Teorien om rettferdighet

Et forsvar av utilitarismen som fordelingsprinsipp

Albert Didriksen

Våren 2021

Masteroppgave i statsvitenskap

Institutt for statsvitenskap

UNIVERSITETET I OSLO

Antall ord: 34 979

Sammendrag

Alle kan ikke få alt de ønsker seg. Godene må fordeles mennesker imellom. Et spesielt viktig spørsmål er hvordan godene fordeles på en rettferdig måte. Krever rettferdigheten at godene fordeles likt, at vi gir prioritet til de som har minst, at alle har tilstrekkelig mye, eller at alle får det de fortjener?

I denne oppgaven argumenterer jeg for at ingen av disse svarene stemmer. En fordeling av goder og byrder er rettferdig om den bidrar til mest mulig velferd. Med andre ord forsvarer jeg påstanden at utilitarismen er den beste rettferdighetsteorien. Jeg viser at utilitarismen kommer seirende ut av en sammenlikning med plausible konkurrenter og at den stemmer overens med våre reflekterte oppfatninger om hvordan goder og byrder bør fordeles i bred reflektert likevekt.

Forord

Først og fremst vil jeg takke Aksel Braanen Sterri for eksepsjonelt god veiledning. Han har vært flink til å dytte meg i riktig retning hver gang jeg har gått meg vill i løpet av det siste året. Det er bra, for det har skjedd stadig vekk. Hjelpsomheten har Aksel kombinert med å vise respekt for at oppgaven er mitt prosjekt, og fremstått oppriktig nysgjerrig på mine refleksjoner. Det har gitt meg selvtillit og gjort skriveprosessen inspirerende. Takk, Aksel!

Jeg flyttet til Norge i 2015, samme år som jeg begynte å studere. Følgelig består hele vennegjengen min av mennesker med minst 60 studiepoeng i statsvitenskap. Alle sammen fortjener egentlig å bli nevnt, men jeg er farlig nære ordgrensen allerede. Spesielt vil jeg trekke frem Johanne, Hans Peter, Marte, Emilia, Sofie, Jon og Eric for deres bidrag i forbindelse med masteroppgaven. Dere har alle vært *nødvendige* for at jeg til slutt kom meg gjennom skoleåret. Sammen håper jeg vi også har gjort en *tilstrekkelig* stor innsats til å besvare problemstillingen min.

Alle gjenstående feil i oppgaven er mine egne.

Albert Didriksen,

Oslo, 2021.06.21

Innholdsfortegnelse

1 Innledning og problemstilling

Godene vi har, som penger og utdanning, styrer i stor grad livene våre. De er avgjørende for hvor mye og hvor godt vi får spise, hvor rask helsehjelp vi får, hvilke sosiale lag vi beveger oss i, om vi kan kjøpe oss hytte på fjellet eller trues av fraflytting ved uforutsett inntektsfall og mye mer.

Kunne alle fått det de ville til enhver tid, ville mye konflikt vært unngått. Slik er det ikke. Knapphet er et ufravikelig faktum. Goder må fordeles mennesker imellom. Siden de fleste ønsker seg mer enn de kan få, må ulike hensyn vektes mot hverandre. Hvordan godene fordeles er av stor betydning. Vi forventer at fordelingen er rettferdig. Hva dette innebærer, er imidlertid et åpent spørsmål. Mens noen mener rettferdighet betyr at alle har det like godt, forventer andre å bli behandlet i tråd med innsatsen de legger ned for samfunnets beste. For at uoverensstemmelser skal kunne løses, kan ikke begrepet være tvetydig. Man trenger en teori om rettferdighet. Det er dette jeg setter meg fore å undersøke her. Min problemstilling er følgende:

Under hvilke betingelser er en fordeling av goder og byrder rettferdig?

I denne oppgaven argumenterer jeg for at en fordeling er rettferdig hvis den gir mest mulig velferd til mennesker. Jeg argumenterer med andre ord for at utilitarismen er den beste teorien om fordelingsrettferdighet.

Rettferdighet er et sammensatt begrep med flere, til tider overlappende, underkategorier, som konservativ rettferdighet, fordelingsrettferdighet og gjengjeldelse. I denne oppgaven tar jeg kun for meg fordelingsrettferdighet. I kapittel 2 viser jeg derfor hvordan dette skiller seg fra andre former for rettferdighet. I tillegg presenterer jeg utilitarismen, som er fordelingsprinsippet jeg forsvarer i denne oppgaven. I kapittel 3 beskriver jeg metoden jeg undersøker problemstillingen med, reflektert likevekt. Denne fremgangsmåten innebærer å sammenligne mitt forslag med alternative prinsipper og se hvilke prinsipper som stemmer best overens med våre reflekterte oppfatninger. Jeg vurderer derfor Rawls' rettferdighetsteori, egalitarianisme, prioritarianisme og tilstrekkelighetsprinsippet i kapittel 4. I tillegg tar jeg for meg noen vanlige ankepunkter ved utilitarismen, som at den legitimerer grusomheter som slaveri. Jeg argumenterer for at utilitarismen ikke svekkes nevneverdig av innvendingene, og fremstår som et mer rettferdig prinsipp enn alternativene jeg vurderer. I kapittel 5 ser jeg på om fortjeneste bør være en faktor i

en rettferdig fordeling av goder. Jeg argumenterer for at fortjeneste ikke bør spille noen rolle, siden konseptet hviler på lite sannsynlige premisser angående moralsk ansvar. Kapittel 6 gir en oppsummering av funnene mine og peker ut veien for videre forskning.

2 Fordelingsrettferdighet og utilitarismen

2.1 Fordelingsrettferdighet

2.1.1 Ideell fordelingsrettferdighet

Rettferdighet er et sammensatt begrep som fanger flere, til tider overlappende, konsepter.

Man bruker gjerne ordet i forbindelse med juridiske prosesser. Det er for eksempel urettferdig å bli straffet hardere enn andre som begår samme forbrytelse som en selv. Mer generelt kan man snakke om det David Miller kaller konservativ rettferdighet, altså et utrykk for i hvilken grad allerede eksisterende normer og praksiser følges opp (Miller, 2017, Kapittel 2.1). Disse spørsmålene setter jeg til side i min oppgave. Som det kommer frem av innledningen, er temaet jeg tar for meg fordelingsrettferdighet. Innenfor dette undersøker jeg hvilket prinsipp som *ideelt sett* bør styre fordelingen av goder, i motsetning til den konservative tilnærmingen, som ville vært å se på hvor godt et allerede akseptert prinsipp etterleves.

2.1.2 Ideelle prosedyrer og utfallsorientert fordelingsrettferdighet

Et annet skille som gjøres, er mellom prosedurale og utfallsorienterte rettferdighetsteorier (Miller, 2017, Kapittel 2.3; se også Rawls, 1999, Kapittel 14).[1] Prosedurale teorier definerer et utfall som rettferdig om det følger av en rettferdig prosess. John Rawls eksemplifiserer dette med gambling (Rawls, 1999, s. 75). Gitt at det ikke er noe juks i bildet, er utfallet av for eksempel et veddemål av ti runder med «kron eller mynt» rettferdig, uansett hva resultatet blir. Utfallsorientert rettferdighet ser derimot kun på resultater. Slike teorier vurderer prosesser som rettferdige om de fører til rettferdige utfall. Også prosedurale teorier legger jeg til side, og fokuserer på utfallsorienterte tilnærminger. Det er to grunner til dette.

[1] Rawls opererer ikke med to, men tre kategorier: Fullkommen-, ufullkommen- og ren prosedural rettferdighet. I denne omgang er derimot en todeling tilstrekkelig.

For det første, så er utfallsorienterte teorier mer åpenbart knyttet til *fordelings*rettferdighet. En gitt fordeling av goder er nettopp det, et utfall. Prosedurale teorier er gjerne mer knyttet til juridiske aspekter av rettferdighetsbegrepet – hvorvidt ulike kontrakter og normer blir fulgt opp, kan være viktig for å avgjøre om en prosess er rettferdig. I deler av litteraturen beskrives følgelig fordelingsrettferdighet og prosedural rettferdighet som kontrasterende kategorier.[2]

For det andre, har de utfallsorienterte prinsippene jeg vurderer større anvendelighet enn de prosedurale teoriene jeg utelater. Selv om også prosedurale teorier peker på en fordeling som den riktige (den som følger av de riktige prosedyrene), gjøres dette gjerne med det Amartya Sen kaller den transcendentale tilnærmingen (Sen, 2006, s. 216–218). Det vil si at teorien kun beskriver hva perfekt rettferdighet innebærer. Slike utopiske scenarioer er imidlertid kun oppnåelige med visse lite realistiske forbehold, slik som at teorien etterleves til det fulle av alle samfunnsmedlemmer. Nozicks libertarianisme beskriver for eksempel en rettferdig fordeling som en der alt eierskap følger av rettferdig ervervelse eller rettferdige byttehandler (Nozick, 1974, s. 151). Fordelingen av goder i dagens verden er imidlertid ikke kun resultat av dette, men en lang historie med krig, tyveri og annen illegitim ervervelse av midler, uten at det er mulig å gjøre seg en fullstendig oversikt over hvordan godene ville vært fordelt om slike forbrytelser aldri fant sted. Nozicks teori fungerer derfor dårlig som en rettesnor i søken etter en mer rettferdig fordeling i dag.[3]

Utfallsorienterte prinsipper kan derimot brukes i det Sen kaller den komparative tilnærmingen (Sen, 2006, s. 216). Den komparative tilnærmingen er egnet til å brukes i den virkelige verden, siden den ikke omhandler hva *perfekt* rettferdighet innebærer, men hvordan *mer* rettferdighet kan oppnås. De utfallsorienterte prinsippene fungerer som kriterier til å rangere ulike utfall fra det beste til det verste. Derfor er de anvendbare, også i en ikke-ideell verden, der moralteorier ikke følges av alle til punkt og prikke. Et egalitært prinsipp vil for eksempel kunne peke på utfallet med mest likhet av tre mulige fordelinger som det mest rettferdige. Slik vil prinsippet kunne fordre politikk som fører til denne fordelingen, til tross for at det, i ideelle

[2] Skillet gjøres riktignok oftest i forskning knyttet til psykologi (se f.eks. Folger, 1987; McFarlin & Sweeney, 1992). Basert på argumentene over mener jeg det også er riktig å gjøre denne distinksjonen i min oppgave.
[3] For en diskusjon av vanskelighetene med den praktiske implementeringen av Nozicks teori, se f.eks. side 18-68 i *Ethics for a Broken World* (Mulgan, 2011, s. 18–68).

omstendigheter, ser på fullstendig likhet som det optimale. På samme måte vil utilitarismen peke på utfallet med mest velferd av de tilgjengelige alternativene, selv om det finnes mer optimale løsninger i en ideell verden. Om beslutningstakere i et velstående land vurderer om de bør gi 0.7% eller 0.8% av BNI til veldedighet, og tenker å velge det mest rettferdige av alternativene, finner de svaret med å kombinere det etiske prinsippet de følger med relevante empiriske fakta om de to alternativene. I min oppgave argumenterer jeg for at riktig etisk prinsipp å følge er utilitarismen, og de derfor må velge alternativet som fører til mest velferd. Det er nærliggende å tro at dette innebærer å gi 0.8%. Til tross for at utilitarismen trolig ville fordret å allokere betraktelig mer til bistand *ideelt sett*, vil kriteriet også kunne vurdere 0.8% som *mer rettferdig* enn 0.7%, og som det *mest rettferdige* av de tilgjengelige alternativene.

Som nevnt er rettferdighetsbegrepet sammensatt, og dets underkategorier er til dels overlappende. I min diskusjon av fordelingsrettferdighet vil jeg ta for meg John Rawls' teori, til tross for at han selv beskriver den som prosedural (Rawls, 1999, s. 76). Det er primært to grunner til at jeg likevel inkluderer denne.

Det første er at Rawls' teori med noen endringer kan gjøres om til å være utfallsorientert. Det prinsippet hans som er mest relevant for fordelingsrettferdighet, døpt forskjellsprinsippet, gir uttrykk for maximin-kriteriet. Maximin-kriteriet rangerer ulike utfall basert på hvor mange goder den dårligst stilte har (i Rawls' teori er dette ressurser, men man kan også anvende maximin-kriteriet på velferd). Flere av Rawls argumenter tyder på at han også anser dette som et mer rettferdig *fordelings*prinsipp enn utilitarismen, til tross for at han selv primært fremmer det i forbindelse med valg av innretningen av samfunnets mest grunnleggende institusjoner, og ikke allokeringssituasjoner direkte. I en slik, modifisert form, kan også Rawls' teori anvendes som et komparativt rettferdighetsprinsipp. For det andre utviklet Rawls sin teori i eksplisitt konkurranse med utilitarismen. Å utelate en såpass sentral bit av rettferdighets-litteraturen som det Rawls' bidrag er fra en diskusjon om fordelingsrettferdighet, som attpåtil inneholder en rekke viktige innvendinger mot utilitarismen, ville vært å glatte over svakheter ved sistnevnte teori.

2.1.3 Fordeling og straff

Et begrep jeg imidlertid undersøker som en del av min diskusjon av fordelingsrettferdighet, er fortjeneste. Deler av litteraturen analyserer også dette som en separat kategori fra fordelingsrettferdighet.[4] Jeg mener derimot aspektene bør sees sammen. Som også John Stuart Mill påpeker, er fortjeneste blant de viktigste konseptene som forbindes med rettferdighet i dagligtalen (Mill, 1863, s. 218). Man er generelt opptatt av å ha goder i tråd med det man fortjener, noe som gjerne poengteres når ulike arbeidsgrupper går ut i streik. Å begrense riktig prinsipp for fordelingsrettferdighet til å kun gjelde i tilfeller der fortjeneste ikke spiller inn, samtidig som man ikke gjør vurderinger angående når fortjeneste er relevant, ville vært å snevre inn prinsippet altfor mye.

En annen underkategori av rettferdighetsbegrepet angår hvorvidt man bør straffe folk for ugjerninger, og i så fall, hvor alvorlig. Denne formen for juridisk rettferdighet er ikke et fokus i denne oppgaven. Når det er sagt, er dette ikke helt avskilt fra fordelingsrettferdighet. Å straffe folk for sine handlinger innebærer å frata dem goder, eller allokere dem onder. I så måte fremstår slik *gjengjeldelse* som en spesifikk form for fortjeneste. I min diskusjon av fortjeneste i kapittel 5 undersøker (og avviser) jeg konseptet moralsk ansvar. I den forbindelse skriver jeg om gjengjeldelse i samme åndedrag som fortjeneste. Siden spørsmålet om juridisk straff er utenfor min oppgaves rekkevidde, undersøker jeg imidlertid ikke om det finnes andre grunner til en slik form for straff enn den som knytter seg til diskusjonen av moralsk ansvar, og om det finnes relevante skiller mellom tematikken fortjeneste og gjengjeldelse. Mine funn kan forhåpentligvis være relevante også på dette feltet, men gir hverken utfyllende eller endelige svar på spørsmål angående juridisk rettferdighet.

2.1.4 Ulike premiss

Rettferdighet er svært viktig, siden fordelingen av goder har store implikasjoner for menneskers liv. Jeg gir imidlertid ingen konkrete råd for politikkutforming, eller til andre dilemmaer der hensynet til rettferdighet bør ivaretas. For å finne den rettferdige fordelingen i en gitt situasjon, trenger man to premisser: Det empiriske og det etiske. Man trenger altså vite alle relevante (empiriske) fakta om

[4] I noen deler av litteraturen, slik som hos ulike egalitære tenkere, poengteres det at ufortjente ulikheter er urettferdige. Noen ganger poengteres det også at fortjente ulikheter kan være rettferdige (se f.eks. Temkin, 2003, s. 767). I andre deler av litteraturen blir fortjeneste satt til siden, eller avvist uten en grundig diskusjon (se f.eks. Parfit, 1991, s. 33).

valgalternativene man har, samt kjenne til hvilke (etiske) kriterier man baserer rangeringen av disse på. Med å se på disse to sammen, vet man hvilket valg som er det rettferdige.

I denne oppgaven vurderer jeg kun det andre, etiske, premisset. Som nevnt tar jeg til ordet for at det riktige fordelingsprinsippet er utilitarismen. Gitt at mine argumenter overbeviser, bør videre forskning undersøke hvilken fordeling av goder og onder som fører til mest mulig aggregert velferd. Politikere på sin side bør gjøre sitt for at slike ressursfordelinger skal være realistiske å enes om.[5]

2.2 Utilitarisme

Utilitarianismen kommer i flere former. Alle slike teorier har imidlertid tre elementer til felles.[6] For det første er de konsekvensialistiske. Utilitarismen vurderer handlinger som riktige eller gale, institusjoner som rettferdige eller urettferdige, og mennesker som dydige eller mindre dydige, *utelukkende* basert på hvilke konsekvenser de frembringer. Den vanligste versjonen av teorien er handlingsutilitarismen. For en handlingsutilitarist er dette forholdet i utgangspunktet direkte. Handlinger er riktige om dets konsekvenser er så gode som mulig, og ellers er de gale.[7] For det andre er utilitarismen velferdsbasert. Det vil si at teorien ser på velferd, altså hvor godt mennesker eller andre følende vesener har det, som det eneste iboende godet. Andre ting kan også være verdifulle, men kun instrumentelt, altså verdifulle i kraft av dets positive virkninger på folks velferd. Sykdom er for eksempel et onde siden det

[5] For en oversikt over ulike deler av debatten om ideell og ikke-ideell teori innenfor politisk filosofi, se (Valentini, 2012). I diskusjonen til nå har jeg berørt alle de tre til dels overlappende skillene Valentini trekker opp: Henholdsvis mellom full og delvis etterlevelse, utopier og realistiske teorier samt skillet mellom endestasjons- og overgangsrelevante prinsipper, som er det Sen kaller transcendentale og komparative rettferdighetssyn.

