selv om vi aksepterer tesen om at preferansene til rasjonelle aktører må kunne representeres ved hjelp av en nyttefunksjon, så er det vanskelig å se at dette alene kan være tilstrekkelig til at vi vil betrakte dem som «rasjonelle». Problemet er, for å si det med den norske vitenskapsteoretikeren Jon Elster, at definisjonen av rasjonalitet i aksiomatisk nytteteori er alt for abstrakt (eller «tynn», som han uttrykker det). Den tar hverken hensyn til hvordan preferanser dannes eller hvordan de endres over tid, bare hvordan de formelt (logisk-matematisk) henger sammen (Elster, 1983; 1985; 2007). Det virker imidlertid rimelig at det må være noen føringer på måten aktører danner og endrer preferanser på, som har betydning for om vi er villige til å betrakte dem som rasjonelle. Litteraturen flommer over av ulike eksempler på dette. For å ta en variant av ett av Elsters egne eksempler: Sett at Fredrik ønsker å ta livet av den irriterende naboen sin og vurderer om han enten skal helle cyanid i kaffen hans eller stikke nåler i en dukke som forestiller ham. Fredrik velger å lage en dukke av naboen og stikke nåler i den. Det er i prinsippet ingenting i veien for at Fredriks preferanser kan representeres ved hjelp av en nyttefunksjon. Likevel opplever vi ham ikke som spesielt rasjonell: Hvis Fredrik ønsker å drepe naboen, så fremstår ikke det å stikke nåler i en dukke som et særlig rasjonelt valg av fremgangsmåte!

Det som kalles «teorien om rasjonelle valg» i samfunnsvitenskap er aksiomatisk nytteteori supplert med forskjellige «føringer» på aktørenes preferanser, noe som er ment å gjøre teorien mer realistisk og dermed unngå denne typen moteksempler. Elsters eksempel ovenfor illustrerer for eksempel at selv om en aktørs preferanser kan tilfredsstille aksiomene i aksiomatisk nytteteori (slik at de kan representeres ved hjelp av en nyttefunksjon), så vil ikke aktøren være «rasjonell» dersom oppfatningene hans (sannsynlighetene han tilordner handlingsalternativene) ikke er «velbegrunnede» i forhold til den informasjonen han besitter. Et vanlig syn i teorien om rasjonelle valg er at slike oppfatninger (eller subjektive sannsynligheter) er velbegrunnede dersom de er «oppdaterte» i forhold til den tilgjengelige informasjonen. (Det finnes en matematisk formel i statistikk kalt *Bayes teorem,* som antas å uttrykke hvordan man bør oppdatere slike sannsynligheter.) I tillegg til at aktørens oppfatninger bør være oppdaterte i forhold til informasjonen han besitter, så bør også mengden informasjon han besitter være «optimal» for treffe et rasjonelt valg. Det er imidlertid ikke så enkelt å si hva dette betyr. Et eksempel fra den norske samfunnsøkonomen Leif Johansen illustrerer noe av problemet (Johansen, 1983). Sett at du er på blåbærtur og målet er å plukke mest mulig blåbær, men du befinner deg i et området du kjenner dårlig. Du kan selvfølgelig slå deg ned på den første tuen du kommer over, men det virker litt dumt. Det kan jo hende at noen enda bedre tuer befinner seg rett rundt neste sving. Samtidig bør du ikke bruke for lang tid på å finne de beste tuene heller. Risikoen med det er jo at det ikke blir noe tid igjen til å plukke. Så hvor mye tid er det «optimalt» å bruke på innhente informasjon? Det kan du ikke vite nettopp fordi du mangler informasjon! I slike situasjoner er det beste du kan gjøre å velge det som er «bra nok» basert på rimelig skjønn, selv om det er fullt mulig at du kunne ha funnet noe enda bedre hvis du hadde brukt litt mer tid og krefter.

