24 Vår fria vilja?[1]

Den fria viljan hör till de klassiska frågeställningar som plågat generationer av filosofer och andra tänkare. Vad som gör problemet så besvärligt är en krock mellan följande båda ytterst naturliga och närliggande tankegångar.

Å ena sidan synes det moraliskt nödvändigt att tillerkänna människan en fri vilja. I annat fall rubbas vissa av grunderna för vårt rättstänkande: vari ligger det moraliskt rimliga i att straffa en person för gärningar som hon inte gjort av fri vilja? En än värre konsekvens som förnekandet av den fria viljans existens riskerar att leda till är en allmän apati: om jag inte har någon fri vilja har jag väl överhuvudtaget inget inflytande över mina gärningar, och vad är det då för poäng med att jag anstränger mig att leva laglydigt, att vara god mot mina medmänniskor, eller att överhuvudtaget stiga upp om morgonen?

Å andra sidan synes det svårt eller rentav omöjligt att finna rum för människans fria vilja i en materialistisk världsbild. Om materien styrs av deterministiska naturlagar, så bestäms ju mina kommande handlingar entydigt av tillståndet hos de elementarpartiklar, atomer och molekyler som bygger upp min hjärna och dess omgivningar, och något utrymme därutöver för någon fri vilja att gripa in står ej att finna. Att förse naturlagarna med ett inslag av slump, t.ex. via kvantmekaniken, båtar föga, ty inte blir väl min vilja friare av att regleras av något slags roulettehjul i min hjärna (eller annorstädes)?

Vad som ändå tycks tala för att vi har en fri vilja är vår starka upplevelse av att, i allehanda beslutssituationer, faktiskt kunna välja fritt mellan olika alternativ. Jag inledde förra meningen med ”Vad som ändå”, men kunde lika gärna ha valt att skriva ”Något som likväl”. Jag tog spårvagnen till jobbet i morse, men kunde lika gärna ha valt att ta cykeln. Jag drack lättöl till middagen igår, men kunde lika gärna ha valt att tacka ja till det vin som erbjöds. Eller?

Det verkar svårt att leda i bevis att dessa val är resultatet av min fria vilja, men den intuitiva känslan av att så är fallet är mycket stark. Författaren och IT-filosofen Clay Shirky frågar sig i sitt bidrag ”Free will is going away” till den infallsrika antologin *What is Your Dangerous Idea?* (Brockman, 2007) hur länge denna känsla kommer att vara. Om en man begått en serie bestialiska våldsbrott, om det visat sig att dessa skedde under

[1] Nyskrivet kapitel för denna bok.

inflytande från en hjärntumör, om tumören opererats bort, och om mannens våldsamma tendenser därvid gått upp i rök, kan han då hållas ansvarig för våldsbrotten? De flesta skulle här hålla med om att svaret är nej: det var inte mannens fel, utan tumörens, att brotten ägde rum.[2] Mannens fria vilja var helt enkelt satt ur spel. Shirky fruktar att i takt med att vi blir allt skickligare på att med allt bättre precision finna orsaker till en människas handlingar – orsaker som kan finnas t.ex. i kemiska obalanser i hjärnan, i traumatiska händelser i hennes barndom, eller i någon defekt gen – så kommer den fria viljans utrymme att krympa för att till slut kanske till och med försvinna helt. Denna utveckling måste vi, menar Shirky, närma oss med eftertanke och framförhållning, för att undvika att vi står utan moralisk kompass när den väl inträffar.

En annan utmaning mot vår föreställning om den fria viljan är en serie experiment av den nyligen bortgångne geniale amerikanske neurofysiologen Benjamin Libet. Denne detekterade hos sina försökspersoner elektrisk aktivitet i hjärnan svarande mot påbörjandet av förbreredelser för en viss handling, upp till en halv sekund *innan* försökspersonen medvetet överväger handlingen (se t.ex. Libet, 1985, eller Dennett, 2003). Detta kan ses som en indikation på att vi överskattar betydelsen av medvetna beslut. Senare experiment pekar i samma riktning. Försökspersoner som ombeds lyfta valfritt pekfinger, varpå deras hjärnor stimuleras elektriskt på ett sätt som framtvingar lyftandet av det ena av dem, rapporterar likväl att de valde detta finger av fri vilja. Wegner (2003) reflekterar över resultaten av dessa och andra försök; han konstaterar att våra hjärnor är skickliga på att skapa illusionen av medvetna viljeyttringar.