[6] Min oppsummering av utilitarismen følger i stor grad innledningskapitlet i *Taking Utilitarianism Seriously*, som selv er basert på Sens *Utilitarianism and Welfarism* (Woodard, 2019, Kapittel 1.1; Sen, 1979, s. 464–471). En kortfattet, men nyttig innføring i utilitarismen finnes f.eks. i *An introduction to moral philosophy*, mens en beskrivelse av utilitarismen som rettferdighetsteori er å finne i *Justice: What's the right thing to do?* (Wolff, 2006, Kapittel 8; Sandel, 2009a, Kapittel 2).

[7] Selv om handlingsutilitarisme er den vanligste versjonen, er det ikke den eneste. Regelutilitarister gjør f.eks. en mer indirekte vurdering. Ifølge denne teorien er en handling riktig hvis den følger av en regel som har de beste mulige konsekvensene. Som det fremkommer i løpet av de neste sidene, er disse skillene ikke spesielt viktige for min oppgave, da jeg primært er opptatt av de andre aspektene ved utilitarismen. I den grad det konsekvensialistiske spiller inn, er handlingsutilitarismen mitt utgangspunkt. Jeg setter derfor de andre alternativene til side i denne omgang.
Et annet skille er mellom utilitaristiske teorier som fordrer maksimering av det totale velferdsnivået, og teorier som fokuserer på det gjennomsnittlige nivået. Gitt at verdens befolkning holdes uendret, er disse to tilnærmingene like. I denne oppgaven tar jeg ikke stilling til dette spørsmålet. For enkelhetens skyld skriver jeg ofte om aggregert velferd, men dette er ikke et uttrykk for at jeg foretrekker det totale.

gjerne medfører mindre velferd, mens ressurser, som f.eks. penger, regnes som goder, siden de ofte bidrar til velferdsøkning. Utilitarismens siste aspekt er at den er additiv. Ulike utfall vurderes ved at mengden goder (altså velferd) summeres opp. Om teorien åpner for negativ velferd, som i *lidelse*, trekkes dette fra summen positiv velferd. Jo mer velferd et scenario inneholder, desto bedre. Ønsket utfall er altså så mye velferd som mulig, fremfor for eksempel en lik fordeling mellom aktørene.

Til sammen er altså utilitarismen en gruppe konsekvensialistiske teorier, som vurderer utfall basert på mengden summert velferd. Som det til dels fremkommer av beskrivelsen over, er utilitaristiske teorier både evaluative og deontiske.[8] Med evaluativ mener jeg at de rangerer utfall fra den beste til den dårligste (basert på velferden i dem). At de er deontiske, vil si at de spesifiserer hva som er moralsk riktig handling, nemlig den som fører til mest mulig velferd.

Vurderer man utilitarismens tre bærebjelker på et prinsipielt nivå, uavhengig av implikasjonene, fremstår teorien rimelig. Konsekvensialisme i seg selv har for eksempel en betydelig intuitiv appell. Som selv en markant kritiker av tilnærmingen, Samuel Scheffler, erkjenner, virker det å stemme at man bør frembringe goder og unngå onder, slik konsekvensialismen fordrer (Scheffler, 1988, s. 1). Det samme kan sies om velferdismen. Det virker riktig at menneskets (eller andre sansende vesens) velvære har iboende verdi – det fremstår bedre om noen har det bra, enn om de har det dårlig. Og, som Christopher Woodard påpeker, så er det i det minste plausibelt ved første øyekast at velferd er det eneste som er av iboende verdi. Man anser gjerne også ting som autonomi, skjønnhet, kunnskap og lignende som verdifulle. Disse bidrar imidlertid gjerne til velferd, så det er i det minste ikke åpenbart at deres verdi ikke kun er instrumentell (Woodard, 2019, s. 9). Til slutt fremstår det også riktig at om man har definert noe som verdifullt, så er det ønskelig med så mye som mulig av det. Også det summative aspektet ved utilitarismen virker derfor svært plausibelt. Samler man de tre punktene, får man altså et umiddelbart tiltrekkende prinsipp, som sier at det moralsk riktige er å maksimere velferd.

Likevel har utilitarismen møtt sterke innvendinger. Noen av prinsippets implikasjoner fremstår uakseptable. Blant de vanligste kritikkene utilitarismen møter, er at den er urettferdig. Siden den kun rangerer utfall basert på aggregert velferd, utelukker for

[8] For definisjoner på henholdsvis evaluativ og deontisk teori, se f.eks. (Tappolet, 2013, s. 1).

eksempel ikke utilitarismen de privilegertes utnyttelse av et mindretall, hvis dette bidrar til økt velferd. I prinsippet kan en utilitaristisk politikk føre til at rike slaveeiere utnytter sine slaver, gitt at de selv nyter godt av dette. Teorier som garanterer et minimumsnivå av velferd, vektlegger likhet eller prioriterer å hjelpe de dårligst stilte, svarer tilsynelatende bedre til noen av de vanligste oppfatninger vi har om hva en rettferdig fordeling innebærer. Følgelig har utilitarismen blitt diskutert relativt lite i forbindelse med fordelingsrettferdighet. Ofte blir det ikke engang nevnt som et seriøst alternativ, mens i andre tilfeller blir det gjerne avvist som en lite plausibel teori (se f.eks. Anderson, 1999; Cohen, 1989; Eyal et al., 2013, s. V–VI; Kymlicka, 2002, Kapittel 2.5; Powers & Faden, 2006, s. 50).

I denne oppgaven tar jeg imidlertid utilitarismen i forsvar. Jeg vurderer noen av de vanligste ankepunktene, og sammenligner teorien med konkurrerende rettferdighetsprinsipper. Min påstand er at ingen av innvendingene svekker utilitarismen nevneverdig, og at det er det beste av fordelingsprinsippene jeg beskriver i oppgaven.

2.2.1 Teoriens omfang

Et aspekt ved teorier angående fordelingsrettferdighet er deres omfang, altså fordelinger mellom hvilke individer eller grupper som bør vurderes som rettferdige eller urettferdige. Som nevnt trenger utilitarismen ikke engang begrenses til mennesker. Snarere tvert imot, siden man kan snakke om velferd i forbindelse med dyr, støtter de fleste utilitarister å ta hensyn til disse (se f.eks. Bentham, 1789, s. 325, fotnote a; Singer, 1995). Selv om jeg selv deler synet om at dyrevelferd er et svært viktig og alt for sporadisk berørt aspekt ved både politikk og moral, setter jeg den diskusjonen til siden i denne omgang.

Spørsmålet om hvem som skal regnes med i en rettferdig fordeling, melder seg også når man kun vurderer mennesker. Man skiller for eksempel gjerne mellom global og nasjonal fordelingsrettferdighet, altså fordelinger mellom alle verdens mennesker (eller mellom land), samt fordelinger innad i et land. Svaret på dette knytter seg til poenget diskutert i del 2.1.4. Utilitarismen, så vel som de fleste andre prinsippene jeg vurderer, kan anvendes globalt. I en ideell verden, der praktiske hensyn ikke satte grenser, ville utilitarismen fordret at alle verdens goder ble fordelt på den måten

som gav mest aggregert velferd globalt sett. Tilsvarende fordrer i hvert fall noen versjoner av egalitarismen at alle verdens mennesker bør være på samme velferdsnivå, ideelt sett. Vi lever derimot ikke i en ideell verden.

Når utilitaristisk fordelingsrettferdighet anvendes i praksis, må som nevnt den etiske dimensjonen kombineres med et empirisk premiss. Deler av dette premisset er nettopp hvem den aktuelle fordelingen gjelder. Når politikere lurer på hva som gir den mest rettferdige bistandspolitikken, er det naturlig at de tenker globalt. Stilles derimot spørsmålet om landet bør ha en flat eller progressiv skattesats, og om inntektene bør gå til utdanningsstøtte blant landets fattigste eller til en allmenn borgerlønn, er disse nasjonale fordelingsspørsmål. Utilitarismen vil fortsatt kunne besvare det med å peke på alternativet som fører til mest aggregert velferd. Tilsvarende kan man for eksempel også vurdere hva som er den mest rettferdige fordelingen av en kasse brus innad i en vennegjeng, altså en fordeling blant enda færre mennesker.[9]

Siden fordeling og omfordeling er ansett som et av statens viktigste oppgaver, og fordelingsrettferdighet, både i dagligtalen og i faglitteraturen, gjerne omtales på nasjonalt nivå, anvender jeg som regel samme språk i oppgaven.[10] Dette er imidlertid kun av praktiske årsaker. Som beskrevet er utilitarismen globalt gjeldende, mens empiriske premisser begrenser dets anvendbarhet.

2.2.2 Utilitaristisk fordelingsrettferdighet

For å illustrere hvordan fordelingsrettferdighet skiller seg fra andre underkategorier av rettferdighetsbegrepet, samt å vise hva utilitarismen fordrer i den sammenheng, låner jeg et eksempel av Christopher Woodard:

[9] At globale prinsipper skaleres ned til å gjelde lokalt, er riktignok ikke like ukontroversielt som i eksempelet hvis det er noen eksterne virkninger av fordelingen på lokalt nivå til det globale nivået. Følgelig bør så mange som mulig av de som påvirkes av en fordeling tas med i vurderingen, før man velger fremgangsmåte.

[10] Merk at en grunn til at fordelingsrettferdighet primært kan diskuteres innad i et land, er at det ellers ville kollidert med andre rettferdighetshensyn. Teoriene jeg vurderer ville trolig alle fordret en svært kraftig omfordeling fra verdens rikeste land til mindre velstående områder i verden. Dette ville derimot vært urettferdig mot de rike landenes beboere, i rettferdighetsbegrepets konservative forstand. Forventninger velstående mennesker har opparbeidet seg basert på eksisterende normer og praksiser, ville ikke blitt fulgt opp, om de plutselig måtte skatte mesteparten av sin inntekt for at ressursene skulle sendes til utlandet.

> Ellen er på vei for å donere alle sine oppsparte midler til en svært effektiv veldedig organisasjon. På veien møter hun Joe, som stjeler Ellens penger, og donerer de til en marginalt mer effektiv veldedig organisasjon (Woodard, 2019, s. 142).

I denne situasjonen vurderer mitt utilitaristiske prinsipp utfallet som en mer rettferdig fordeling enn alternativet, der Joe ikke stjal Ellens penger. Siden pengene er omfordelt til en mer effektiv organisasjon enn det de ellers ville vært, bidrar denne ressursfordelingen til mest mulig velferd (av alternativene), som er det utilitarismen som fordelingsprinsipp fordrer.

Woodard er uenig i at en utilitaristisk teori om fordelingsrettferdighet innebærer dette.[11] Han mener det kan være en implikasjon av prinsippet at handlingen til Joe *riktig*, siden det bidrar til maksimering av velferd. Imidlertid mener han at et utilitaristisk fordelingsprinsipp trenger å gjenspeile intuisjonen om at Joe gjorde noe urettferdig (Woodard, 2019, s. 142).

Woodard blander her fordelingsrettferdighet med andre deler av rettferdighetsbegrepet. Det stemmer trolig at Joe gjorde noe urettferdig.[12] Hans handlinger er ulovlige, og følgelig urettferdige, i juridisk forstand. Handlingene er i tillegg urettferdige i konservativ forstand. Gitt eksisterende normer og praksiser, burde Ellen kunne forvente å få velge mottakerorganisasjon for pengene sine selv. Det kan også hende at Joes handlinger er gale fra et utilitaristisk standpunkt. Siden det å bli frastjålet penger er svært opprørende, og siden hans handlinger kan bidra til normaliseringen av tyveri, og dermed kan ha negative ringvirkninger, er det ikke gitt at å stjele og donere videre pengene bidrar til mer velferd enn om han lot Ellen styre pengene sine selv.

Både urettferdigheten og de nevnte ondene knytter seg imidlertid til handlingen (tyveriet), ikke til fordelingen av ressurser i seg selv. I eksemplet er effekten av fordelingen maksimering av velferd. En utilitaristisk teori om fordelingsrettferdighet

[11] I forbindelse med eksemplet skriver han riktignok om rettferdighet generelt. Tidligere i boken fastslår han imidlertid at det er fordelingsrettferdighet han sikter til (Woodard, 2019, s. 138).
[12] Jeg skriver «stemmer trolig», siden jeg ikke undersøker andre aspekter av rettferdighetsbegrepet enn fordelingsrettferdighet, og følgelig ikke har grunnlag til å være for bombastisk. Selv om det virker svært plausibelt at Joe opptrer på en måte som medfører juridisk urettferdighet, tenker man f.eks. intuitivt ikke det samme om Robin Hood, som stjeler fra de rike for å gi til de fattige.

fordrer nettopp det. Slik kan konseptet holdes avskilt fra andre deler av rettferdighetsbegrepet, eller andre aspekter av moralsk vurdering mer generelt.

På dette punktet er det verdt å adressere en innvending. En mulig tolkning av rettferdighetsbegrepet er at det nødvendigvis utgjør et separat hensyn fra hva som er best i nyttemaksimerende forstand. I følge en slik tolkning finner man den moralsk riktige fordelingen ved å undersøke hvilken av alternativene som best mulig balanserer det utilitaristiske hensynet med rettferdighet – for eksempel hvilken fordeling som kombinerer likhet og den totale mengden velferd på best mulig måte. Maksimering av velferd vurderes tross alt som det *riktige* av utilitarismen som moralteori. Derfor kan det fremstå overflødig å betegne det som *rettferdig* i tillegg.

Ifølge et slikt rettferdighetsbegrep er en utilitaristisk fordelingsteori ikke en teori om rettferdighet. Snarere innebærer en utilitaristisk tilnærming til fordeling av goder å avvise at rettferdighet i det hele tatt er et relevant hensyn i fordelingssammenheng, og derfor kun lete etter *beste* fordeling, fremfor det *mest rettferdige*. Dette er i tråd med synet til den klassiske utilitaristen Jeremy Bentham, som kalte rettferdighet et innbilt konsept, som mennesker kun har funnet opp for å enklere kunne snakke om nyttemaksimering (Bentham, 1789, s. 139, i bunnteksten).

Det er mulig det ikke går et konseptuelt skille mellom hva som er bra i fordelingssammenheng, og hva som er rettferdig. Jeg mener likevel at å beholde begrepet er riktig, i det minste for praktiske formål. Begrepet rettferdighet er svært innarbeidet både i dagligtalen og i faglitteraturen. Utilitarismen gir svar på spørsmål som vanligvis besvares i rettferdighetstermer, nemlig «hvordan bør goder fordeles?». Å holde fast ved at utilitarismen ikke har et alternativ for fordelingsrettferdighet, men snarere avviser hele konseptet, vil i praksis kunne føre til at tilnærmingen uteblir fra viktige debatter angående dette feltet. Det er derfor, av praktiske grunner, best å holde seg til de samme uttrykkene som resten av litteraturen innenfor fordelingsrettferdighet, og følgelig bruke rettferdighetsbegrepet om egen tilnærming.[13] Videre i oppgaven setter jeg derfor denne innvendingen til

[13] En annen grunn til at å holde på rettferdighetsbegrepet er metoden jeg bruker, beskrevet i del 3.1. Jeg starter oppgaven med noen intuisjoner om rettferdighet. Teorien som best forklarer disse, bør følgelig kunne kalles et rettferdighetsprinsipp.

side, og argumenterer for utilitarismen som det mest rettferdige fordelingsprinsippet. Med det avklart, går jeg videre til å presentere oppgavens metode.

3 Fremgangsmåte

Det finnes flere mulige måter å undersøke moralske spørsmål på. Tilhengere av utilitarismen bruker gjerne en aksiomatisk metode i sin argumentasjon (Burns, 2005, s. 46; Singer, 1974, s. 517).[14] Det innebærer å finne noen, ifølge dem, selvforklarende grunnprinsipper, og utlede utilitarismen fra disse via deduktive slutninger. Fremgangsmåten tilsvarer den René Descartes tok i bruk i sin tenkning (Kirkeby, 2009, s. 19). Descartes starter fra det han anser som en selvsagt grunnsetning: «Jeg tenker, altså er jeg» (Descartes, 2003, s. 28). Videre resonnementer om verdens eksistens utleder han fra dette. På lignende vis legger for eksempel Henry Sidgwick aksiomer om upartiskhet og omtenksomhet til grunn, og utleder utilitarismen fra disse (Sidgwick, 1981a, Kapittel XIII). Det er to problemer med denne fremgangsmåten. For det første, så er det lettere sagt enn gjort å finne selvinnlysende sannheter. Siden disse er ment å være berettigede av kun seg selv og ikke av annen informasjon, har mennesket ikke mulighet til å kvalitetssjekke egne aksiomer. De begrunnes av dem selv, altså er oppfatningen og dets begrunnelse den samme. Så om man har godtatt et feilaktig aksiom, har man også godtatt den feilaktige begrunnelsen. Bygger man et etisk system på noen antatte sannheter som egentlig er misvisende, vil teorien mest sannsynlig være feil. For det andre, tar metoden ikke bruk av moralske intuisjoner, som generelt anses å være det viktigste datagrunnlaget for teoretisering innenfor moralfilosofien. Som jeg utdyper i del 3.1.4, fremstår dette som å la verdifull informasjon ligge ubrukt.

I denne oppgaven velger jeg derfor en annen tilnærming til normativ analyse. For å kunne ha en berettiget oppfatning om hva som er det mest rettferdige fordelingsprinsippet, forsøker jeg å oppnå såkalt reflektert likevekt. Folke Tersman påstår metoden ikke kan føre til aksepten av utilitarisme (Tersman, 1991, s. 398). Som det kommer frem av oppgaven min, mener jeg dette er feil.

3.1 Reflektert likevekt

3.1.1 Smal reflektert likevekt

[14] For eksempler på dette, se Jeremy Benthams *A Fragment on Government* og Henry Sidgwicks *The Methods of Ethics* (Bentham, 1977, s. 393; Sidgwick, 1981a, s. 379).