I teorien om rasjonelle valg er aktørenes informasjon og ressurser eksempler på «føringer» som innarbeides i aktørenes nyttefunksjon. I tillegg baserer teorien seg vanligvis på en mer substansiell tolkning av begrepet om «nytte» enn den operasjonelle tolkningen som Von Neumann og Morgenstern la til grunn (Bermúdez, 2009). Det betyr at påstanden om at *a1* har «større nytte» enn *a2,* ikke bare tolkes som at «*a1* foretrekkes over *a2*» (noe som strengt tatt er forenlig med at «nytte» kan være en hvilken som helst verdi), men at *a1* er mer «tilfredsstillende» enn *a2*, mer «ønskverdig», eller kanskje «fremmer mer velferd» enn *a2*. Kort sagt, *det* rasjonelle aktører maksimerer er en konkret psykologisk verdi. (I den engelske litteraturen brukes ofte begreper som «satisfaction», «desirability» eller «welfare» om denne verdien.) Mot denne bakgrunnen er det vanlig å betrakte teorien om rasjonelle valg ikke bare som et redskap for prediksjon av aktørers handlinger. Like mye betraktes det som en teori om hva som faktisk forklarer disse handlingene. Mange tilhengere av teorien har for eksempel ment at «preferanser» bør forstås som kombinasjoner av ønsker og oppfatninger – altså psykologiske tilstander – som

forårsaker aktørenes handlinger i tråd med Donald Davidsons handlingsteori (Elster er én av dem).

Til tross for at teorien om rasjonelle valg har hatt stor betydning i samfunns- og helsevitenskap, er den langt fra ukontroversiell. En innvending som ofte dukker opp, er at teorien er absurd fordi ingen virkelige mennesker går rundt og «regner» ut forventningsnytte på den måten teorien forutsetter. Hvem driver og multipliserer nytte med sannsynligheter når de for eksempel skal velge hva de skal ha til middag, eller hvilken film de skal se på kino? Ingen, selvfølgelig. Men dette er en misforstått kritikk. Teorien om rasjonelle valg forutsetter ikke at folk bevisst og eksplisitt går rundt og «regner» ut forventningsnytte. Det den sier er at hvis du er rasjonell, så kan preferansene dine *beskrives* ved hjelp av en nyttefunksjon. Det betyr at du velger *som om* du maksimerer forventningsnytte. Generelt er det mye vi gjør «automatisk» uten å tenke over det, som kan gis matematiske beskrivelser. Alt fra å trekke enkle logiske slutninger (noe vi vanligvis gjør helt automatisk) til å ta imot baller som blir kastet mot oss, krever former for beregning. Selv om vi ikke utfører disse beregningene bevisst og eksplisitt, er det grunn til å tro at de involverer prosesser i hjernen som kan gis matematiske beskrivelser. La oss derfor se nærmere på noen mer interessante kritikker av teorien.

Normer, forpliktelser og forklaringskraft

Kritikken av nyttemaksimeringsteorien om rasjonalitet kan deles inn i en normativ og en empirisk kritikk. Mens den normative kritikken retter seg mot teoriens begrep om rasjonalitet og menneskelig aktørskap, retter den empiriske kritikken seg mot teoriens forklaringskraft. Det finnes mange ulike eksempler på begge typer kritikk, og vi kan bare diskutere et lite utvalg her. En betydelig del av den normative kritikken har kommet fra teoretikere som selv arbeider innenfor rammene av nyttemaksimeringsteorien, og har dreid seg om aksiomene den baserer seg på. Sentralt i denne kritikken har vært spørsmål om hvorvidt disse aksiomene gir en rimelig beskrivelse av rasjonalitet. Dette er en type kritikk (ofte av nokså teknisk karakter) som har bidratt til forskjellige justeringer av teorien, men også til viktige videreutviklinger av den. For våre formål er kanskje den mest interessante kritikken den som hevder at teorien bygger på en feilaktig antagelse om menneskelig aktørskap. Sentralt i denne kritikken er en påstand om at teorien impliserer at rasjonelle mennesker nødvendigvis er «egoister». Det skyldes at den forutsetter at rasjonelle aktører alltid er drevet av nyttemaksimeringsmotiver. Virkelige mennesker er imidlertid ofte drevet av helt andre motiver uten at det trenger bety at de er «irrasjonelle». For eksempel kan de være drevet av motiver om å hjelpe andre mennesker selv om det påfører dem selv store kostnader. Mange gir penger til nødhjelp uten at de får noe igjen for det. De investerer tid og krefter i miljøarbeid med tanke på velferden til fremtidige generasjoner.