Många invecklade eller rent ut sagt konstlade försök har gjorts att förena idén om fri vilja med ett av naturlagar styrt universum, men jag föredrar hur Douglas Hofstadter summariskt behandlar frågan i sin nya bok *I Am a Strange Loop* (2007).[3] Han menar att insisterandet på *fri* vilja är ett missförstånd. Visst har jag en vilja, men (och här ansluter jag mig till Hofstadters argumentation) den är knappast fri, då mina val ju ständigt stöter på hinder: margarinet är slut, jag glömde förrådsnyckeln i lägenheten och jag sak-

[2] Se Zaremba (2006) för en läsvärd och kritisk diskussion kring det svenska rättssystemets sätt (inte helt i linje med andra länders) att förhålla sig till liknande fall.

[3] I förordet till denna underbara och djupt personliga behandling av frågan om mänskligt medvetande bekänner Hofstadter att en av de saker som drev honom att skriva boken var ett missnöje med hur hans klassiska debutbok om samma ämne, *Gödel Escher Bach* från 1979, togs emot – vilket framstår som en smula paradoxalt då den senare ju rimligtvis är att betrakta som en av de största populärvetenskapliga sensationerna på många årtionden. Hur som helst är *I Am a Strange Loop* mycket mer än blott en ompaketering av idéerna i *Gödel Escher Bach*.

nar förmåga såväl att flyga som att bevisa Riemannhypotesen. Livet är en labyrint vars väggar utgörs av dessa bivillkor och frustrationer. En del av bivillkoren dyker upp i min egen hjärna i form av starka eller ibland rentav oemotståndliga känslor: kättja, sockersug, impulsen att klia på gårdagens myggbett och det maniska undvikandet av trottoarens betongskarvar. Dessa drivkrafter finns där jämte andra mer komplexa sådana: mina ambitioner och förhoppningar om att vara välförberedd på torsdagens seminarium, om att färdigställa denna bok, om att åter nå ned till min personliga rekordtid på Göteborgsvarvet, om att vara snäll mot vänner och familj, om att betala av bostadslånet och om att leva någorlunda miljövänligt. Vad är då min vilja annat än det stycke neurologisk informationsbehandling som väger samman alla dessa villkor, drivkrafter och önskningar för att nå fram till ett entydigt svar om vad jag skall göra härnäst? Och vad i hela fridens namn skulle det innebära att min vilja är *fri* – att jag, även sedan sammanvägningen resulterat i att jag skall ta spårvagnen, *ändå* kunde ha valt cykeln? Vad vore poängen med det? Jag är glad åt att ha en vilja, men varför låta mig störas av det faktum att denna är uppbyggd av nervknippen och molekyler som liksom all annan materia har att rätta sig efter naturlagarna? Jag vågar påstå att det går att vänja sig vid denna tanke utan att ens moraluppfattning eller tro på det meningsfulla i att leva slås i spillror.

*

Fri eller inte är den funktion hos våra hjärnor som vi kallar viljan värd att studera närmare, och kanske är det mer konstruktivt att försöka förstå den utan att hänga upp sig på frågan om den är fri (vad det nu egentligen betyder). Den amerikanske psykiatriprofessorn George Ainslie gör så i sin idérika, nydanande och synnerligen läsvärda bok *Breakdown of Will* (2001).

En utgångspunkt för Ainslies bok är hur vi människor värderar framtiden i förhållande till nuet. Ett grundläggande mänskligt beteende är att föredra omedelbar behovstillfredsställelse framför en längre fram i tiden. Om du erbjuder mig valet mellan en hundralapp nu och en om ett år så tar jag den nu genast, och i själva verket slår jag till direkt till och med i fallet du erbjuder mig valet mellan 99 kr nu och 100 kr om ett år.[4] Det finns goda skäl till att vi beter oss på detta vis, något jag diskuterar lite mer utförligt i Kapitel 28.[5]

För att förklara Ainslies centrala tankegång behövs en matematisk formalisering av denna tidspreferens. Låt $V(t)$ vara värdet vi fäster vid utsikten

[4] Jag gör det även om vi antar att inflationen är noll.