Reflektert likevekt som metode i moralfilosofien ble, i hvert fall med dette navnet, først tatt i bruk av John Rawls i *A Theory of Justice* (Rawls, 1971). I sin opprinnelige form, senere døpt smal reflektert likevekt, er det en metode for å systematisere moralske intuisjoner angående en gitt problemstilling. Prosessen starter med å ta utgangspunkt i intuisjonene, og se hvilket prinsipp som best mulig kan forklare disse. Siden det sjelden er mulig å finne et prinsipp som svarer til alle intuisjonene, må noen av disse forkastes. Dette er gjerne de oppfatningene man var minst sikker på fra før, og eventuelt også andre intuisjoner, om de står i veien for koherens mellom prinsippet og de resterende intuitive oppfatningene. Til slutt står man igjen med en moralteori og et sett med veloverveide vurderinger som samsvarer med hverandre (Daniels, 1980, s. 22). Som jeg beskriver i sidene som følger, antas en slik koherens å gi oppfatningene en viss grad av berettigelse.

3.1.2 Intuisjoners feilbarlighet

Utfordringen med smal reflektert likevekt er at moralske intuisjoner er feilbarlige. Som Peter Singer påpeker, kan de være pregede av for lengst «forkastede religiøse systemer og forvridde syn på sex» (Singer, 1974, s. 516). Man skal ikke langt tilbake for å finne en tid der oppfatninger om at homofili er moralsk forkastelig var svært vanlige, og i store deler av verden gjelder dette fortsatt. Det samme kan sies om fremmedfiendtlige holdninger. Det fremstår likevel ukontroversielt å påstå at hverken homofobiske eller rasistiske intuisjoner bør følges om man vil gjøre moralsk riktige refleksjoner. Når man vet at moralske intuisjoner hos andre har feilet, både tidligere og i dag, kan man heller ikke stole blindt på sine egne intuitive oppfatninger.

En naturlig innvending mot bruken av smal reflektert likevekt som metode er derfor at en systematisering, og eventuell finpolering, av våre opprinnelige oppfatninger er av liten verdi, gitt at disse oppfatningene var feil i utgangspunktet. Moralske teorier er ment å gi en rettesnor, siden intuisjoner kan være motstridende og ikke alltid gir de nødvendige svarene. Men når teorier kun baseres på intuisjoner som er kjent for å være feilbarlige, er det liten grunn til å stole på teoriene til å gi grunnlag for velbegrunnede oppfatninger (Hare, 1982, s. 25; Singer, 1974, s. 515).

3.1.3 Bred reflektert likevekt

Løsningen på dette er å etterstrebe bred fremfor smal reflektert likevekt. Også i den brede versjonen er målet å oppnå koherens mellom intuitive oppfatninger og teori, men her inngår i tillegg bakgrunnsteorier i ligningen (Daniels, 1979, s. 258). Disse kan blant annet være teorier angående epistemologi, metafysikk eller nevrovitenskap (Lycan, 2019, s. 109). Her er det altså ikke to, men tre nivåer av oppfatninger som må være konsistente med hverandre. Prosessen begynner også her med å ta utgangspunkt i intuisjonene. Men i motsetning til den smale versjonen av reflektert likevekt, ender man ikke nødvendigvis opp med teorien som forklarer flest av disse. Man vurderer heller en rekke alternative forklaringer på intuisjonene man har, og velger det prinsippet som har størst forklaringskraft, sett i lys av bakgrunnsteoriene (Daniels, 1980, s. 25). Et prinsipp som samsvarer med færre av intuisjonene enn alternativene kan fortsatt bli valgt som det beste, gitt at bakgrunnsteoriene gir en god forklaring på denne inkonsistensen mellom intuisjoner og prinsipp.

Bakgrunnsteorienes rolle kan illustreres med et eksempel. Om en person i dag fortsatt har homofobiske intuisjoner, og eventuelt andre, relaterte fordommer, kan en øvelse i smal reflektert likevekt tilsi at homofili er umoralsk. Det finnes imidlertid forklaringer på de homofobiske intuisjonene som ikke innebærer at homofili er galt. En lang tradisjon med, primært religiøs, lærdom har fortalt mennesker at kjærlighet kun skal finne sted mellom mann og kvinne. Sex som ikke fører til reproduksjon, har tidvis vært ansett som «unaturlig» eller «skittent». Mennesket har i tillegg lett for å være avvisende mot både meninger, personer og grupper som er forskjellige fra en selv (se f.eks.Greene, 2013, s. 26, 2017, s. 73). Denne oss mot dem-holdningen er trolig en konsekvens menneskets historiske røtter fra stammesamfunn, der det var nødvendig for overlevelsen å utkonkurrere andre grupperinger. Tankegangen gjør minoriteter spesielt utsatte. Det er derfor ikke rart om slike intuisjoner er til stede, selv om de ikke gjenspeiler noen sannhet om hva som er moralsk rett eller galt. Denne kunnskapen gir personen grunn til å forkaste intuisjonen om at homofili er galt om den ikke kan støttes av andre, mer troverdige, intuisjoner.

Personen kan for eksempel slutte seg til et prinsipp om at å elske er en menneskerett, uten å bryte med den reflekterte likevekten, siden bakgrunnsteorien gir en forklaring på intuisjonen som går imot prinsippet.

Man oppnår bred reflektert likevekt når de tre nivåene gir hverandre gjensidig støtte:

- Intuisjonene kan forklares med kombinasjonen av prinsippene og bakgrunnsteoriene,
- prinsippene fremstår riktige, sett i lys av bakgrunnsteori og intuisjonene, og
- bakgrunnsteoriene stemmer overens med det man forventer ut ifra intuisjonene og prinsippene man har.

3.1.4 Hvorfor reflektert likevekt?

Peter Singer har argumentert for at en så bred variant av reflektert likevekt er innholdsløs, og ikke lenger har noe eget å tilby (Singer, 2005, s. 247–249). Med en bakgrunnsteori om at moralske intuisjoner er lite pålitelige kan man forkaste alle de intuitive oppfatningene metoden var ment å systematisere. Det vil i så fall ikke lenger være snakk om å finne noen likevekt mellom prinsipper og intuisjoner, men en fullstendig dominans av prinsipper over intuisjoner. I et slikt tilfelle vil reflektert likevekt sammenfalle med den aksiomatiske metoden, beskrevet i starten av dette kapittel 3.

Jeg mener derimot det er nyttig å beholde dette rammeverket for normativ argumentasjon. Metoden tydeliggjør hva det er man gjør i alle typer forskning, også den normative, nemlig å forsøke å finne den beste sannhetskandidaten. Uansett hvilke vitenskaper man forsker på og hvor pålitelige metoder man bruker, er en sunn grad av skepsis på sin plass (se f.eks. Popper, 2002, Kapittel 79). Selv de mest dagligdagse antakelsene hviler på en rekke forutsetninger, og deres troverdighet avhenger blant annet av om de inngår i et koherent system. Når man danner seg en oppfatning av hva det er man ser foran seg, baseres dette på en bakgrunnsteori om hvordan øynene og synet fungerer. Også denne bakgrunnsteorien trenger å være konsistent med andre oppfatninger for å bli akseptert. Troen på øynenes pålitelighet skyldes for eksempel at det man ser vanligvis er i overenstemmelse med det man hører eller føler – og tankerekken kan fortsettes i det uendelige. Og selv om øynene vanligvis er til å stole på og er en stabil kilde til kunnskap, så skjer det stadig vekk at man ser feil, og må forkaste den første oppfatningen man danner seg.

Man har altså aldri 100% sikker tilgang til sannheten.[15] Det man derfor er ute etter i forskning, er oppfatningene som man har mest grunn til å tro er sanne. Et av aspektene ved sannsynliggjøringen av en oppfatning, er at den må gå inn i et koherent system av holdninger som gir hverandre gjensidig støtte – slik som oppfatningen man får via øynene som samsvarer med andre sanseinntrykk, og bakgrunnsteorier angående hvordan synet fungerer.

Dette tydeliggjør også hvorfor jeg mener reflektert likevekt-metoden er å foretrekke fremfor den aksiomatiske fremgangsmåten. Moralske intuisjoner har riktignok ikke den samme epistemiske statusen i filosofien som sansedata har i empirisk forskning. Dette er fordi vi har bakgrunnsteorier som gir mer grunn til å betvile at intuisjoner følger av moralske sannheter enn at sanseinntrykkene samsvarer med omverdenen (se f.eks. Greene, 2014; Haidt, 2001). En aksiomatisk metode forkaster imidlertid alle intuisjoner. Det gjøres uten å vurdere om man kan skape et koherent bilde av intuisjoner, bakgrunnsteori og prinsipper, slik som med sansedata, teori om sansenes pålitelighet og prinsippene man slutter seg til i empirisk forskning. Dette forsikrer riktignok at en moralteori ikke blir bygget på intuisjoner som peker i feil retning. Samtidig får man heller ikke nyttiggjort seg av potensielt riktige innsikter slike oppfatninger kan bidra med. Å ignorere alle moralske intuisjoner, uten å engang ha vurdert hvorvidt de kan være nyttige, fremstår som sløsing med data. Heller enn å starte søken etter moralske prinsipper fra «bar bakke», tar jeg utgangspunkt i at intuisjoner i hvert fall kan gi noe innsikt i hva som er moralsk riktig. Om sannsynlige bakgrunnsteorier tyder på at visse intuitive oppfatninger ikke er til å stole på, forsøker jeg å ignorere disse intuisjonene.[16]

3.1.5 Hverken nødvendig eller tilstrekkelig – men likevel en indikasjon

Koherens er ikke tilstrekkelig for at en oppfatning skal være sann, eller engang berettiget (Elgin, 1996, s. 107). Mellom to punkter på et papir kan man tegne

[15] I formalvitenskap, slik som for eksempel matematikk eller logikk, kan det påstås at man har sikker kunnskap om hva som er sant. Dette er uansett ikke et viktig poeng i denne sammenheng.

[16] Tilnærmingen er i tråd med fenomenologisk konservatisme: Gitt at noe virker å være tilfellet, så tar jeg utgangspunkt i at det i det minste er noen grunn til å tro at så faktisk er tilfellet, med mindre jeg har informasjon som motstrider dette. For en grundigere beskrivelse av fenomenologisk konservatisme, se for eksempel Michael Huemers *Compassionate Phenomenal Conservatism* (Huemer, 2007, definisjon på s. 30).

uendelig mange forskjellige streker. Tilsvarende kan man forklare de samme intuisjonene med utallige ulike prinsipper og bygge koherente rammeverk rundt. Altså kan man oppnå forskjellige sett med reflekterte oppfatninger, prinsipper og bakgrunnsteorier i gjensidig støtte. Når man for eksempel ser barn i fattige land lide av sult, enkelt forebyggbare sykdommer eller på grunn av krig, tenker de fleste intuitivt at noe er galt, og man føler gjerne en trang til å bidra med noe. Det samme gjelder når man ser noen bli negativt forskjellsbehandlet i nabolaget sitt, for eksempel på grunn av legning eller hudfarge. Disse oppfatningene kan forklares av en rekke ulike teorier. For eksempel kan de skyldes at nevnte hendelser ikke maksimerer nytte, altså har de en utilitaristisk forklaring. En egalitær tankegang innebærer de samme oppfatningene: Siden personene i eksemplet trolig hadde det dårligere enn gjennomsnittsmennesket, ville det ført til mer likhet (i velferd) å hjelpe dem. Alternativt kan det være at man har en plikt til å støtte mennesker som opplever onder de ikke selv har forårsaket – og denne listen er langt fra utfyllende.

Målet med den refleksive prosessen er å vise hvilken av de mulige teoriene som er mest troverdig, for eksempel med å se hvilken som stemmer best overens med oppfatningene man antar er de mest berettigede, og som må gjøre færrest ad hoc løsninger for å beholde sin indre konsistens. Man undersøker hvilken teori som best samsvarer oppfatningene man har, og som best forklarer «feilen» med de intuisjonene som går imot prinsippet. Kandidaten som skaper likevekt på den mest troverdige måten, antar man er mest sannsynlig til å være riktig.

I tillegg til å ikke være tilstrekkelig, er koherens ikke engang nødvendig for å ha riktig oppfatning om et konkret spørsmål. Likevel fungerer det som en viktig indikasjon på at oppfatningen er berettiget. Det er mulig for en person å ha rett angående et tilfelle, samtidig som det fullstendige systemet av oppfatningene hens. Men siden vedkommende ikke kan ha rett i alt hen mener samtidig, er den manglende koherensen bevis på at hen tar feil om noe – kanskje også den aktuelle oppfatningen som diskuteres. Derfor er det mer grunn til å være skeptisk til en holdning som ikke er del av en reflektert likevekt, enn til en holdning som er det. Det er riktignok for mye å forlange at alle oppfatningene til en person skal harmonere perfekt. Tilnærmingen mister likevel ikke sin appell. Jo «nærmere» inkonsistensen er det aktuelle dilemmaet, altså jo viktigere og mer relevant for det man diskuterer inkoherensen er, desto mer svekkes en oppfatnings sannsynlighet til å være sann (Quine, 1981, s.

71). Man forventer for eksempel ikke at en kokk skal ha internt konsistente oppfatninger angående jordbruk, men om vedkommende også motsier seg selv i forbindelse med matlaging, gir det grunn til å betvile hens egenskaper som kokk.

I oppgaven tar jeg derfor utgangspunkt i noen intuisjoner angående fordelingsrettferdighet, og ser etter prinsipper som, i lys av bakgrunnsteorier, kan forklare disse. Å sammenligne mitt prinsipp med absolutt alle eksisterende alternativer er umulig i praksis. Jeg tar, som Rawls foreslår, utgangspunkt i kjente, konkurrerende prinsipper, og ser om noen av disse fremstår mer plausibel enn min kandidat, utilitarismen (Rawls, 1999, s. 43). Når jeg vurderer de ulike alternativene, ser jeg på hvilke implikasjoner de har, også utover de mest åpenbare. Med å undersøke disse kan jeg vurdere hvilke av de forskjellige prinsippene som, i lys av relevante bakgrunnsteorier, krever minst ad-hoc løsninger for å bli akseptert. På denne måten kommer jeg frem til den mest sannsynlige sannhetskandidaten. Før jeg setter i gang med det, derimot, er noen avklaringer på sin plass.

3.2 Konsekvensialisme, velferdisme og summering

Jeg beskrev utilitarisme som en teori som hviler på tre grunnpilarer, henholdsvis konsekvensialisme, velferdisme og summering. I denne oppgaven er det spesielt det siste aspektet jeg forsvarer.

Det jeg skriver minst om er konsekvensialismen. Jeg har allerede begrunnet hvorfor jeg primært ser på utfallsorienterte prinsipper. Sett i lys av dette er konsekvensialisme riktignok ikke en forutsetning, men i hvert fall en lite kontroversiell egenskap av utilitarismen i denne sammenheng.

I forbindelse med fordelingsrettferdighet er et naturlig spørsmål hva som bør fordeles. Diskusjonen til nå tyder på ressurser. I politikken er det som regel penger og rettigheter man får (om)fordelt mellom mennesker. Ressurser trenger derimot ikke være den relevante enheten for fordelingsrettferdighet. Som nevnt er et av de sentrale aspektene ved utilitarismen velferdisme. Siden jeg har satt ikke-menneskelige dyr til side i denne omgang, kan velferdisme oppsummeres med et sitat fra Joseph Raz: «Enhver forklaring og rettferdiggjørelse av hva som er godt eller ondt kommer til slutt fra dets faktiske eller potensielle bidrag for menneskelivet

og dets kvalitet» (Raz, 1986, s. 194, oversatt av meg).[17] En utilitaristisk fordelingspolitikk omfordeler riktignok ressurser mellom mennesker, men tilegner disse kun verdi i den grad de fører til velferd.

Teorier om hva som avgjør et menneskelivs kvalitet, altså hva som velferd er, deles gjerne inn i tre kategorier (Parfit, 1984, Kapittel I). Disse er henholdsvis hedonistiske teorier, altså teorier som definerer menneskelig lykke (og nytelse) som det verdifulle, teorier om preferansetilfredsstillelse, som anser det som verdifullt om mennesker får sine preferanser tilfredsstilt, samt objektive liste-teorier, som har en liste av ting som er goder eller onder for mennesket, uansett hvordan de selv opplever det.[18] Når jeg skriver om velferdisme i denne oppgaven, mener jeg de to første kategoriene, altså hedonisme og preferansetilfredsstillelse. Det virker umiddelbart sannsynlig at hvor godt et menneskeliv til syvende og sist, bør vurderes fra den aktuelle personens perspektiv. Med andre ord fremstår det sannsynlig at hvor lykkelig hen er, eller i hvor stor grad livet til vedkommende samsvarer med hens preferanser, gir uttrykk for hvor godt livet hens er. Siden livet er hens eget, fremstår det rart å skulle ha objektive kriterier til å vurdere det, fremfor å «høre på» hen selv.

Heller ikke velferdismen vier jeg spesielt mye plass. Det er flere grunner til dette. For det første, så faller en del av argumentene mot velferdisme bort, gitt oppgavens tema. Det kan være grunner til å tilegne iboende verdi til ting som ikke relaterer til mennesker og deres velvære. Slike hensyn virker derimot lite relevante i forbindelse med fordelingsrettferdighet. Gitt at jeg bare undersøker fordelinger mellom mennesker, virker velferdisme, som knytter verdi til følende vesen (f.eks. mennesker), som en svært plausibel teori.[19] Det er også en forutsetning for flere av de andre prinsippene jeg vurderer.