Noen ofrer til og med eget liv for å redde livet til andre mennesker. Kort sagt handler virkelige mennesker ofte altruistisk. Problemet med nyttemaksimeringsteorien er at den impliserer at rasjonelle mennesker *ikke kan* handle altruistisk. Men dette er usant.

Formulert på denne måten vil imidlertid tilhengerne av teorien ha et enkelt svar på kritikken. De vil påpeke at teorien på ingen måte utelukker at rasjonelle mennesker kan handle altruistisk. Hvis en person verdsetter å gi penger til nødhjelp *mere* enn hun verdsetter å la være, eller verdsetter å redde et annet menneskes liv *mere* enn hun verdsetter å redde sitt eget, så vil det være rasjonelt for denne personen å gi penger til nødhjelp eller ofre sitt eget liv. En slik person vil maksimere forventningsnytte på vanlig måte, helt i tråd med hva nyttemaksimeringsteorien påstår om «rasjonelle aktører». Er dermed problemet unngått? I den klassiske artikkelen «Rational Fools: A Critique of the Behavioral Foundations of Economic Theory» (1976) setter den nobelprisvinnende økonomen og filosofen Amartya Sen fingeren på det som etter manges mening er det dypere problemet med nyttemaksimeringsteorien om rasjonalitet. Det dypere problemet er ikke at teorien impliserer at rasjonelle mennesker ikke kan gi penger til nødhjelp eller ofre livet for å redde andre. Det består snarere i teoriens forklaring på *hvorfor* de gjør det. Ifølge teorien gjør de det fordi de «verdsetter» det mere enn alternativene. Problemet er teoriens syn på hva det betyr å verdsette noe mere enn noe annet. Ifølge teorien betyr det at personen tillegger handlingen større «nytte» enn alternativene. «Nytte» er imidlertid nødvendigvis knyttet til et individ. «Nytte» er alltid *et bestemt individs* «nytte». Med det mener Sen at hvis jeg velger å beskytte barna mine mot skade fremfor meg selv, så er det fordi nytten *min* av å beskytte meg selv mot skade fremfor barna mine er mindre enn nytten *min* av å beskytte barna mine mot skade fremfor meg selv. Med andre ord: *Grunnen* til at jeg beskytter barna mine, er at nytten *for meg* av at de ikke påføres skade, er stor, ikke at skade på dem reduserer *deres* nytte! Essensen i Amartya Sens innvending er at mennesker noen ganger helt rasjonelt gjør ting til tross for at nytten deres av å avstå kan være større enn nytten deres av å gjøre dem, simpelthen fordi de opplever at de *bør* gjøre dem. Det skyldes at *grunnene* vi mennesker har til å handle, ofte involverer normer, prinsipper eller det han kaller «forpliktelser», snarere enn preferanser.

Sen illustrerer forskjellen på disse to typene motivasjon med en berømt historie: To gutter, la oss kalle dem Ole og Petter, finner to epler, ett stort og ett lite. Ole sier til Petter: «Du kan velge hvilket du vil ha først.» Petter kaster seg glupsk over det største eplet. Ole blir synlig oppbrakt over Petters oppførsel, noe som får Petter til å spørre: «Men hvilket eple ville *du* ha valgt hvis du hadde fått velge først?» Ole svarer at han selvfølgelig ville ha valgt det minste eplet. Da sier Petter: «Men hva er det da du klager over? Det var jo det minste eplet du fikk!» Her antar Petter at handlinger alltid er motivert av preferanser. Fordi Ole sier at han ville ha valgt det minste eplet, så antar Petter at det var dette eplet han foretrakk. Dermed har han ingenting å klage over. Han fikk jo det eplet han foretrakk! Sens poeng er imidlertid at vi føler at Ole har rett til å klage på