[5] Se även Häggström (2007d) där jag diskuterar fenomenet ur utbildningssynpunkt.

att t tidsenheter fram i tiden få något som vi skulle ge värdet 1 om vi fick det omgående. Ett annat sätt att uttrycka saken är att $V(t)$ betecknar hur mycket vi skulle vara beredda att idag betala för att undvika en kostnad av storlek 1 vid tiden t. Per definition har vi $V(0) = 1$, och vår preferens för omedelbar belöning framför en i framtiden gör att $V(t) < 1$ för $t > 0$. Därtill är det naturligt att tänka sig att $V(t)$ är en avtagande funktion av t, vilket betyder att ju förr vi får vår belöning desto bättre.

Under naturliga antaganden om rationalitet kan man gå betydligt längre i att härleda hur $V(t)$ ser ut. Antag att jag sätter $V(1)$ till 0,9, dvs att jag idag, måndag, fäster 10% lägre avseende vid vad som händer imorgon jämfört med vad som händer idag. Hur bör jag då värdera vad som händer i övermorgon, dvs vad bör $V(2)$ vara? Med tanke på att jag imorgon, tisdag, kommer att vara samma person med samma slags tidspreferens som idag, vet jag att jag då kommer att värdera onsdagen 10% lägre än tisdagen, och jag bör därför sätta $V(2)/V(1) = V(1) = 0,9$, dvs $V(2) = V(1)^2 = 0,81$. Varje annat val av $V(2)$ leder till att jag gör en prioritering mellan i morgon och i övermorgon som jag vet att (och hur) jag kommer att ändra på i morgon, och då gör jag ju klokast i att ändra mig redan nu. Med detta slags resonemang kan man mer allmänt komma fram till att $V(t)$ bör vara en funktion på formen

$$V(t) = a^t \tag{1}$$

för något fixt $a < 1$. Endast om $V(t)$ har formen (1) förtjänar värderingen av framtiden att kallas *tidskonsistent*.

Det intressanta här är att psykologiska experiment och annan empirisk erfarenhet visar att vi har en benägenhet att värdera framtiden på ett sätt som inte alls passar in i formeln (1). Ett exempel: om jag erbjuder dig en middag värd 1000 kr om en vecka eller en värd 950 kr redan i kväll, så är chansen stor att du väljer att slå till på en middag redan i kväll; låt oss för resonemangets skull anta att du gör det. Antag nu att jag istället erbjuder dig en middag värd 950 kr om ett år, eller en värd 1000 kr om ett år och en vecka. Om du är tidskonsistent i din värdering av framtiden bör du i konsekvensens namn slå till på 950-kronorsmiddagen om ett år, men poängen här är att mycket få människor skulle göra ett sådant val: skillnaden mellan 52 och 53 veckor känns så oväsentlig att du hellre väntar en vecka extra för att få den aningen finare middagen. Du värderar alltså framtiden *tidsinkonsistent*.

Ainslie pekar i *Breakdown of Will* på data från en rad undersökningar som tyder på att den tidspreferens vi i praktiken tillämpar inte liknar (1)

särskilt väl, utan snarare har formen

$$V(t) = \frac{1}{1+ct} \tag{2}$$

för något fixt $c > 0$. En sådan tidspreferens kallas *hyperbolisk*.

Ett annat viktigt begrepp är den så kallade *diskonteringsräntan*, som betecknas $r(t)$ och definieras som den relativa värdeminskningen mellan tid t och tid $t+1$, dvs

$$r(t) = \frac{V(t) - V(t+1)}{V(t)}\,.$$

I fallet (1) med tidskonsistent diskontering är $r(t)$ konstant i tiden (och lika med $1-a$). I fallet (2) med hyperbolisk diskontering visar det sig att $r(t)$ blir avtagande i t.[6]

Vår tendens till hyperbolisk – eller mer allmänt tidsinkonsistent – diskontering har intressanta och långtgående psykologiska konsekvenser.[7] Den innebär att jag idag, måndag, inte är överens med mitt framtida jag i morgon, tisdag, om den relativa värderingen av vad som händer på tisdag och vad som händer på onsdag. Detta leder till att vi, analogt med det mer välbekanta psykologiska fenomenet att vi i många situationer söker manipulera personer vi inte är överens med, har anledning att söka manipulera våra framtida jag. Den hyperboliska diskonteringen kan förstås som ett slags "närsynthet" inför

[6] Hyperbolisk diskontering i begreppets strikta betydelse definieras av (2), men ofta används begreppet mer allmänt för att beteckna diskonteringsfunktioner där räntan $r(t)$ är avtagande i t.