Velferd virker for eksempel umiddelbart som en riktigere kandidat til å være iboende verdifullt enn det ressurser er. Ressurser er riktignok viktige for mennesket. Å kunne spise og drikke nok, utdanne seg til det man ønsker og lignende, er ting man gjerne tenker på som verdifullt. Dette kan imidlertid forklares med at de fører til økt velferd. Mennesker får det bedre av å ikke være sultne og å studere det de ønsker. Ressurser forklarer derimot ikke andre ting vi anser for å være verdifulle, slik som

[17] Raz kaller dette for «det humanistiske prinsippet».
[18] Det er mulig å tolke muligheter-tilnærmingen jeg snart tar for meg som en form for objektiv-liste teori.
[19] Robert Shaver kaller f.eks. velferdismen som det mest tiltrekkende aspektet ved utilitarismen (Shaver, 2004).

vennskap, kjærlighet og artige opplevelser, den negative verdien man tilegner kjedsomhet eller fysisk- og psykisk smerte. Gitt at velferd er mer sannsynlig til å ha iboende verdi for mennesket enn ressurser, fremstår det riktig å basere rettferdighetsprinsipper på hvordan velferd fordeles, og la ressurser ha en instrumentell rolle.[20]

For det andre, så er de praktiske implikasjonene til velferdisme omtrent de samme som den andre markante kandidaten, muligheter-tilnærmingen («capability approach»), og følgelig er ikke denne distinksjonen like viktig. Sistnevnte tilnærming skiller seg fra velferdismen spesielt på to måter. Den er opptatt av hvilke reelle alternativer mennesker har, fremfor hva de faktisk gjør eller opplever (Robeyns & Byskov, 2020, Kapittel 0.0). Så mens velferdismen for eksempel tilegner det å spise et godt måltid en verdi, siden dette bidrar til velvære, vil det ifølge muligheter-tilnærmingen være verdifullt bare å ha muligheten til å spise måltidet, uansett om man benytter seg av dette.[21] Denne forskjellen mellom de to verdisynene har få praktiske implikasjoner, med mindre man vurderer fordelinger i svært nære relasjoner, der alle kjenner hverandre godt. I politikken (og andre liknende settinger), der avgjørelser påvirker mange ukjente mennesker, vet man ikke hva som fører til velferd for den enkelte. Det man derimot vet, er at mennesket foretrekker å gjøre det som gir dem selv velferd. Politikeres beste måte å fremme dette på er derfor å holde så mange muligheter som mulig åpne foran befolkningen. Slik får hver enkelt forhåpentligvis friheten til å velge det hen selv får økt velferd av. Siden alle muligheter ikke kan gjøres tilgjengelige samtidig, kreves det prioriteringer. I slike tilfeller er det naturlig å begynne med de frihetene som antas å gi mest velferd, slik som muligheten til god helse, kroppslig integritet og fritt følelsesliv. Dette er også punkter som gjerne nevnes på listen av friheter tilhengere av muligheter-tilnærmingen anser verdifulle (se f.eks. Nussbaum, 2003, s. 41–42). Frihet og fraværet av paternalisme fører trolig i tillegg til økt velferd, noe som også gjør de to tilnærmingene tilnærmet like i praksis.

[20] Et argument for ressurser fremfor velferd som jeg ikke diskuterer i oppgaven, er at det kan fremstå riktigere å fordele ressurser, for å så la det være menneskers eget ansvar å utnytte disse på en måte som fører til velferd. Sett i lys av kapittel 5, der jeg kommer frem til at mennesker ikke kan holdes moralsk ansvarlige, virker dette som et dårlig argument.

[21] Argumentet mot ressurser jeg beskrev i fotnote 20, gjelder også i forbindelse med muligheter-tilnærmingen. At jeg avviser konseptet moralsk ansvar i kapittel 5, gjør det lite sannsynlig at det er reelle friheter som har en iboende verdi, og ikke velferd. Heller ikke dette poenget diskuterer jeg videre i oppgaven.

Det samme praktiske poenget gjelder angående den andre forskjellen mellom synene. Muligheter-tilnærmingen har den fordelen overfor velferdisme at den fanger opp våre intuitive oppfatninger bedre når mennesker får velferd på en ukonvensjonell måte. Dette er spesielt iøynefallende i noen tilfeller der adaptive preferanser gjør at mennesker under dårlige kår lever med høyt nivå av velferd.[22] Martha Nussbaum eksemplifiserer fenomenet med å beskrive kvinner i ørkenområdet utenfor Mahabubnagar, som var alvorlig underernærte, og hvis landsby ikke hadde noen pålitelig forsyning av rent vann (Nussbaum, 2000, s. 113–114). Før bevissthets-program ankom, følte kvinnene tilsynelatende ikke på noe sinne eller stor lidelse, og de protesterte ikke på grunn av sin fysiske situasjon. De vurderte ikke forholdene sine som usunne eller for lite renslige, og de anså ikke seg selv for å være feilernært. Man kan altså si at de, målt i velferd, ikke hadde dårlige liv, tross sine omstendigheter. Dette fremstår feil.

Det er god grunn til å være skeptiske til våre intuisjoner i forbindelse med slike situasjoner. Underernæring og lite drikkevann er noe som, for meg og alle lesere av denne oppgaven, forbindes med uutholdelig lidelse. For mennesker vant med minst tre gode måltider om dagen, er det ofte ille nok å måtte stå over et av disse. Å stadig måtte leve på langt mindre ernæring enn det, i tillegg til å ikke kunne vaske seg ordentlig, er for mennesker som ikke er vant til dette helt forferdelig. Følgelig reagerer vi umiddelbart negativt bare til tanken. Når noe som vanligvis er et onde i en enkeltsituasjon ikke fører til lidelse likevel, er det ikke rart at velferdismens vurdering virker kontraintuitiv. Det er fordi intuisjonene våre fanger opp hvordan ting vanligvis er, ikke hvordan det er i noen kontraintuitive scenario. Når man vurderer adaptive preferanser med uendrede intuisjoner, er det ikke overaskende med upålitelige resultater. Adaptive preferanser tilbakeviser følgelig ikke velferdisme som en sannsynlig teori om iboende verdi.[23]

Samtidig gjelder poenget over også her. For alle praktiske formål sammenfaller muligheter-tilnærmingen og velferdismen, også på dette området. I noen enkeltsituasjoner der preferansene er endret i stor nok grad, kan liv være gode uten rent vann eller tilstrekkelig med mat. I de aller fleste tilfeller er imidlertid dette

[22] Adaptive preferanser viser til at mennesker tilpasser sine egne forventninger det de antar er realistisk basert på tidligere erfaringer.
[23] Jeg kommer tilbake til Nussbaums eksempel i del 4.2.1.

forutsetninger for at mennesker skal ha det godt. Følgelig er dette noe som bør prioriteres, også med en velferdistisk tilnærming til (fordelings)politikk.

I oppgaven skriver jeg derfor i velferdstermer. Jeg forutsetter at velferdisme er en svært plausibel teori angående iboende verdi, og den riktige enheten å bry seg om i forbindelse med fordelingsrettferdighet. Å fordele velferd direkte er riktignok umulig. Derfor bør ressurser fordeles på måten som antas å føre til ønsket fordeling av velferd. Jeg forblir agnostisk angående hva som er riktig av hedonisme og teorier av preferansetilfredstillelse. I del 4.3.2, der jeg sammenligner utilitarismen med Larry Temkins egalitarisme, gjør hans argumenter det enklere å omtale en spesifikk form for velferd. Denne gangen anvender jeg en hedonistisk argumentasjon for å forsvare utilitarismen. Dette er imidlertid kun gjort av pragmatiske årsaker. Jeg kunne like godt argumentert fra et preferansetilfredsstillelses-ståsted. Å gjøre begge ville derimot vært for repetitivt.

En begrepsavklaring er samtidig på sin plass i forbindelse med velferd. I diskusjon angående utilitarisme bruker jeg gjerne uttrykk som «lykke» og «glede», samt «smerte» og «lidelse». De to første er ment som uttrykk for positiv velferd, mens de to sistnevnte uttrykker negativ velferd. Det finnes trolig nyanseforskjeller mellom hva disse begrepene dekker, men for enkelhetens skyld bruker jeg de om hverandre. Jeg bruker i tillegg «nytte» som synonymt med «velferd».

Jeg diskuterer altså hverken velferdismen eller konsekvensialisme-aspektet i detalj. Det som primært skiller de ulike prinsippene jeg vurderer, er hvordan de rangerer forskjellige utfall av velferdsfordeling. Jeg argumenterer for at utilitarismens summering er det riktige, fremfor for eksempel en lik fordeling, det å vekte de med lavest velferdsnivå tyngre eller det å ta hensyn til fortjeneste. Når det er sagt, så ønsker jeg å oppnå bred reflektert likevekt i forbindelse med problemstillingen. Metoden innebærer nødvendigvis å noen ganger gå utenfor de nevnte avgrensningene, for å se om mitt svar inngår i et koherent bilde. Også de to andre aspektene ved utilitarismen berøres, så vel som flere underkategorier av rettferdighetsbegrepet. Fokuset er imidlertid utilitarismens summering i forbindelse med fordelingsrettferdighet.

Med dette starter jeg den refleksive prosessen, og sammenligner utilitarismen med alternative fordelingsprinsipper. Jeg tar først for meg Rawls' teori, som trolig er den

mest kjente på området. Deretter vurderer jeg henholdsvis egalitarisme, prioritarianisme og teorier om tilstrekkelighet. Etter å ha diskutert disse prinsippene, fortsetter jeg med å undersøke om fortjeneste er et relevant konsept for fordelingsrettferdighet.

4 Alternative fordelingsprinsipper

4.1 Rawls' teori

For å etablere en teori om rettferdige samfunnsstrukturer, beskriver Rawls et tankeeksperiment. I en hypotetisk utgangsposisjon skal representanter for de forskjellige samfunnsgruppene resonnere seg frem til en samfunnskontrakt, som de velger ut fra en liste med alternativer. Valgene skjer bak et tykt slør av uvitenhet. Dette uvitenhetsløret gjør at aktørene ikke kjenner til informasjon om seg selv som kan skille dem fra andre samfunnsmedlemmer (Rawls, 1999, s. 118). De vet for eksempel ikke hvilket kjønn, etnisitet eller livssyn de har, hvilken sosial/økonomisk klasse av samfunnet de tilhører, hvilke talenter og ferdigheter de innehar. De vet heller ikke hvor risikovillige eller risikoaverse de er. De mangler også mye kunnskap om det aktuelle samfunnet de skisserer grunnprinsippene for. De vet ikke hvordan den politiske eller økonomiske situasjonen er, eller hvor langt sivilisasjonen deres er kommet i den kulturelle utviklingen.[24] Representantene i utgangsposisjonen er stipulert å være rasjonelle, men har kun tilgang til veldig generell informasjon, slik som kunnskaper om menneskelig psykologi, om hvordan sosialt samarbeid fungerer og om økonomiske teorier (Rawls, 1999, s. 119).[25]

Med disse begrensningene på aktørenes kunnskapsgrunnlag, unngår Rawls at prinsippene de ender opp med å gjenspeile noen personers egeninteresser. Om representantene kjente til sine talenter eller sin sosioøkonomiske posisjon, ville de bedre stilte kunne utnyttet sin konkurransefordel til å forhandle frem en mer skreddersydd kontrakt. Moralsk arbitrære forhold ved personer kunne endt opp som svært avgjørende for hvor godt de ble representert i samfunnskontrakten.[26]

[24] Det antas riktignok at samfunnet kun har en moderat knapphet av ressurser (Rawls, 1999, s. 110).
[25] Gitt deres rasjonalitet kombinert med manglende kunnskaper om seg selv, gir det strengt tatt like mye mening å snakke om én representant som flere representanter, da alle antas å velge likt. I det videre vil jeg omtale det som flere representanter.
[26] Moralsk arbitrære forhold ifølge Rawls, i hvert fall. Bakgrunn for hele teorien er Rawls' syn på det han kaller «det naturlige lotteriet», ofte kalt «moralsk flaks» (Rawls, 1999, s. 63–64). Foruten personers bakgrunn og talenter mente Rawls heller ikke at for eksempel egenskaper som arbeidsmoral faller utenfor denne

Uvitenhetens slør fjerner disse mulighetene, da personer ikke får kunnskaper om hvem de selv er, og følgelig ikke kan ta hensyn til deres egenskaper. Sløret fremstår som et svært godt verktøy for å illustrere viktigheten av upartiskhet når en skal ta moralske avgjørelser – et hensyn som har vært poengtert av mange filosofer tidligere, ikke minst innenfor den utilitaristiske tradisjon.[27]

Under disse omstendighetene mener Rawls at representantene vil komme frem til følgende prinsipper:

> *Det første prinsippet*:
> Enhver person skal ha den samme rett til det mest omfattende system av grunnleggende friheter som er forenlig med et tilsvarende system av friheter for alle.
>
> *Det andre prinsippet*:
> Sosiale og økonomiske ulikheter skal tillates
>
> > (a) hvis og bare hvis de er til størst mulig fordel for de dårligst stilte i samfunnet, og
> >
> > (b) bare hvis de er knyttet til stillinger og posisjoner som alle kan konkurrere om på rimelige og like vilkår (Rawls, 1999, s. 266).

De to prinsippene er i leksikalsk rekkefølge, altså har det første prinsippet absolutt prioritet over det andre. Innenfor det andre prinsippet har (b) absolutt prioritet over (a). Prinsipp 2(a), også kalt forskjellsprinsippet, er det mest relevante i forbindelse med fordelingsrettferdighet.

4.1.1 Forskjellsprinsippet

kategorien, da det til syvende og sist er forhold utenfor ens egen kontroll, slik som familiesituasjon og hvem man omgås med som påvirker dette. Ikke alle vil være enige med Rawls i at disse er moralsk irrelevante faktorer, men siden jeg selv deler hans syn, vil jeg ikke problematisere det i denne omgang. Mer om dette i kapittel 5. For en diskusjon av Rawls' syn på moralsk ansvar, se side 213-223 av *Equal Opportunity and Moral Arbitrariness* (Barry, 2012).

[27] Uten å bruke det samme uttrykket innfører John C. Harsanyi i praksis uvitenhetsløret når han argumenterer for sin versjon av utilitarismen (Harsanyi, 1953, s. 435, 1955, s. 316). Mer generelt poengteres moralens upartiskhet (og menneskers likeverd) hos alle velkjente utilitarister, som for eksempel Jeremy Bentham, John Stuart Mill, Henry Sidgwick, R. M. Hare og Peter Singer (Bentham, 1966, s. 559; Mill, 1863, s. 91; Sidgwick, 1981b, s. 417; Hare, 2003, s. 10; Singer, 2011, s. 20).

Forskjellsprinsippet gir uttrykk for maximin-kriteriet. Som navnet tilsier innebærer dette kriteriet at man maksimerer gevinsten ved det verste mulige utfallet (Haddad, 2005, s. 167).[28] Altså er det den fordelingen der den dårligst stilte har mest, som vurderes som best av maximin-kriteriet.[29]

En ulikhet mellom Rawls' prinsipp 2(a) og utilitarismen, er at førstnevnte på fordeler ressurser, mens sistnevnte fokuserer på velferd. I forbindelse med dette kritiserer Rawls utilitarismen for å være for vag, og lanserer sin teori som et alternativ med klarere svar (Rawls, 1999, s. 149–150, 182, 281). En utfordring med utilitarismen er å skulle sammenligne forskjellige menneskers lykkenivåer.[30] Rawls' argument kan oppsummeres slik: Det er umulig å kjenne til de konkrete opplevelsene til andre personer enn en selv. Vi reagerer ulikt i forskjellige situasjoner, men har kun introspektive bevis for våre egne erfaringer. Følgelig er det i beste fall en grov forenkling å sette alle menneskers velferdsnivåer på en felles skala, noe utilitarismen forutsetter for å kunne rangere alle utfall etter hvor mye nytte de skaper.

Som jeg var inne på i del 3.2, kan denne innvendingen legges til side. Velferd fremstår som et riktigere mål på hva som er verdifullt enn ressurser. Verdien vi gjerne tilegner ressurser kan forklares med at de er et middel til å oppnå velferd. Det motsatte er derimot ikke sant: Ikke alle velferd-relaterte momenter vi tilegner verdi kan forklares med at det egentlig er ressurser som har iboende verdi. Nettopp på grunn av vanskelighetene med å måle velferd Rawls poengterer, bygger imidlertid en utilitaristisk fordelingspolitikk på antakelser, der man fordeler ressurser på måten

[28] Siden Rawls omtaler regelen som «maximin», og uttrykket gjerne forbindes med «A Theory of Justice», vil også jeg bruke denne betegnelsen. I presisjonens navn er det verdt å nevne at det er nærliggende å tro at Rawls faktisk argumenterer for «leximin-kriteriet». Dette er i utgangspunktet det samme som maximin, mht. at det er det dårligste scenarioet som maksimeres. Men ved eventuelle «uavgjortresultat» fører leximin til en sammenligning av de nest dårligste scenarioene (og deretter evt. tredje dårligste osv., om nest dårligste også var likt) (Haddad, 2005, s. 169).

[29] Samme tankegang er også en del av begrunnelsen for de to andre prinsippene. Som Rawls selv skriver, kan hele hans teori sees på som maximin-løsningen på «problemet» sosial rettferdighet utgjør (Rawls, 1999, s. 132). En viktig del av hans argumentasjon for å gi like friheter leksikalsk prioritet, virker å være en antakelse om at dette gagner de dårligst stilte mest. Dette fordi det sikrer at de kan fortsette å etterstrebe sine mest fundamentale interesser, og fordi det bidrar til en følelse av selvrespekt i befolkningen (Rawls, 1999, s. 476–478). Å i stor grad tolke verdien av frihet som instrumentell for velferd, gir også mening mtp. at prinsippene *velges* fra bak uvitenhetens slør, altså er de uttrykk for preferanser. Kritikk mot maximin-kriteriet kan derfor trolig være relevant for hele Rawls' rettferdighetsteori, ikke kun forskjellsprinsippet. Både å vurdere verdien av frihet, og andre aspekter ved Rawls' teori enn det rent fordelingsrelevante, er imidlertid utenfor min problemstilling, og jeg legger det derfor til siden. For mer om dette, se f.eks. kapittel 8-9 av *Facing up to Scarcity* (Fried, 2020, Kapitler 8–9).