Petters oppførsel: Selv om det var det største eplet Petter *foretrakk,* så burde han ha tatt hensyn til Ole *og valgt* det minste! Det å glupsk bare kaste seg over det største eplet er veldig usosialt. Petter fremstår som det Sen kaller en «sosial idiot». Og det er her det dypere problemet med teorien om rasjonelle valg ligger, mener Sen: Den gjør rasjonelle mennesker til *sosiale idioter*!

Holder denne innvendingen? Til forsvar for nyttemaksimeringsteorien kan det kanskje påpekes at i historien om Ole og Petter er ikke «gevinsten» nødvendigvis begrenset til eplene. Det å «ofre sine egne behov» og tilby det største eplet til Petter kan *også* være en del av Oles gevinst. Det er fordi det kan tenkes å gi ham en følelse av «selvtilfredshet», en positiv psykologisk opplevelse av å være et «godt menneske». Denne opplevelsen vil endre Oles nyttekalkyle. Hvis vi tar Oles nytte av å velge det minste eplet og legger til nytten denne «selvtilfredsheten» gir ham, så vil summen av nytten hans av å velge det minste eplet bli større enn nytten hans av å velge det største. Med andre ord: Ole maksimerer nytte akkurat som Petter. Han har bare en annen nyttefunksjon enn Petter. Denne typen respons reiser spørsmål om hvorvidt det er korrekt at motivene som forklarer hva vi velger, alltid inneholder en eller annen form for positiv psykologisk opplevelse. Det er i bunn og grunn et empirisk spørsmål. Noen har tatt forskning i moderne biologi og psykologi til inntekt for at dette ikke er tilfelle. Blant annet har det blitt argumentert for at det er gode *evolusjonære grunner* til å anta at mennesker må ha utviklet en direkte form for motivasjon til å handle altruistisk, altså en form for motivasjon som ikke motiverer *via* noen positiv psykologisk opplevelse (Sober & Wilson, 1998; Stitch, Doris & Roedder, 2010). Vi kan ikke forfølge denne interessante diskusjonen videre her. Isteden skal vi avslutte med å se på noen eksempler på den empiriske kritikken av nyttemaksimeringsteorien. En viktig del av denne kritikken kommer fra det som kalles «atferdsøkonomi».

Atferdsøkonomi er en retning innen moderne økonomi som kombinerer økonomisk teori med eksperimentell psykologi (særlig kognitiv psykologi). Målet er å utvikle mer realistiske teorier om menneskelige beslutningsprosesser. Mye av forskningen i atferdsøkonomi har fokusert på eksperimentell testing av nyttemaksimeringsteorien om rasjonalitet. Resultatene fra denne forskningen kan tyde på at folks valg ofte avviker ganske mye fra det denne teorien predikerer. I ett eksperiment (Clark & Sefton, 2001) testet man for eksempel ut en såkalt «dynamisk» variant av fangens dilemma på forsøkspersoner. I dynamiske spill gis spillerne anledning til å observere og reagere på de andre spillernes valg. Etter at den første spilleren har valgt samarbeid eller ikke-samarbeid velger den andre spilleren (basert på full informasjon om førstespillerens valg) enten å samarbeide eller ikke. I denne typen eksperimenter brukes ofte penger som mål på nytte. (Det vil si at gevinst og kostnad ved de ulike strategiene er bestemt av hvor mye penger man kan vinne eller tape.) Som vi husker er den dominante strategien i fangens dilemma *ikke-samarbeid*. Eksperimentet viste imidlertid at i mer enn 50 prosent av tiden valgte førstespilleren tvert imot å

samarbeide. Andrespillers respons på førstespiller som valgte samarbeid, var å samarbeide i mer enn 30 prosent av tiden.