[7] Även inom nationalekonomin har begreppet hyperbolisk diskontering väckt en del intresse. Det har visat sig att de vanliga valen av diskonteringsräntor – ett typiskt sådant är en årlig ränta om 3% – som fungerar bra vid kalkyler på några få års sikt, ger egendomliga resultat om de används i det slags betydligt mer långsiktiga samhällsekonomiska kalkyler som är aktuella i samband med exempelvis global uppvärmning. T.ex. får vi $(1-0{,}03)^{100} \approx 0{,}05$, vilket innebär att vi med den 3%-iga diskonteringsräntan blott är beredda att betala 5 öre idag för att undvika en kostnad om 1 kr om 100 år – en hårresande likgiltighet inför framtiden. Som lösning på detta problem (vilket jag för övrigt diskuterar med ingående i Kapitel 28 och 29) har ibland föreslagits hyperbolisk diskontering; se t.ex. Weitzman (2001), Karp (2005) och Bostedt m.fl. (2006). Jag är dock mycket tveksam till en sådan "lösning", ty en samhällsplanering baserad på hyperbolisk diskontering innebär i praktiken att vi kräver av framtida generationer att ta mer långsiktig hänsyn i sitt agerande än vad vi själva gör, något som vi inte rimligtvis har vare sig maktmedel eller moralisk rätt till. Sinnebilden av vår tendens till hyperbolisk tidspreferens är min önskan om att å ena sidan nu i kväll få festa loss på gräddtårta, gåslever och champagne och å andra sidan leva mer dygdigt och sparsamt kommande kvällar. Detta är ett psykologiskt fenomen vi alla har att försöka betvinga; att istället göra det till förebild för långsiktig samhällsplanering synes mig groteskt.

framtiden, och vad vi försöker göra är att förmå våra framtida jag att bete sig mindre närsynt.

Sådan självmanipulation ägnar vi oss i själva verket dagligen åt. Jag pensionssparar för att göra det svårare för mig själv att under den närmaste 25-årsperioden komma åt pengar jag kan väntas behöva på äldre dar, och jag köper ett orimligt dyrt årskort till gymmet i syfte att få mig själv att välja att träna oftare. Det allra mest patetiska exempel på sådan manipulation jag vid självrannsakan funnit är hur jag beter mig, halvt proppmätt men ändå godissugen, på lördagkvällen framför TV:n. Jag tar då ännu en godisbit, men skjuter iväg godisskålen så att jag när nästa sug uppstår någon minut senare inte skall orka sträcka mig efter ytterligare en bit och därmed riskera överskrida den hårfina gränsen mellan mättnad och illamående (en strategi som lyckas ibland men inte alltid).

Ainslie ger i sin bok många belysande exempel på den här sortens självmanipulation och dess konsekvenser. Särskilt intressant är hans demonstration av hur vi i vissa situationer handlar långsiktigt som ett led i strategin i att skapa en inre bild av oss själva som långsiktiga, något som kan påverka våra framtida jag till bättre långsiktigt tänkande. Han går sedan vidare och begagnar idén om hyperbolisk diskontering som nära nog ett universalverktyg för att förstå olika aspekter på det mänskliga psyket, inklusive smärta, humor och livets mening. Det tycks mig som om han här går längre än idén egentligen förmår bära, men detta menar jag är karaktäristiskt för banbrytande tänkare: att oblygt och utan rädsla för misslyckande söka pressa sina idéer så långt det går i försöken att kasta ljus över svåra frågor om oss själva och vår värld.

*

Människans fria vilja räknas som sagt till de stora filosofiska frågorna, och det är naturligtvis oklart i vad mån någon av de ovan refererade tankarna bidrar till att lösa problemet. Men i mitt tycke finns det andra filosofiska frågeställningar som är *ännu* besvärligare. En sådan går jag över till att diskutera i nästa kapitel, nämligen den om det mänskliga medvetandet. Hur kan det i ett materialistiskt universum finnas plats för subjektiva upplevelser? Skälet till att jag finner medvetandet mer gåtfullt än den fria viljan är följande. Frågan om den fria viljan medger det enkla och koherenta (om än möjligen moraliskt anstötliga) svaret att den helt enkelt är inbillad. Motsvarande undanmanöver fungerar inte då frågan istället handlar om medvetandet, ty vem är i stånd att inbilla sig något utan att först och främst vara medveten?