[30] Her låner Rawls refleksjoner gjort av Amartya Sen (Sen, 1970, s. 92–99).

man regner med fører til mest aggregert velferd. Sett i lys av at det er velferd som virker iboende verdifullt, bør også forskjellsprinsippet gjenspeile dette. Heller enn å fokusere på at de dårligst stilte bør få så mange ressurser som mulig, er en mer plausibel maximin-tilnærming til fordelingsrettferdighet å velge ressursfordelingen som antas å føre til mest mulig velferd til den dårligst stilte.[31] I videre diskusjon er det slik jeg tolker forskjellsprinsippet.

Utilitarismen og forskjellsprinsippet skilles derfor av hvilken strategi som brukes for å rangere forskjellige utfall. For utilitarismen er dette summering av velferd. I Rawls' teori er det maximin-kriteriet. Rawls mener at hans prinsipper er det rasjonelle å godta i den hypotetiske utgangsposisjonen. I de neste sidene undersøker jeg hans grunner til dette, og argumenterer for at utilitarisme er et bedre valg, også bak uvitenhetens slør.

Før det er en presisering på sin plass. Som allerede nevnt, skriver Rawls en teori om rettferdige samfunns*strukturer*. Hans prinsipper, inkludert maximin-kriteriet, er ment å gjelde i forbindelse med valg av innretningen av basis-institusjoner, ikke allokeringssituasjoner direkte. I sistnevnte situasjoner åpner han til og med for at å gjøre fordelinger basert på hva som maksimerer nytte kan være det riktige (Rawls, 1999, s. 77). Hans begrunnelse for å velge en slik prosedural rettferdighetsteori virker imidlertid være praktiske årsaker. Men et slikt system slipper man å holde oversikt over det uendelige mangfoldet av omstendigheter som oppstår (Rawls, 1999, s. 76). Fordelingsrettferdighet kan «overlates til seg selv», og det går likevel bra, så lenge prosedyrene følges.

I del 2.1.2 beskrev jeg derimot hvorfor utfallsorienterte prinsipper er å foretrekke når det gjelder fordelingsrettferdighet. Tatt i betraktning at Rawls forutsetter full etterlevelse av hans prinsipper i sin rettferdighetsteori, bør prosedyrene han foreslår basert på maximin-prinsippet stort sett resulterer i de samme fordelingene som et utfallsorientert maximin-prinsipp ville gjort. Dette siden institusjonene og strukturene som er best for den dårligst stilte bør føre til de beste utfallene for den dårligst stilte, gitt full etterlevelse. Videre i oppgaven behandler jeg derfor forskjellsprinsippet som

[31] Det er nærliggende å tro at dette faktisk er en del av begrunnelsen for forskjellsprinsippet. Kombinerer man frihetene Rawls' andre prinsipper omhandler med ressursene fra forskjellsprinsippet, får man i praksis en slags muligheter-tilnærming til verdi, der ressursene gjør at de formelle frihetene som er sikret blir til reelle muligheter. Som beskrevet i del 3.2, overlapper denne tilnærmingen med velferdismen i praksis.

et utfallsorientert fordelingsprinsipp, basert på maximin-kriteriet. Sett i lys av refleksjonene over, mener jeg dette er en relevant tolkning av Rawls' teori. Følgelig poengterer jeg ikke videre i oppgaven at Rawls' refleksjoner i utgangspunktet omhandler prosedyrer fremfor utfall. I rettferdighetens navn er det imidlertid verdt å huske at innvendingene jeg gjør mot Rawls primært gjelder min egen fordelingsrettferdighet-relevante tolkning av hans prinsipper.

4.1.2 Grunner til å velge maximin-kriteriet

I det hverdagslige ville maximin-kriteriet vært lite rasjonelt å bruke. John Harsanyi eksemplifiserer dette med en person som har fått to jobbtilbud (Harsanyi, 1975, s. 595). Den ene jobben virker kjedelig og er svært dårlig betalt, men er i vedkommendes hjemby. Den andre betaler godt og fremstår spennende. Arbeidsplassen er i en by langt unna, men personen har ikke noe imot å flytte dit. Tar vedkommende i bruk maximin-kriteriet når hen tar avgjørelsen, kan hen imidlertid ikke flytte. Siden flyturen dit innebærer en svært liten, men virkelig risiko for flystyrt og død, er det verste mulige utfallet dårligere ved å ta den gode jobben, enn å bli i hjembyen. Maximin-kriteriet rangerer ulike alternativer kun basert på det verste mulige utfallet. Siden dette ved den dårlige jobben kun er å være fattig og kjede seg, noe som er bedre enn å dø, fordrer maximin-kriteriet å ra den dårlige jobben.[32] Dette fremstår direkte feil. Om vedkommende derimot tar i bruk den forventede nyttemaksimeringen som utilitarismen bygger på, får vedkommende flyttet til den gode jobben, da dette så å si alltid fører til det beste utfallet.

Rawls innrømmer selv at maximin-kriteriet generelt ikke er en regel til etterfølgelse, og beskriver sine prinsipper som de en person ville valgt, gitt at hen visste at hens verste fiende fikk bestemme hvor hen selv ble plassert i det aktuelle samfunnet (Rawls, 1999, s. 133). Slike forbehold er derimot ikke tatt i den hypotetiske utgangsposisjonen – aktørene har ingen informasjon om hvor de havner i det nyetablerte samfunnet. Sjansen for å havne «nederst» er ikke kjent. Rawls må derfor sannsynliggjøre at det er riktig strategi i akkurat denne situasjonen. Han peker på

[32] En mulig innvending er at det også finnes en risiko for død ved å bli værende i hjembyen, altså er det verste mulige utfallet i de to alternativene de samme. Dette gjør imidlertid maximin-kriteriet uegnet til å ta en rekke avgjørelser, og gir grunn til å heller anvende leximin-kriteriet, som i slike tilfeller sammenligner nest verste utfall. Siden flyreisen øker risikoen for død, gjelder argumentet jeg beskriver.

spesielt tre aspekter som gjør en slik tankegang spesielt passende bak uvitenhetens slør (Rawls, 1999, s. 134):

1. At aktørene ikke engang kjenner til sannsynlighetene for å havne i de forskjellige samfunnsposisjonene,
2. at deres verdisyn er slik at de ikke bryr seg nevneverdig om mulige gevinster over minimumsnivået som maximin-prinsippet sikrer
3. at alternative tilnærminger innebærer en mulighet for uakseptable utfall.

Aspekt nr. 2 kan man umiddelbart se bort ifra som et argument for maximin-prinsippet fremfor utilitarisme. For det første er det verdt å merke seg at Rawls' argumentasjon fremstår noe ad hoc. I beskrivelsen av utgangsposisjonen og uvitenhetens slør nevner han eksplisitt at representantene ikke kjenner til sitt verdisyn eller interesser/tanker om «det gode» (Rawls, 1999, s. 11). Sett i lys av det fremstår det som en motsigelse å basere teorien på at aktørene har spesifikke tanker om det gode liv likevel. Om man derimot aksepterer verdisynet Rawls tilegner representantene, vil dette også bli fanget opp av utilitarismen. At nytten av for eksempel penger avtar i det man får mer av det er velkjent (fenomenet kalles gjerne avtakende utbytte). Om aktører får betydelig mer lykke ut av et minimumsnivå av ressurser enn det de vinner på tilsvarende mengder over dette nivået, vil maximin-prinsippet og utilitarismen sammenfalle. En fattig person som spiser to måltider om dagen får for eksempel en adskillig større lykkeøkning av ett kilo brød i uken ekstra, enn det en mett millionær ville gjort.

Det er her verdt å merke seg at Rawls viser til økonomen William Fellner når han peker på de tre forholdene ved utgangsposisjonen som gjør maximin-prinsippet passende. Fellner selv poengterer imidlertid at argumentet faller bort i det man vurderer lykke fremfor ressurser, siden en fordeling av lykke allerede har tatt høyde for ressursenes avtagende utbytte (Fellner, 1965, s. 142). Dette er altså en fordel Rawls' teori har sammenlignet med en «forventet ressursmaksimering», men ikke sammenlignet med utilitarismen. Jeg utdyper dette poenget mer i del 4.6.

Aspekt nummer 1, altså at aktørene ikke kjenner til sannsynlighetene ved ulike utfall, og argument nummer 3, at alternative tilnærminger innebærer muligheten for uakseptable utfall, sees best sammen. At man ikke kjenner til sannsynlighet ved utfallene, er spesielt problematisk hvis det finnes uakseptable utfall. Og at det finnes

uakseptable utfall er ekstra bekymringsfullt idet man ikke kjenner til deres sannsynlighet, og dermed ikke kan utelukke at den er høy. Samlet tilsier disse punktene at vi bør være risikoaverse. Rawls antar at aktørene bak uvitenhetens slør velger bort alle potensielle gevinster, for å unngå at verste mulige utfall blir for ille. Det tar altså i bruk maximin-kriteriet, og gjør det verste utfallet så godt som mulig. Han beskriver flere grunner til å anse dette som en rimelig hypotese. En av disse er hans refleksjoner rundt sannsynlighetsberegninger i mangelen på informasjon.

I tilfeller der man ikke har noe informasjon å ta utgangspunkt i, slik som utgangsposisjonen sier «prinsippet om utilstrekkelig grunnlag» at man bør ta utgangspunkt i at alle mulige scenarioer har akkurat samme sannsynlighet til å inntreffe.[33] Som Rawls påpeker er dette en betingelse for at en selvsentrert aktør skal velge utilitarismen bak uvitenhetens slør. Rawls referer til John Maynard Keynes når han avviser det som en legitim logikk å bruke bak uvitenhetens slør (Rawls, 1999, s. 146, fotnote 26).

I eksemplet stipulerer Keynes at vi ikke kjenner til befolkningstallet i forskjellige land, og prøver å gjette hvor en tilfeldig person kommer fra (Keynes, 1952, s. 44). I dette tilfellet vil prinsippet av utilstrekkelig grunnlag tilsi at vedkommende er like sannsynlig å være fra Frankrike som Storbritannia. Prinsippet tilsier også at sjansen for å være fra Frankrike og Irland er lik. Også Frankrike og De britiske øyer vil tilegnes samme sannsynlighet.[34] Dette mener Keynes er uholdbart, da de første to premissene samlet tyder på at det er dobbelt så sannsynlig at personen er fra De britiske øyer som fra Frankrike.

Keynes virker her å gjøre en feilslutning. Anslaget om at De britiske øyer og Frankrike er like sannsynlige, er gjort uten kunnskapen om at førstnevnte er en samlebetegnelse på to land. Så lenge vi ikke vet noe mer enn at Frankrike og De britiske øyer er eksisterende land, er det ingen grunn til å tro at personen er mer sannsynlig å komme fra det ene, enn det andre. Når vi derimot får vite at vi med «De britiske øyer» ikke mener ett land, men to, er det naturlig å gjøre en ny kalkulering. I

[33] Prinsippet kalles gjerne også Laplaces teorem, siden det først ble beskrevet av Pierre-Simon Laplace (Laplace, 1902, s. 11)

[34] «De britiske øyer» er for øvrig ikke et land. Problemet ville derfor i utgangspunktet aldri oppstått, da det ikke er et av landene det ville vært naturlig for oss å tippe at personen er fra. Dette poenget kan likevel settes til side for eksempelets skyld.

denne situasjonen vet vi tross alt at personen er fra De britiske øyer både om han er fra Irland og om han er fra Storbritannia. Dette er ikke motstridende med det første anslaget, da de to ulike resultatene er basert på forskjellig kunnskapsgrunnlag. At man, ved å lære ny relevant informasjon, oppdaterer sine anslag, skulle bare mangle. Keynes skriver riktignok at kunnskapen om at De britiske øyer består av Irland og Storbritannia ikke fremstår som en god nok grunn til å gjøre den oppdaterte gjetningen. Han begrunner derimot ikke sin aversjon mot dette, og denne skepsisen virker derfor lite berettiget

På samme måte virker prinsippet om utilstrekkelig grunnlag å være et nyttig verktøy i Rawls' hypotetiske utgangsposisjon. Så lenge aktørene bak uvitenhetens slør ikke kjenner til informasjon om sannsynlighetene til å være en spesifikk person av de forskjellige menneskene i samfunnet, har de ingen grunn til å gå ut ifra at det ene er mer eller mindre sannsynlig enn det andre. Det rimeligste er da å anta at sannsynlighetene for de forskjellige utfallene er like, heller enn å opptre som om det var sikkert at verste mulige scenario inntreffer, som maximin-tankegangen i praksis innebærer.

Rawls' umiddelbare mistro til prinsippet om utilstrekkelig grunnlag virker altså ikke å være passende, noe som er et viktig poeng, all den tid dette fremstår som et av hans hovedgrunner til å foretrekke maximin-kriteriet i utgangsposisjonen fremfor utilitarisme. Han nevner riktignok også to andre grunner til å være så risikoaverse bak uvitenhetens slør: Henholdsvis at valget har svært omfattende konsekvenser, og at det må kunne forsvares overfor resten av samfunnets borgere, som også påvirkes av avgjørelsen. Ingen av disse virker spesielt overbevisende. Som John Harsanyi påpeker i sin kritikk av maximin-prinsippet, ville det vært rart om spørsmålets størrelsesorden skulle ha noe å si for hva som er riktig prinsipp å tilnærme seg det med (Harsanyi, 1975, s. 605). Rawls stipulerer at aktørene bak uvitenhetens slør er rasjonelle. De behøver de ikke å opptre med maksimal risikoaversjon, selv i så betydelige avgjørelser.

Heller ikke at andre påvirkes av valgene en representant tar, stiller maximin-prinsippet i bedre lys. Igjen virker argumentet noe ad hoc, ettersom aktørene bak uvitenhetens slør er ment å være likegyldige til hvordan prinsippet påvirker medborgere. De prøver hverken å hjelpe eller skade andre enn dem selv (Rawls,

1999, s. 125). Man kan kanskje se bort fra denne inkonsistensen, med å argumentere for at en selvsentrert aktør ønsker å unngå ubekvemmelighetene det medfører å måtte forsvare valget de tok overfor skuffede personer som ble negativt påvirket av samfunnskontrakten. Men i så fall fremstår det merkelig å velge et prinsipp som kun tar hensyn til én person i fordelingen av goder. Den dårligst stilte vil riktignok ikke ha grunn til å klage til representanten som stemte for maximin-prinsippet. Men alle andre borgere blir nedprioritert av prinsippet for å gjøre det verste mulige scenarioet bedre, og vil kunne skylde på representanten for dette. Rawls kritiserer utilitarisme for å være altfor avhengig av at mennesker er empatiske med hverandre til at teorien skal være en fristende strategi bak uvitenhetens slør (Rawls, 1999, s. 154). Dette fordi noen uheldige potensielt må «ofre seg» for at det totale nivået av nytte skal maksimeres. Men i praksis fremstår det som at det er hans teori som er mest avhengig av slike egenskaper i befolkningen for å kunne bli akseptert. Med maximin-prinsippet ofrer hele resten av samfunnet egne gevinster for at den dårligst stilte skal få det så bra som mulig. Det krever urealistisk mye velvilje fra samfunnet som helhet – og gjør at utilitarismen fremstår som mindre belastende å velge bak uvitenhetens slør, om man er redd for å bli kritisert i etterkant.

Foruten risikoaversjon viser Rawls til enda et teoretisk argument for sine prinsipper overfor utilitarismen, med å vise til at hans teori er mer i overenstemmelse med Immanuel Kants kategoriske imperativ.[35] Humanistformuleringen av Kants kategoriske imperativ er som følger: «Handle slik at du alltid bruker menneskeheten både i din egen person og i enhver annens person samtidig som et formål og aldri bare som et middel.» (Kant, 1970, s. 42). Rawls velger å «tolke formuleringen fritt i lys av kontraktdoktrinen» (Rawls, 1999, s. 155–156). Minstekravet for å behandle mennesker som mål i seg selv, er ifølge Rawls å behandle dem i tråd med prinsipper de ville samtykket til i en utgangsposisjon av likeverdighet. Dette endrer imidlertid

[35] Rawls fremmer i tillegg et argument for eget prinsipp fremfor utilitarismen som jeg ikke adresserer. Innvendingen Rawls gjør mot utilitarismen er at å offentlig skulle enes om en utilitaristisk samfunnskontrakt kan være demotiverende for befolkningen, siden den ikke gir noen formelle forsikringer mot dårlige utfall, i form av rettigheter eller lignende. Følgelig vil den utilitaristiske samfunnskontrakten føre til tap av velferd. Hvis Rawls' premiss stemmer, fordrer utilitarismen å enes om noe annet enn utilitarisme, noe som ifølge Rawls er selvmotsigende. Dette er imidlertid ikke intern kritikk av utilitarismen – en utilitarist kan akseptere at å handle av ikke-utilitaristisk motivasjon noen ganger vurderes riktig av utilitarismen. Derfor bryter det ikke med utilitaristers reflekterte likevekt. Følgelig skriver jeg ikke mer om innvendingen. For detaljerte svar, *se Secrecy in consequentialism: A defence of esoteric morality*, samt første kapittel av *Reasons and Pearsons* (de Lazari-Radek & Singer, 2010; Parfit, 1984, Kapittel 1).

ikke på noen ting. *A Theory of Justice* er allerede et forsøk på å finne prinsippene mennesker ville enes om i hans hypotetiske utgangsposisjon. Men Rawls går videre, og mener at hans prinsipper er mest i tråd med Kants idé, da de sikrer at de dårligst stilte ikke blir «ofret» for andres gevinst - slik som teknisk sett kan skje i utilitarismen. Som Harsanyi imidlertid påpeker, tillater også Rawls' prinsipper at noen personers interesser velges bort til fordel for andre – noe som ikke er så rart (Harsanyi, 1975, s. 597). I en verden med begrensede ressurser, der ikke alle kan få alle godene de ønsker, vil slike valg alltid være nødvendige. Ifølge maximin-prinsippet må alle menneskers gevinster ofres, uansett størrelse, om de ikke bidrar til at den dårligst stilte får det bedre. Å påstå at dette ikke er å behandle alle utenom sistnevnte som et middel, samtidig som man kritiserer utilitarisme for ikke å være i tråd med Kants doktrine, fremstår svært lite konsekvent.