Et annet eksperiment illustrerer ganske godt Amartya Sens poeng. Her lot man forsøkspersonene spille et enkelt spill som kalles «diktator». I «diktator» gis førstespiller et valg mellom to måter å fordele en sum penger på mellom seg selv og en annen spiller. Den andre spilleren må akseptere førstespillers valg. I eksperimentet fikk førstespiller et valg mellom enten å gi andrespiller $ 400 og beholde $ 400 selv, eller gi andrespiller $ 750 og beholde $ 400 selv (Charness & Rabin, 2002). Gevinsten for førstespiller var altså den samme uansett hvilke av alternativene han valgte. Forskerne bak eksperimentet antok at forskjellen på disse valgene derfor ville måtte reflektere noe annet enn førstespillers nytte. Faktisk viste det seg at hele 69 prosent av forsøkspersonene valgte det andre alternativet, altså å gi andrespiller $ 750 og beholde $ 400 selv. (For en god diskusjon av betydningen av dette og andre eksperimenter for spillteori, se Guala, 2006.) Akkurat som Ole i Sens eksempel, verdsatte forsøkspersonene å gi andre fordeler uten at de fikk noe igjen for det. Det finnes mange varianter av denne typen eksperimenter. En vanlig oppfatning i atferdsøkonomi er at de viser at preferanser varierer mye fra situasjon til situasjon, og at *hva* en person velger ikke bare er avhengig av abstrakt nytte (slik den klassiske nyttemaksimeringsteorien forutsetter), men også er påvirket av kontekst og omstendigheter. Mye atferdsøkonomi har dreid seg om å studere nærmere hvordan slike kontekstuelle faktorer påvirker folks beslutningsprosesser.

Konklusjon

Dersom denne kritikken av nyttemaksimeringsteorien er korrekt, hva er konklusjonen? Bør samfunns- og helsevitenskapene oppgi teorien om rasjonelle valg? Kritikken gir ikke grunnlag for en slik konklusjon. Denne teorien bør betraktes som en «modell». Det betyr at den er basert på forenklinger av fenomenene for å få frem viktige aspekter ved dem. Det kan være liten tvil om at den opererer med en form for «idealiserte» aktører. I virkeligheten er menneskelig motivasjon og rasjonalitet høyst sannsynlig mer komplekse fenomener enn det denne modellen klarer å fange inn. Men det er ikke noen drepende kritikk av en modell i vitenskap (enten det er en modell i fysikk eller i samfunnsvitenskap) å påpeke at fenomenene som modellen beskriver, er mer komplekse i virkeligheten enn i modellen. Det ligger i sakens natur. Hvis modellen likevel gir gode forklaringer eller prediksjoner i mange nok situasjoner, så fungerer den etter hensikten. I hvilken grad nyttemaksimeringsteorien gir grunnlag for forklaring eller prediksjon i samfunns- og helsevitenskapene, er et åpent empirisk spørsmål. Det kan bare vises gjennom anvendelse. Men det kan være liten tvil om at den har blitt anvendt med stor suksess i veldig mange sammenhenger. «Modeller» kan dessuten videreutvikles og raffineres på ulike måter for å gjøre

dem mer realistiske (noe man har forsøkt å gjøre med nyttemaksimeringsteorien i for eksempel atferdsøkonomi). En rimeligere konklusjon er derfor at nyttemaksimeringsteorien om rasjonalitet fremdeles har en naturlig plass som en av de viktigste teoriene i studiet av sosiale fenomener. Det utelukker ikke at man bør reflektere over begrensningene ved den.

Litteratur

Anscombe, G. E. M. (2000). *Intention*. Cambridge, Massachusetts: Harvard University Press.

Bermúdez, J. L. (2009). *Decision Theory and Rationality*. New York: Oxford University Press.