Alternativt er Rawls' poeng at representanter i en posisjon av likeverdighet i utgangsposisjonen ikke ville samtykket til utilitarismen, men til maximin-kriteriet og leksikalsk prioritet til frihet. I dette tilfellet tilfører ikke hans henvisning til Kants kategoriske imperativ noe ekstra – poenget hans er fortsatt avhengig av at maximin-kriteriet faktisk er rasjonelt å velge bak uvitenhetens slør.

4.1.3 Implikasjoner i praksis

Til nå har jeg tatt for meg teoretiske argumenter knyttet til «A Theory of Justice», som omhandler prosessen med valg av prinsipper bak uvitenhetens slør. Avslutningsvis er det også interessant å undersøke hva henholdsvis utilitarismen og maximin-prinsippet impliserer i praksis, for å se hvilke av de to teoriene som fører til de rimeligste konsekvensene. Rawls advarer riktignok mot å gi argumentasjon i form av moteksempler for mye vekt, da disse har en tendens til å påpeke små feil i teorien, heller enn å fokusere på det store bildet (Rawls, 1999, s. 45). Samtidig går en betydelig del av hans kritikk mot utilitarismen ut på at den for eksempel ikke utelukker slaveri (Rawls, 1999, s. 144, 444). Å se på hva de forskjellige prinsippene innebærer i praksis, både av sannsynlige og mindre sannsynlige scenarioer, er relevant for å finne ut av hvilken teori som bør velges bak uvitenhetens slør Jeg tar først for meg Rawls' kroneksempel mot utilitarismen, at det potensielt tillater slaveri. Deretter beskriver jeg noen uheldige implikasjoner av maximin-kriteriet.

Rawls og andre kritikere har rett i at utilitarismen ikke utelukker slaveri som sådan. Teorien sier at den totale velferden bør maksimeres, men ikke noe om hvordan. Ingen institusjoner, handlinger, motiver, ulikheter eller lignende er forbudt uten videre. Bedømmelsen av alt dette kommer an på hvor mye lykke det fører til, eller rettere sagt, hvor mye det fører til i forhold til alternativene. Så slaveri er virkelig ikke utelukket av utilitarismen uten videre. Dette er en egenskap ved prinsippet som umiddelbart fremstår avskrekkende. For at utilitarismen skal fortsette å være et fristende alternativ, bør den ikke være kompatibelt med slikt. Har Rawls avdekket en avgjørende svakhet i teorien?

Det korte svaret er nei. Utilitarismen er ikke kompatibel med slaveri slik vi kjenner den i den virkelige verden, rett og slett fordi slaveri fører til mindre aggregert velferd enn det mangelen på slaveri gjør. Det fremstår direkte umulig at det å innføre slaveri i et hvilket som helst land i dag ville ført til en økning i total lykke der. Og motsatt, det er ingen grunn til å tvile på at det er et betydelig gode, også fra et utilitaristisk perspektiv, at slaveri ble avviklet i Storbritannia 1807, de danske koloniene og i Frankrike i 1848, i USA i 1863 og så videre. I den sammenheng viser Joshua Greene til et tankeeksperiment (Greene, 2013, s. 277): Hvis du måtte leve halve livet ditt som slave, for å kunne være slaveeier den andre halvparten, ville du gjort det?[36] Det fremstår rimelig klart at svaret på det spørsmålet er nei. Og at en slave kan eies av flere, endrer lite på dette. Selv om det fører til færre lidende slaver, må slaveeierne i så fall dele på godene. Så spørsmålet endres da for eksempel til dette: Ville du levd en tredjedel av livet ditt som slave, for å kunne være slaveeier i en annen tredjedel? Dette er også et alternativ man umiddelbart takker nei til.

Man ønsker ikke å bytte x antall år av livet sitt for å være slave mot like mye tid som slaveeier, rett og slett fordi det å være slave ville kostet mer i lykke enn det å eie en slave ville veid opp for. For å forklare hvorfor, viser Greene til hva en slave først og fremst er for en slaveeier, nemlig gratis arbeidskraft. Med andre ord kan det å eie en slave i stor grad «veksles» i penger, altså en form for ressurs. Ressurser gir som kjent stadig mindre utbytte i form av lykke, jo rikere man er. Pengene slaveeierne tjener gir altså adskillig mindre lykke for dem, enn det slavene taper på gratis arbeid. I tillegg kommer alle andre grusomheter det medfører å være slave, som er adskillig

[36] Greene bruker flere sider av boken sin på «slave-innvendingen» til Rawls, så tankeeksperimentet og bakgrunnen er noe mer nyansert. For hans refleksjoner rundt temaet, se Greene (Greene, 2013, s. 275–284).

større tap enn gevinstene ved å være eieren.[37] Med andre ord er eksistensen av slaveri et netto tap for utilitaristen. Selv om utilitarismen ikke tilegner den dårligst stilte større vekt enn resten av befolkningen, er slaveri noe den aldri ville anbefalt i den virkelige verden.[38]

Som utilitarismen, vil maximin-prinsippet heller ikke tillate slaveri. Samfunnets dårligst stilte prioriteres alltid når man følger maximin-kriteriet, og å tillate slaveri vil i praksis aldri være positivt for vedkommende. I dette tilfellet trenger ikke slavenes lidelse å overskygge slaveeiernes gevinster engang for at slaveri skal fordømmes, med maximin-prinsippet er det nok å vurdere tiltaks effekt på den dårligst stilte.

Diskusjonen over viser at utilitarismen og maximin-prinsippet ofte kan sammenfalle. Dette er ikke så rart, av grunnene allerede nevnt. Siden ressurser gir avtakende avkastninger for folks velferd, vil det som regel være billigere å hjelpe de som har det verst. I en verden uten uendelig med ressurser betyr det at man har mulighet til å hjelpe de dårligere stilte mer enn de som har det bedre. Problemet er at de gangene de to teoriene faktisk skilles, virker maximin-prinsippet å gi de dårligere svarene. Harsanyi beskriver noen gode eksempler (Harsanyi, 1975, s. 596–597, 605–606).[39] Man kan se for seg at landets sykehus har færre doser igjen av antibiotika enn det som er nødvendig for å gi til alle som trenger. De innlagte varierer mye i typen sykdom, og følgelig også i hvor mye en dose antibiotika kan være til hjelp. Mange av pasientene, for eksempel de som har lungebetennelse, kan gjøres helt friske med hjelp av antibiotika. Samtidig er det også noen pasienter som kun får det marginalt bedre av medisinen – for eksempel en person med uhelbredelig kreft og lungebetennelse samtidig, som kun får noen dager ekstra å leve av medisinen. Følger man maximin-prinsippet, er det kun å bedre de(n) dårligst stiltes situasjon som gjelder. Å gi antibiotikaen til en med «kun» lungebetennelse, ville vært å øke forskjellene mellom hen og den kreftsyke pasienten, uten at det bidro til sistnevntes beste. Dette tillater ikke maximin-prinsippet, og medisinen må derfor allokeres til hen

[37]Her er en naturlig innvending at også slaveeierne kan bli lykkelige av maktfølelsen en slave kan gi. Det virker uansett usannsynlig at dette veier opp for lidelsen slavene opplever.

[38] En annen måte Rawls' teori har en «sikring» mot slaveri på, er frihetene som sikres med hans øvrige rettferdighetsprinsipper. Dette nevner jeg ikke i teksten, da jeg her kun er opptatt av forskjellsprinsippet. All den tid utilitarismen heller ikke innebærer slaveri er dette poenget uansett uviktig.

[39] Igjen forsvinner noen nyanser i min beskrivelse. Harsanyi fikk blant annet svar på tiltale av Rawls, som igjen fikk svar fra Harsanyi. I teksten ovenfor oppsummerer jeg, med et litt modifisert Harsanyi-eksempel, ca. det han poengterte over to eksempler. Originalteksten hans har jeg referert til i teksten.

som har det verst, uavhengig av hvor lite vedkommende får ut av det. Ikke bare det, men den dårligst stilte vil ha krav på all antibiotikaen, så lenge hen får *noe* godt ut av den. Så om, hypotetisk sett, én dose antibiotika gav personen én dag mer å leve, og hen potensielt kunne levd 100 dager ekstra, måtte vedkommende få sykehusets hundre siste doser av antibiotika. Dette på bekostningen av hundre pasienter med lungebetennelse, som heller enn å bli friske fort, fikk lange sykdomsforløp. Dette fremstår som en uholdbar løsning. Dette problemet med maximin-prinsippet møter man i alle deler av politikken der man forsøker å anvende det. Alle ressurser i skolene må allokeres til de med størst lærevansker (gitt at disse personene også kan minst), alle midler innen samferdsel må bevilges områdene med færrest veier, uansett hvor få som bor der (gitt at befolkningstallet er over 0), alle bistandsmidler måtte hjelpe den som har det dårligst, selv om mange flere kunne blitt hjulpet mye mer for de samme ressursene, også videre.[40]

4.2 Lykkelige slaver og enda lykkeligere monstre

Før jeg tar for meg andre teorier av fordelingsrettferdighet, er det verdt å stoppe opp litt ved slaveeksemplet. Rawls postulerer tross alt at slaveri faktisk innebærer velferdsmaksimering i hans eksempel, så det kan fremstå litt for enkelt å avfeie innvendingen med at slikt ikke skjer i virkeligheten. Siden dette og lignende tankeeksperiment er vanlige former for kritikk av utilitarismen, er det relevant å si litt mer om disse, også med tanke på diskusjonen mellom utilitarismen og andre alternative fordelingsprinsipper.

For at slaveri faktisk skal være et utilitaristisk gode, må en kombinasjon av to ting være sant.

1, Slaver er mye nærmere frie mennesker i lykkenivå enn antatt.

og

2, Slaveeiere får mye mer lykke ut av å eie slave(r) enn antatt.

Jeg tar for meg de to mulighetene i nevnte rekkefølge.

[40] Som nevnt i innledningen av kapitlet, er det ikke gitt at Rawls hadde foreslått alle disse tiltakene. Han aksepterte selv at maximin-kriteriet sjelden er riktig, og primært passende for basis-institusjoner. Tendensen over er likevel slående: Maximin-kriteriet gir enten de samme svarene som utilitarismen, eller dårligere løsninger. Det virker derfor irrasjonelt å velge det fremfor utilitarismen.

4.2.1 Lykkelige slaver[41]

At slaver skal være minimum tilnærmet så lykkelige som fri mennesker, betyr at verdien av frihet er adskillig mindre for et enkeltmenneskets liv enn antatt, eller at det, tross stor verdi, bringer med seg mye negativt. For at det skal være tilfellet, må enten slavene være betydelig lykkeligere enn antatt, eller så må frie mennesker være mye mindre lykkelige (alternativt en kombinasjon av disse to grunnene).

I sin selvbiografi skriver den afroamerikanske menneskerettighetsforkjemperen Booker Taliaferro Washington om overgangen til status som fritt menneske (Washington, 1986, s. 24–27). Washington ble født som slave i USA, og var fortsatt et barn da president Abraham Lincoln kunngjorde Emansipasjonserklæringen i 1863. Da nyheten ble kjent blant (de tidligere) slavene, brøt de umiddelbart ut i glede. Endelig var friheten de hadde drømt og sunget om i årevis blitt en realitet. Etter at ekstasen hadde lagt seg, forteller imidlertid forfatteren om en rådvillhet og bekymring i befolkningen, spesielt blant de eldste. De tidligere slavene hadde riktignok ingen eier lenger, men manglet også kunnskaper og ferdigheter til å klare seg som frie mennesker. Å plutselig ha alt ansvar for seg selv og sine nærmeste ble fort til en byrde for mange. Som Washington skriver: «I løpet av få timer var de store spørsmålene som den angelsaksiske rase hadde kjempet med i århundrer blitt kastet over disse menneskene for å løses» (Washington, 1986, p. 27, min oversettelse). Noen av slavene endte opp med å gå tilbake til sine tidligere eiere, for å jobbe med omtrent de sanne betingelsene som tidligere.

I Washingtons fortelling hadde de tidligere slavene lite eller ingenting å gå til da de ble selveiende, og noen av de eldste fikk det kanskje til og med dårligere av frigjøringen. Likevel var slaveriet i USA langt fra noe positivt fra et utilitaristisk ståsted, noe som også gjenspeiles i hvor mye medlidenhet Washington uttrykker overfor nasjoner som har vært gjennom slaveri i sitt land (Washington, 1986, s. 23). For det første, så ble emansipasjonen til de fleste afroamerikanernes beste allerede da, og generasjonene senere nyter godt av å ikke lenger måtte bli født som slaver.

[41] Det kan fremstå noe merkelig at jeg vier såpass mye plass til en diskusjon angående slaveri, da jeg tidligere har sett diskusjoner angående frihetens verdi til side, og i tillegg er opptatt av ressursfordeling, ikke institusjoner, som heller forbindes med prosedural rettferdighet. Slaveri kan imidlertid også anses som en svært uheldig fordeling av ressurser for slaven. Frihet antas i tillegg å ha en stor instrumentell verdi for velferdistiske teorier. Innvendingen mot utilitarismen er derfor relevant, og nødvendig å besvare for å kunne opprettholde bred reflektert likevekt.

For det andre, så var frigjøringen kun et onde for noen tidligere slaver på grunn av skadene slaveriet allerede hadde gjort på dem. Grunnen til at den nyervervede friheten ble en byrde er at de gjennom hele livet hadde levd som slaver, og ikke ble gitt noen kunnskaper eller egenskaper å jobbe med som frie. Slaveriet gjorde ikke livet deres lykkelig, og hadde de vokst opp uten slaveri ville de ikke vært ulykkelige. For noen få fristet det riktignok mer å oppsøke sin tidligere eier igjen, men majoriteten nøt godt av frigjøringen. Dette «redder» ikke slaveri som et utilitaristisk gode.

Dette eksemplet tydeliggjør forholdet mellom velferdisme og muligheter-tilnærmingen, diskutert i del 3.2. De slavene som endte opp med å gå tilbake til sine tidligere eiere, hadde fått preferansene sine tilpasset sin dårlige levestandard over lang tid, på en måte som minner om kvinnene i ørkenområdet utenfor Mahabubnagar i Nussbaums eksempel. Følgelig medførte ikke lenger statusen som fritt menneske en betydelig velferdsøkning, slik det ville gjort om deres preferanser var som tidligere. At velferd er mer sannsynlig til å ha en iboende verdi enn muligheter, tydeliggjøres av følgende: Om det kun var mulig å frigjøre noen av slavene, men ikke alle, virker det åpenbart at man latt de med tilpassede preferanser forbli slaver, og frigjort resten. Det til tross for at de med tilpassede preferanser ville fått like mange nye muligheter med å bli frigjort som de andre slavene.[42] Man tar altså denne avgjørelsen basert på hvor stor velferdsøkning alternativene fører til. Samtidig sammenfaller velferdismen og muligheter-tilnærmingen i praksis, da frigjøring av slaver generelt antas å føre til økt velferd. Med dette sidesporet ute av veien, fortsetter jeg med et eksempel der slaveri *faktisk* er et utilitaristisk gode.

For å finne en situasjon der friheten har liten nok verdi, trenger man å se for seg et scenario der det å være slave innebærer flere goder enn dets alternativ. For å bruke Isaiah Berlins terminologi, må den negative friheten man vinner med å være

[42] En potensiell forskjell mellom eksemplene er at det ikke nødvendigvis er hvordan de tidligere slavene opplevde det å være fri som endret seg, men hva det å være fri medførte. Det er mulig at slavene ville fått en betydelig velferdsøkning av å faktisk klare seg som fri mennesker, men at deres manglende erfaring og utdanning gjorde dem uegnede til et slikt liv. Dette endrer imidlertid ikke poenget. Om det var slavenes preferanser som endret seg, gir velferdismen, som beskrevet over, en bedre vurdering av situasjonen enn muligheter-tilnærmingen. Om det imidlertid ikke er slavenes evne til å oppleve velferd som endret seg, men deres egnethet til et liv uten eier, sammenfaller muligheter-tilnærmingen og velferdismen. I dette tilfellet medfører nemlig ikke det å bli frigjort en økning i antall reelle friheter, slik muligheter-tilnærmingen baserer seg på.

selveiende ikke veie opp for den positive friheten man taper, for eksempel med å ikke lenger engang få det minimale av ernæring slaver får (Berlin, 2002, s. 166–187). Dette er vanskelig å forestille seg. Richard M. Hare har likevel prøvd, i form av et tankeeksperiment (Hare, 1979, s. 111–113). Kort fortalt forestiller Hare seg at Frankrike vant slaget ved Waterloo, men at Napoleon døde. Det påfølgende kaoset i Europa fører til at to hypotetiske øyer i Karibia, Juba og Kamaika, plutselig blir uavhengige fra tidligere imperialister. De to øyene er tilnærmet like, men opptrer forskjellig i dagene etter frigjøringen. I Juba tar lederen blant det som til da var imperialistenes slaver over som statsoverhodet, og beholder slaveri som institusjon, dog med noen modifikasjoner. Slavene får svært fordelaktige kår sammenlignet med tidligere, både i form av materielle goder og noen grunnleggende rettigheter, for eksempel beskyttelse mot grov mishandling. De fortsetter likevel å være slaver. Staten, med den nye lederen på topp, kan bestemme hvor og hva de skal jobbe, og kan straffe dem om de nekter – riktignok kun med fengsel, ikke vold eller lignende, som eiere kunne med tidligere slaver. Det nye statsoverhodet er svært begavet, og har i tillegg full kontroll over befolkningen. Etter få år blomstrer økonomien på Juba.

Kamaikanerne gjorde ikke samme maktovertakelse. Første året etter imperialistenes forsvinnelse råder det fulle kaos, med borgerkrig og ødeleggelse. Årene etter utvikles riktignok et demokrati på øyen. De frislupne slavene har imidlertid ført til en markant befolkningsvekst, og staten er generelt svak i møte med store utfordringer. Innen kort tid er ressursene alt for knappe, og en dyp fattigdom rår over hele øyen. Frie mennesker begynner å flykte over havet til Juba fra Kamaika, for å heller kunne være slaver der – men møtes av strenge grensevakter (slaver, sendt til grensen av staten) i det de går i land. Menneskene på Kamaika fortsetter å lide, mens de jubanske slavene, som med Rousseaus uttrykk er «tvunget til å være fri», nyter tilværelsen (Rousseau, 2011b, s. 167).

Selv i Hares tankeeksperiment er slaveri ikke et ubetinget gode. Situasjonen på begge de fiktive øyene kunne vært unngått, om ikke imperialistene hadde innført slaveri i første omgang. Likevel, gitt at skaden er gjort må utilitarister innrømme at slaveri var positivt i akkurat dette tilfellet. Hvorvidt man trenger å støtte at det skjedde kommer også an på hvilke konsekvenser det har over lenger tid. For eksempel kan det å tillate slaveri en gang svekke menneskers generelle aversjon mot dette, og føre til at flere land benytter seg av slaver - noe som i så fall kan gjøre

at slaveriet på Juba likevel fører til mindre lykke enn lidelse.[43] Uansett, på den skisserte måten er det mulig å postulere scenario der utilitaristen må akseptere slaveri som et gode, noe som umiddelbart virker ille. Samtidig trenger ikke dette være en stor kamel å svelge for utilitarister. Hares tankeeksperiment er svært spesielt, der både det å være slave og det å være fri innebærer noe helt annet enn det man forventer. Man reagerer riktignok negativt i det man hører om slaveri, men i dette tilfellet er det lett å se at intuisjonene tar feil. Refleksene skyldes at man er vant til at henholdsvis slaveri eller frihet innebærer noe helt annet enn i eksemplet. Man bør derfor i dette tilfellet ignorere at det intuitivt virker feil, og akseptere slaveri i det skisserte tankeeksperimentet.

4.2.2 Enda lykkeligere monstre

På dette punktet vil kritikere av utilitarismen kunne fremme enda en innvending. I eksemplet over endrer Hare innholdet av begrepet «slave» såpass mye, at det nesten er en feilkategorisering å fortsatt kalle beboerne på Juba for dette. Med de store endringene som er gjort, kunne man like godt kalt den nye statusen noe annet enn slaveri – og dermed ikke kunnet avfeie kritikken utilitarismen møter. For å illustrere den egentlige innvendingen, kan man se for seg at slaveeiere faktisk får mer glede ut av ordningen enn det slavene taper på det. Som allerede fastslått må disse eierne i så fall oppleve adskillig mer lykke enn man normalt vil anta.

Dette er ikke en uvanlig måte å gjøre innvendinger mot utilitarismen på. Michael Sandel ber leseren se for seg at publikums lykke overgikk de kristnes lidelse, da sistnevnte gruppe ble kastet til løvene i Romerriket (Sandel, 2009b, s. 24–25). Et lignende tankeeksperiment gjøres av Robert Nozick, som også kritiserte utilitarismen for dets uakseptable implikasjoner. I eksemplet beskriver Nozick et «lykkemonster», som får veldig mye mer lykke ut av ressurser enn det vanlige mennesker gjør (Nozick, 1974, s. 14).[44] Utilitaristen må i dette tilfellet allokere alle verdens ressurser til dette monsteret, og avslutningsvis la det spise alle mennesker i tillegg. Utilitarismen støtter altså i de nevnte eksemplene både slaveri, det å ofre kristne for

[43] Av lignende grunner skal det enda mer til før en regelutilitarist, som for eksempel Hare, tillater slaveri, enn det gjør fra et handlingsutilitaristisk perspektiv. I denne omgang er derimot ikke dette skillet viktig.
[44] Nozick beskriver ikke eksemplet i veldig grundig detalj, og det er et tolkningsspørsmål om han faktisk snakker om monster, eller egentlig beskriver spesielt «lykkesensitive» mennesker her. For tankeeksperimentets implikasjoner utgjør dette imidlertid ingen forskjell.

gladiatorglade romeres glede og å la et lykkemonster spise (og dermed drepe) alle verdens mennesker. Alt dette virker intuitivt å være feil svar gitt av utilitarismen på de moralske dilemmaene.

4.2.3 Dårlig data

Selv om disse eksemplene ser annerledes ut «utenfra» enn Hares tankeeksperiment, som var i overkant liberal i bruken av begreper, kan de innvendingene besvares på lignende måte. Derek Parfits svar til Nozicks monster illustrerer dette godt (Parfit, 1984, s. 389). Et slikt monster er rett og slett umulig i den virkelige verden, og er ikke engang oppnåelig for mennesket å se for seg. For å kunne frata alle verdens mennesker sine ressurser i tillegg til å spise dem, men likevel bidra til velferdsmaksimering heller enn det motsatte, måtte monsteret kunne oppleve lykke som er flere milliarder ganger større enn det mennesker opplever.[45] Dette er rett og slett ikke noe man har mulighet til å oppfatte med en menneskehjerne. Å bli fratatt ressurser, og attpåtil å bli spist opp, er selvsagt forferdelig, men samtidig innebærer tankeeksperimentet lykke av dimensjoner som for mennesker er umulig å forestille. Forholdet mellom mennesket og lykkemonstret kan sammenlignes med det mellom innsekter og mennesket. Selv om innsektene trolig ville protestert, er de fleste mennesker enige om at man kan drepe milliarder av innsekter. Det er derfor ikke rart om menneskelige intuisjoner ikke er til å stole på i dette tilfellet.

Lignende logikker kan anvendes for å forstå de resterende eksemplene. Mennesket har i slike situasjoner det Joshua Greene kaller et problem med for dårlig data (Greene, 2017, s. 75). Man har rett og slett ingen erfaring med situasjoner der det å ofre mennesker til løvene, eller å innføre/beholde slaveri som institusjon fører med seg flere goder enn onder. I de tre situasjonene beskrevet fremprovoserer disse illustrasjonene heller intuisjoner knyttet til konvensjonelle situasjoner i oss, der slaveri fører til lidelse, og det ikke er verdt å ofre mennesker, hverken til løver eller monst

4.2.4 En analogi til den empiriske statsvitenskapen

[45] Parfit skriver «flere millioner», men med dagens befolkningstall på kloden kan man ta enda hardere i.

For å si det i statsvitenskapelige termer, har de nevnte tankeeksperimentene dårlig målevaliditet.[46] Målevaliditet gir uttrykk for hvor godt de operasjonaliserte variablene (i indikatorform) måler de teoretiske begrepene de er ment å måle (Adcock & Collier, 2001, s. 529). Målet med eksperimentene er å finne ut hva som er moralsk riktig i de skisserte casene, for eksempel i scenarioet der slaveeiere opplever mer lykke enn deres slaver opplever lidelse. Det moralsk riktige operasjonaliseres som våre intuisjoner når vi blir eksponert for tankeeksperimentet, intuisjonene tilegnes med andre ord rollen som indikator på det teoretisk interessante (men umålbare) begrepet (Adcock & Collier, 2001, s. 531). Problemet er at intuisjonen tankeeksperimentet fremprovoserte stammer fra den klassiske situasjonen, der slavens tap er betydelig verre enn eierens vinning. Indikatoren gir altså systematisk uttrykk for noe annet enn det forskeren ønsker å dra slutninger om – i dette tilfellet vil intuisjonene stadig være «for sensitive». Man kan derfor ikke bruke kombinasjonen av denne metoden og dette datasettet til å besvare det aktuelle spørsmålet.

Dette resonnementet er ikke bare relevant for spesielle tankeeksperiment, men har overføringsverdi til mer livsnære hendelser. Utilitarisme godkjenner ikke ting som slaveri, rasisme, seksisme eller lignende forskjellsbehandling, rett og slett fordi disse aldri er måter å maksimere lykke på i den virkelige verden. I hypotetiske eksempler kan man riktignok skape situasjoner der utilitarismen går imot det som fremstår intuitivt riktig. Men om man i slike eksempler endrer på de vanlige empiriske forholdene, har man ikke lenger grunn til å stole på sine intuisjoner som stammer fra den vanlige verden. Gjør man det, baserer man sine moralske dommer på informasjon som er irrelevant for situasjonen man feller en dom over. At utilitarismen i kontraintuitive scenario fremstår kontraintuitiv er ikke en svakhet, men en styrke.

Jeg har nå vurdert Rawls rettferdighetsteori. Maximin-kriteriet gir ikke bedre svar enn utilitarismen i hverdagens eller politikkens moralske spørsmål, og bør heller ikke velges i en hypotetisk utgangsposisjon bak uvitenhetens slør. I tillegg til å ta for meg

[46] I filosofi-sjargong er det vanligere å kalle intuisjoner for «lite reliable» i slike situasjoner, i betydningen «ikke til å stole på».
Anvender man statsvitenskapelige uttrykk derimot, er validitet et bedre treffende begrep. Som jeg beskriver i teksten, fører våre intuisjoner i et slikt scenario til systematiske feil. I statsvitenskap brukes reliabilitet heller om presisjonen i målinger, der lav reliabilitet betyr mye tilfeldige feil (Trochim et al., 2016, Kapittel 5.2a).

Rawls' prinsipp har jeg forsvart utilitarismen mot noen sentrale innvendinger. Jeg går videre til neste teori om fordelingsrettferdighet.

4.3 Egalitarianisme

Egalitarianisme, altså teorier om likhet, er trolig den vanligste tilnærmingen til spørsmål om rettferdighet. Det virker åpenbart at rettferdighet krever likhet på en eller annen måte. Det er imidlertid vanskeligere å avgjøre akkurat hvilken rolle likheten bør spille. Egalitære teorier kan for eksempel gå ut på at mennesker får samme tilgang på ressurser, behandles likt eller har samme sosiale status (Arneson, 2013, Kapittel 0.0). Av grunnene beskrevet i del 2.1.2, er en type egalitarianisme spesielt relevant for min diskusjon: Utfallsorienterte (altså teliske) likhetsprinsipp med fokus på velferd.

Telisk egalitarianisme kjennetegnes av følgende holdning:

> *Prinsippet om likhet:* Det er negativt i seg selv om noen er dårligere stilt enn andre (Parfit, 1991, s. 4).

De fleste teliske egalitære er pluralister, og godtar også dette:

> *Prinsippet om nytte*: Det er positivt i seg selv om mennesker har det bedre.

4.3.1 Innvendingen om utjevning nedover

Prinsippet om nytte er i praksis en formulering av utilitarisme. Det er altså *Prinsippet om likhet* som er særegent for egalitarisme. Derek Parfit beskriver flere utfordringer med dette prinsippet, som får ham til å avvise egalitarianisme til fordel for prioritarianisme. Det er spesielt et poeng som har vist seg å være svært kraftig. Ifølge *Prinsippet om nytte* er mer likhet i samfunnet alltid en positiv utvikling. Følgelig er det også et gode om de som har det bedre fra før får det verre, selv dersom det ikke gjør at de dårligst stilte får det bedre. For å illustrere det numerisk, viser følgende tabell to mulige utfall av velferdsnivåene til en søskenflokk i Andeby:

Tabell 1

	Utfall A			Utfall B		
Navn	Ole	Dole	Doffen	Ole	Dole	Doffen
Velferdsnivå	20	20	10	10	10	10

Ifølge *Prinsippet om likhet* ville det vært positivt om beboerne i Andeby endte på utfall B, heller enn utfall A. Det til tross for at ingen av Onkel Donalds nevøer har det bedre i Utfall B enn i utfall A. Tvert imot, to av de har det verre.

Som nevnt er tilhengere av likhetsteori som regel pluralister, og trenger derfor ikke påstå at den skisserte endringen i Andeby totalt sett er til det bedre. Mange egalitære vil nok påstå at Ole og Doles tap av velferd utgjør et større onde enn det den nyervervede likheten utgjør et gode. Likevel, de må godta at utfall B i hvert fall i noen henseende er bedre enn A, selv om ingen har det bedre i B enn i A. I praksis vil det bety at å ta fra, eller på andre måter ødelegge for dem som har det bedre, uten at det på noen måte fremmer de dårligere stiltes velferd, i hvert fall delvis kan sies å være et gode. Dette virker direkte feil. Fenomenet, som blir kalt *Innvendingen om utjevning nedover*, har skapt mye hodebry for tilhengere av telisk egalitarianisme. Noen har endt opp med å heller støtte lignende prinsipper, som unngår dette problemet. Andre har tatt likhetsteori i forsvar mot *Innvendingen om utjevning nedover*. Jeg skal undersøke to av disse argumentene, henholdsvis fremmet av Larry Temkin og John Broome.

4.3.2 Temkins forsvar av egalitarianisme

Larry Temkin anerkjenner at hans foretrukne teori treffes av innvendingen, men forsøker å tilbakevise den. Som første steg ser han på bakgrunnen til motargumentet. Som Temkin påpeker, må man godta følgende regel for at *Innvendingen om utjevning nedover* skal være et avgjørende poeng mot egalitarismen:

Slagordet: For at en situasjon skal være verre (eller bedre) enn en annen, må det finnes *noen* den er verre (eller bedre) for (Temkin, 2000, s. 132, min oversettelse). [47]

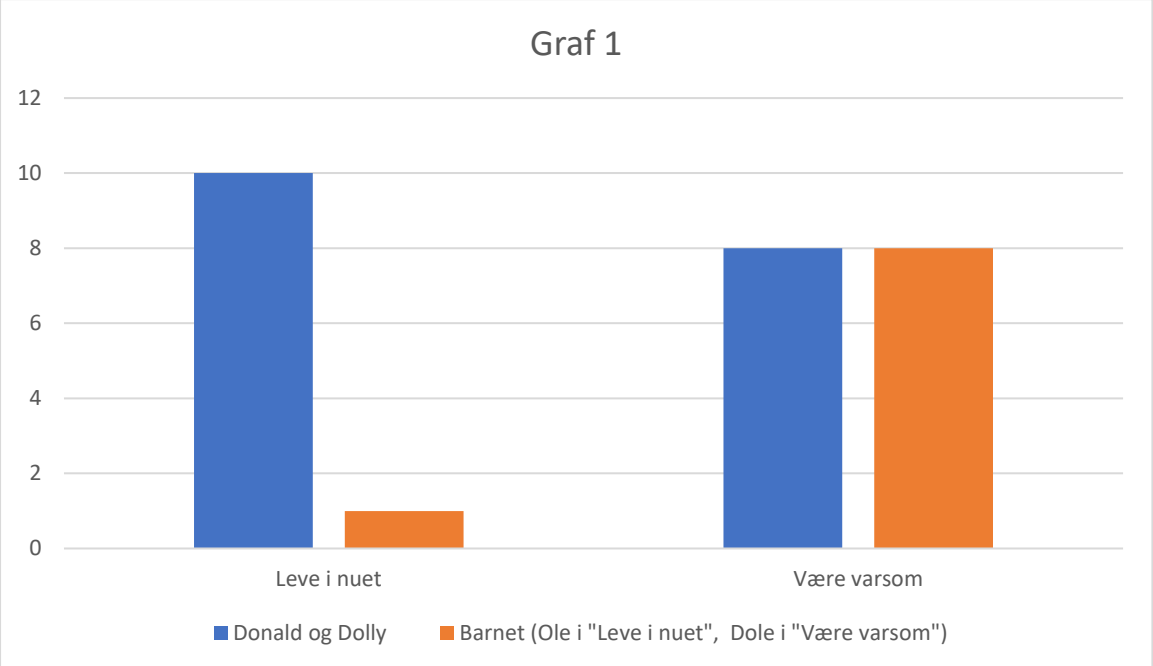
Når dette er etablert, argumenterer Temkin for at *Slagordet* ikke er så selvsagt som mange virker å påstå. Er *Slagordet* svakt, er heller ikke *Innvendingen om utjevning nedover* så alvorlig, og telisk egalitarianisme kan igjen etableres som et plausibelt alternativ om fordeling av goder. Temkin lister opp en rekke punkter for å vise at *Slagordet* er inkompatibelt med andre anerkjente teorier. I en sammenligning med

[47] Parfits «*Prinsipp om påvirkning på mennesker»* oppsummerer samme poeng på en mer presis måte: Hvis et resultat ikke er dårligere for noen, kan det ikke være dårligere (Parfit, 1991, s. 32). Siden Temkin viser til «*Slagordet*» i sin argumentasjon, fortsetter også jeg å referere til det.

utilitarisme, er det spesielt to argumenter som er interessante. Det første er et tankeeksperiment som angår Parfits *Manglende-Identitet Problem*, mens det andre er en påstand om at *Slagordet* ikke er forenlig med hedonisme.[48] Jeg tar de for meg etter tur, og begynner med førstnevnte.

Slagordet og mangelen på identitet[49]

I eksempelet diskuterer et godt stilt par, Donald og Dolly, hvilken tilnærming de skal velge for fremtiden. Donald har lyst til å *leve i nuet*. Det betyr å få barn med en gang, reise mye som familie og kose seg. Dette vil føre til at Donald og Dolly får det enda litt bedre enn de har det fra før. Samtidig vil de bruke opp både familiens og jordens ressurser med all reisingen, og barnet (Ole) får det mye dårligere enn sine foreldre. Dolly ønsker heller å *være varsom*. Hun ønsker å spare penger noen år før de stifter familie, for å så leve et roligere liv enn i Donalds forslag. Dette scenarioet innebærer at Donald og Dolly ofrer litt av sin velferd, i stedet for å *Leve i nuet*. Barnet deres, i dette tilfellet Doffen, får det like bra som Donald og Dolly. Valget illustreres i *Graf 1*.



[48] Temkin beskriver argumenter både mot hedonisme og teorier om preferansetilfredsstillelse. For å opprettholde utilitarismens plausibilitet er det tilstrekkelig å avvise det ene. Som beskrevet i del 3.2, er mitt valg av hedonisme fremfor preferansetilfredsstillelse er vilkårlig, jeg holder meg agnostisk angående hvilken teori om velferd som er riktig.