Broom, J. (1999). Utility. I J. Broom, *Ethics out of Economics* (s. 19–28). Cambridge: Cambridge University Press.

Bickle, J. W. (1998). *Psychoneural Reduction*. *The New Wave*. Cambridge, MA: The MIT Press.

Charness, G. & Rabin, M. (2002). Understanding Social Preferences with Simple Tests. *The Quarterly Journal of Economics*, *117*(3), 817–869.

Churchland, P. M. (1970). The Logical Character of Action Explanations. *Philosophical Review*, *79*(2), 214–236.

Churchland, P. M. (1989). A Neurocomputational Perspective: The Nature of Mind and the Structure of Science. Cambridge, MA: The MIT Press.

Clark, K & Sefton, M. (2001). The Sequential Prisoner's Dilemma: Evidence on Reciprocation. *The Economic Journal*, *111*(468), 51–68.

Davidson, D. (1980). Action, Reasons, and Causes. I D. Davidson, *Essays on Actions & Events* (s. 3–19). New York: Oxford University Press.

Dray, W. (1957). *Laws and Explanations in History*. Oxford: Oxford University Press.

Elster, J. (1983). *Sour Grapes. Studies in the Subversion of Rationality*. New York: Cambridge University Press.

Elster, J. (1985). The Nature and Scope of Rational-Choice Explanations. I E. LePore og B. McLaughlin (red.), *Actions and Events: Perspectives on the Philosophy of Donald Davidson* (s. 60–72). Oxford: Blackwell Publishers.

Elster, J. (2007). *Explaining Social Behavior. More Nuts and Bolts for the Social Sciences*. Cambridge: Cambridge University Press.

Friedman, M. (1953). The Methodology of Positive Economics. I M. Friedman, *Essays in Positive Economics* (s. 3–44). Chicago: University of Chicago Press.

Guala, F. (2006). Has Game Theory Been Refuted? *The Journal of Philosophy 103*(5), 239–263.

Hampshire, S. (1959). *Thought and Action*. New York: Viking Press.

Hempel, C. (1942). The Function of General Laws in History. *Journal of Philosophy 39*(2), 35–48.

Hovi, J. (2008). *Spillteori. En innføring*. Universitetsforlaget.

Johansen, L. (1983). *Opptak til sosialøkonomien*. Universitetsforlaget.

Luce, R. D. & Raiffa, H. (1989). *Games and Decisions. Introduction and Critical Survey*. New York: Dover Publications, INC.

Melden, A. I. (1961). *Free Action*. London: Routledge & Kegan Paul.

Moya, C. J. (1990). *The Philosophy of Action. An Introduction*. Cambridge: Polity Press.

Risfjord, M. (2014). *Philosophy of Social Science*. New York: Routledge, Taylor & Francis.

Ruben, D.-H. (1993). Introduction. I D.-H. Ruben, *Explanation* (s. 1–16). New York: Oxford University Press.

Sen, A. (1977). Rational Fools: A Critique of the Behavioral Foundations of Economic Theory. *Philosophy & Public Affaires*, *6*(4), 317–344.

Sober, E. & Wilson, D. S. (1998). *Unto Others: The Evolution and Psychology of Unselfish Behavior*. Harvard University Press.

Stitch, S. (1983). *From Folk Psychology to Cognitive Science. The Case against Belief*. Cambridge, MA: The MIT Press.

Stitch, S., Doris, J. M. & Roedder, E. (2010). Altruism. I J. M. Doris (red.), *The Moral Psychology Handbook* (s. 147–206). Vol. 1. Oxford: Oxford University Press.

Taylor, C. (1971). Interpretation and the Sciences of Man. *Review of Metaphysics*, *25*, 1–51.

Von Neumann, J. & Morgenstern, O. (1972). *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Weber, M. (1995). *Den protestantiske etikken og kapitalismens ånd*. Oslo: Pax forlag.

Winch, P. (1958). *The Idea of a Social Science*. London: Routledge & Kegan Paul.