[49] Kapittel 16 av Reasons and Persons omhandler *Manglende-Identitet Problemet* (Parfit, 1984, Kapittel 16). Tankene om at ufødte mennesker ikke kan ha det bra/dårlig, og at foreldres utsettelse av å få barn fører til at et annet menneske blir født enn om de ble foreldre med en gang, henter Temkin fra Parfit.

De fremstår åpenbart at *Leve i nuet* er et dårligere scenario enn *Vær varsom*. Men, som Temkin påpeker, er det ingen som får det dårligere fra førstnevnte til sistnevnte utfall. Donald og Dolly har det til og med bedre i *Leve i nuet* enn i *Vær varsom*.

Siden det er flere år mellom tidspunktene for når de velger å få barn i de to scenarioene, er det for eksempel forskjellige sæd- og eggceller som vil resultere i unnfangelse. Det er altså forskjellige barn i de to scenarioene, Ole i *Leve i nuet*, og Doffen i *Vær varsom*. Ingen av de to ville noen gang eksistert i den andre verdenen, så ingen ville kunne sies å ha det dårligere enn alternativet. At *Vær varsom* kan være dårligere enn *Leve i nuet*, uten at det er dårligere for *noen*, virker å svekke *Slagordet*. Og siden Slagordet i sin tur er motivasjonen bak *Innvendingen om utjevning nedover*, som er den viktigste kritikken mot telisk egalitarianisme, virker Temkins eksempel å styrke sistnevnte prinsipp.

Jeg mener imidlertid at eksemplet ikke svekker *Slagordet*. Temkins tankeeksperiment, og Parfits *Manglende-Identitet Problem* mer generelt, viser til en svært spesiell situasjon. Ole og Doffen vokser riktignok opp til å bli forskjellige personer. Når Donald og Dolly skal ta avgjørelsen, eksisterer imidlertid ingen av de to. Det er derfor ingen grunn til å gjøre et moralsk relevant skille mellom dem. Fra foreldrenes perspektiv er det kun «barnet» de kan forholde seg til i sin planlegging for fremtiden. Siden det ikke finnes en forskjell på de to i valgets øyeblikk, gir det mening å omtale dem som en og samme person. «Barnet» fikk det dårligere i *Leve i nuet* enn i *Vær varsom*. Temkins eksempel svekker derfor ikke *Slagordet*, og at den setter pris på utjevning nedover fremstår fortsatt som en sterk innvending mot telisk egalitarianisme.

Slagordet og hedonisme[50]

I det andre argumentet drar Temkin opp et skille mellom *teorier om egeninteresse* og *teorier om gode utfall* (Temkin, 2000, s. 141). Førstnevnte er ment å illustrere hva som er godt for noen, mens sistnevnte sier noe om hva som gjør en situasjon god. Temkin godtar at kun deres egne opplevelser kan være gode for individer, og anerkjenner med det hedonisme som en plausibel *teori om egeninteresse*. Han

[50] I Temkins argument diskuterer han *mentale tilstand-teorier* mer generelt. Men siden hedonisme er den vanligste versjonen av slike teorier, og den mest relevante for utilitarisme, forholder jeg meg kun til den. Som nevnt tidligere, er valget om jeg forsvarer utilitarismen mot Temkins innvending mot hedonisme- eller preferansetilfredsstillelse vilkårlig.

avviser derimot at hedonisme også kan være riktig *teori om gode utfall*. Med å vise til blant annet frihet, dyd og rettferdighet, beskriver Temkin det han mener er en bred enighet blant filosofer: At det finnes ting utenom opplevelser som er av iboende moralsk verdi. Om man først aksepterer at noe er av moralsk verdi, utenom dets påvirkning på noen, så må *Slagordet* forkastes (Temkin, 2000, s. 142). Når den barrieren er brutt er veien kort til å også godta likhet som iboende verdifullt.

Temkin har rett i at mange tilegner frihet, dyd, rettferdighet og lignende iboende verdi, uavhengig av hvordan det påvirker forskjellige personer. Likevel overbeviser han neppe utilitarister (eller prioritarianister, som Temkin i all hovedsak argumenterer mot).[51] Også utilitarister godtar at konseptene Temkin har en instrumentell verdi, uten at det behøver å bety at det er noe iboende godt i dem. At mennesker er fri, lever dydig, opplever samfunnet sitt som rettferdig og så videre, er alle forhold som ofte fører til maksimering av lykkenivåer. Men det iboende verdifulle finner de likevel kun innenfor hedonismens rammer – frihet er for eksempel et gode fordi mennesker blir lykkelige av å være fri.

Likhet kan også ha en slik instrumentell rolle i hedonistisk utilitarisme, men ikke på en måte som svekker *Slagordet*, eller gjør *Innvendingen om utjevning nedover* mindre alvorlig. Ulikheter i ressurser fører ofte med seg tap av velferd blant borgere i et samfunn. Jean-Jacques Rousseau beskriver dette i *Discourse on the Origin of Inequality*. Der peker han på *amour-propre*, altså menneskes trang til å bli anerkjent av hverandre, som roten til menneskelige onder (Rousseau, 2011a, s. 64, 106; Neuhouser, 2008, s. 31). Ulikheter i nivået av ressurser fører til dårlig selvrespekt blant den fattige delen av befolkningen. De blir i tillegg utnyttet («dominert», med Rousseaus ord) av de rike, som lever under et konstant press grunnet konkurransen seg imellom om å være den med flest ressurser.[52] At mennesker har tilnærmet like ressurser og goder antas derfor å ha gode konsekvenser for det totale lykkenivået, siden det gjør at fattige ikke er sjalu på andres rikdom, og at de velstående ikke kan

[51] Det finnes flere veier til prioritarianismen enn den konsekvensialistiske. Men teorien er nært beslektet med utilitarisme, så fra Temkins standpunkt vil argumenter mot det ene som regel fungere i en diskusjon med det andre også. Mer om prioritarianismens forhold til utilitarisme i del 4.4.2.

[52] Tilsvarende tanker forbindes også gjerne med Karl Marx, som skriver om både menneskers tendens til å vurdere seg selv basert på sammenligning med andre, og hvordan ulikheter fører til utnyttelse (Marx, 1976, s. 32–33; Cohen, 2001, s. 322).
Jeg påstår her hverken at Rousseau eller Marx var utilitarister eller hedonister, kun at deres argument forklarer hvorfor likhet i ressurser kan være et utilitaristisk gode.

(eller trenger å) utnytte fattigere mennesker. Å jevne ut de materielle forskjellene, også ved å kun jevne ut nedover, kan derfor være et gode. I slike tilfeller jevnes ikke velferdsnivået ut *nedover,* i motsetning til ressursene. Når *Slagordet* eller *Innvendingen om utjevning nedover* diskuteres, omhandler dette velferdsnivåer. Mangelen på utnytting eller sjalusi er allerede kalkulert inn i folks velferdsnivåer, og en slik form for likhet har derfor hverken iboende eller instrumentell verdi.

Tar man et steg tilbake, kan man også sette spørsmålstegn ved hva Temkins poeng egentlig er i forbindelse med hedonismen. Det er noe sirkulært ved å postulere en forskjell mellom *teorier om egeninteresse* og *teorier om gode utfall*, i et forsøk på å tilbakevise *Slagordet*, som sier at de to er det samme.[53] Jeg konkluderer med at Temkin ikke lykkes i å styrke egalitarianismens posisjon sammenlignet med utilitarisme (eller prioritarianisme).

4.3.3 Broomes forsvar av egalitarianisme

John Broome forsvarer også egalitarisme i møte med *Innvendingen om utjevning nedover* (Broome, 2002, s. 2–3).[54] I motsetning til Temkin gjør han det ikke med å betvile innvendingens viktighet, eller med å avvise *Slagordet*.[55] Han argumenterer i stedet for at telisk egalitarianisme ikke jevner ut nedover, og at innvendingen derfor ikke treffer teorien. Poenget kan illustreres som følger:[56]

Ifølge teliske egalitarianisme skal et utfall vurderes ut fra hvor mye velferd de forskjellige aktørene har, og hvor lite ulikhet det er mellom dem. Formelen for å vurdere et gitt utfall ser slik ut:

Formel 1

$$G = V - \sigma U ,$$

[53] Eventuelt er *teorier om gode utfall* en funksjon av *teorier om egeninteresse*. Temkin anerkjenner riktignok at utilitarister har dette synet på de to teoriene, men fremlegger ikke flere enn de nevnte argumenter for at de tar feil.

[54] En nyttig oversikt over Broomes argument finnes i *Egalitarianism* (Hirose, 2015, s. 74–76). Kapittel 3.4 av samme bok presenterer også en teori som er nært beslektet Broomes prinsipp.

[55] Han argumenterer selv for en svært lignende regel, som han kaller *Prinsippet om personlige goder* (Broome, 1991, s. 165). Broomes regel skiller seg kun fra *Slagordet* på samme måte som hans teori om menneskelig velferd skiller seg fra tradisjonelle syn på dette. Sistnevnte distinksjon diskuteres i teksten.

[56] Ligningene presenteres i Broomes artikkel og Hiroses bok (Broome, 2002, s. 2; Hirose, 2015, s. 66, 74). Jeg har gjort noen små justeringer, samt lagt til et steg, for å gjøre matten enklere å forstå. Meningsinnholdet er uendret.

der G står for utfallets «godhet», V står for aktørenes velferdsnivåer, U står for ulikheten mellom aktørene, og σ er vekten den aktuelle versjonen av egalitarianisme gir ulikhet.

Anvender man Broomes versjon av prinsippet på et scenario («scenario A») med to personer, får man følgende ligning:

Formel 2

$$\mathrm{G}_A \;=\; \mathrm{V} - \frac{1}{2}\,\mathrm{U}$$

Fyller man inn de to personene i ligningen (person 1 og person 2), så skrives det slik:

Formel 3

$$G_A = (V_1 + V_2) - \frac{1}{2}|V_1 - V_2|$$

Broome vurderer altså godheten av et scenario med to personer som summen av deres velferd, minus halvparten av ulikhet mellom de to personenes velferdsnivå. Formel 1, 2 og 3 tydeliggjør *Innvendingen om utjevning nedover*. Siden mengden ulikhet skal trekkes fra det totale nivået av godhet, så er utjevning nedover eksplisitt *godt på en måte* i disse tilfellene, selv om det med Broomes vekting av ulikhet aldri vil gjøre totalen høyere. Men, som Broome viser, kan Formel 3 skrives om på følgende måte.

Formel 4

Hvis person 2 har like mye eller mer velferd enn person 1, så:

$$G_A = \frac{1}{2}V_1 \;+\frac{3}{2}\,V_2$$

Hvis person 1 har like mye eller mer velferd enn person 2, så:

$$G_A = \frac{2}{3}V_1 \;+\frac{1}{2}\,V_2$$

Formel 4 gir akkurat samme resultat som de forrige variantene. Her er imidlertid ulikhet aldri nevnt. I denne formelen er det ikke noe som tyder på at utjevning nedover i det hele tatt er et gode. Tanken bak dette er at ulikheter i velferd ikke er negativt som sådan, uavhengig av personer, men at det er negativt for dem som har mindre velferd. Slik mener Broome at telisk egalitarianisme unngår å bli truffet av *Innvendingen om utjevning nedover*.

Mitt svar til Broome ligner noe på det jeg skrev angående Temkins argument.[57] Om ulikheter er negativ for en person, ville det gjenspeilet seg i vedkommendes velferdsnivå. Altså trenger den ikke å tas høyde for utover det som allerede er synlig i fordelingen av velferd. Den nye ligningen til Broome går riktignok opp matematisk, men den underliggende tankegangen er feil.

For å gjøre poenget klarere, kan man se for seg at en teatersjef ønsker å ansette et brødrepar på jobben. Hun er svært opptatt av at arbeidstakerne har lang livserfaring, så høy alder spiller en stor rolle i rangeringen av søkere. I tillegg vil hun at de to nyansatte skal kunne forstå de samme kulturelle referansene, og mener derfor det er viktig at det er liten aldersforskjell mellom de to. Teatersjefen ender opp med å vekte likt som Broome. Når brødrepar vurderes, begynner hun med deres adderte alder, og trekker fra halvparten av parets aldersforskjell. Beregningen er derfor lik Broomes formel:

Formel 5

$$B_D = (A_e + A_y) - \frac{1}{2}|A_e - A_y|$$

Der B_D er totalverdien til brødreparet D, A står for alder, mens e og y står henholdsvis for den eldre og den yngre broren. Også dette kan skrives om slik:

Formel 6

$$B_D = \frac{1}{2}A_e + \frac{3}{2}A_y \ .$$[58]

[57] Interessant nok påpeker Temkin ca. samme svakhet i Broomes argument, selv om Temkin og jeg ender opp på motstridende konklusjoner angående temaet (Temkin, 2000, Kapittel XI.). For diskusjonen mellom Temkin og Broome, se også side 182-184 av *Weighing Goods* (Broome, 1991, s. 182–184).

[58] Merk at det i formel 6 er nok å skrive en ligning, i motsetning til formel 4. Dette fordi mennesker ikke kan bytte plass angående hvem som er eldst, i motsetning til hvem som har mest velferd.

Selv om regnestykket nå ikke viser spor av ulikhet som sådan, er personenes alder den samme. Det gir ikke mening å tenke at den yngre personen i brødrepar D plutselig er blitt 50 prosent eldre enn tidligere, eller at hans eldre bror er halvparten så gammel i Formel 6 som han var i Formel 5. Selv om det matematisk går opp, så kan man ikke bare omfordele ulikheten til å være noe iboende hos de to aktørene.

På dette punktet kan forsvarere av Broomes teori skyte inn at han her ikke ville ment at de potensielle ansatte ble yngre eller eldre. Snarere ville Broome ment at for eksempel $\frac{3}{2}$ av den eldre brorens alder er hans *alderspoeng* i jobbsøknaden.[59] Det løser derimot ikke problemet. Alderspoengene vil fortsatt være noe situasjonsspesifikt og eksternt tilegnet, som er avhengig av den andre broren. Om det plutselig var en av den yngre brorens andre søsken som møtte opp på audition med ham, ville denne verdien straks vært en annen. Alderspoeng er med andre ord ikke noe iboende ved broren, på samme måte som menneskers «bidrag» i Broomes ligning (Formel 4) ikke er noe iboende hos disse aktørene. Man kommer ikke unna at dette egentlig er funksjoner av ulikheter mellom personene.

Å «omorganisere» disse ulikhetene til å være en del av mennesket fremstår som en fiffig matematisk vri, men for ad hoc til å redde egalitarianismen fra *Innvendingen om utjevning nedover*. Følgelig virker fordelingsprinsippet lite plausibelt. Jeg gjør derfor som Parfit, og beveger meg videre til prioritarianismen.

4.4 Prioritarianisme

Prioritarianismens vurdering av forskjellige utfall er nesten lik utilitarismens rangeringer. Forskjellen er at prioritarianismen vektlegger personers velferdsnivå i større grad, i takt med hvor lite velferd de har. Med andre ord sier prinsippet at fordeler betyr mer jo dårligere stilt en aktør er (Holtug, 2017, p. 1). Maximin-kriteriet kan sees på som en spesielt radikal versjon av prioritarianismen. Siden jeg har argumentert mot dette i detalj i forbindelse med John Rawls' *A Theory of Justice*, gjentar jeg ikke de samme poengene her. I denne omgang tar jeg først for meg en analogi gjort Parfit, og viser at også utilitarismen kan forklare hans oppfatninger i det tilfellet. Etter dette tydeliggjør jeg forskjellen mellom de to fordelingsprinsippene, med

[59] På samme måte som han ikke mener at velferdsnivået til den dårligere stilte blir høyere av mer likhet, men at hva som er godt for et menneske er en bredere kategori enn velferd. Jeg har allerede argumentert mot dette i del 4.3.2, i tillegg til teksten over. Som nevnt påpeker også Temkin at prinsippet ikke virker å fungere (Temkin, 2000, Kapittel XI.).

hjelp av en annen analogi av samme forfatter. Siden mitt hovedargument for utilitarisme fremfor prioritarianisme er det samme jeg bruker mot tilstrekkelighetsteori, fortsetter jeg så med å presentere det prinsippet. Deretter forklarer jeg hvorfor intuisjoner som er mer kompatible med prioritarianisme eller tilstrekkelighetsteori enn utilitarismen er vanlige, og hvorfor jeg likevel mener at disse intuisjonene bør ignoreres.

4.4.1 Parfits argument

I sin artikkel *Equality or priority* presenterer Parfit det som er blitt kjent som prioritarianisme, som et (ifølge Parfit, bedre) alternativ til telisk egalitarianisme (Parfit, 1991).[60] Parfit argumenterer særlig for prinsippet siden den unngår *Innvendingen om utjevning nedover*. Riktignok sammenfaller de to prinsippene ofte, da begge vektlegger å hjelpe de dårlig stilte. Forskjellen mellom prinsippene er at dette for prioritarianister ikke nødvendiggjøres av personers relative velferdsnivå, men på grunn av det absolutte. Mennesker bør ifølge teorien prioriteres fordi de er *dårlig* stilt, ikke fordi de er *dårligere* stilt enn andre.

For å illustrere poenget, bruker Parfit en analogi (Parfit, 1991, s. 23).[61] En person som er høyt oppe på fjellet har større problemer med å puste enn en som ikke er kommet like langt i sin klatring. Følgelig er det viktigst å hjelpe den som er øverst. Dette er ikke på grunn av hennes forhold til den andre turisten på fjellet. Om klatreren var alene, hadde hun fortsatt hatt akkurat de samme pustevanskene, og hatt like mye bruk for en hjelpende hånd (eventuelt en hjelpende oksygenmaske). Relative nivåer er altså ikke viktige, mens absolutte nivåer er det.

4.4.2 Likheter mellom prioritarianisme og utilitarisme

Parfits eksempel tydeliggjør forskjellen mellom prioritarianisme og egalitarianisme. Likevel er det verdt å merke seg at også en utilitarist gjenkjenner behovet for å hjelpe klatreren. Om klatreren er alene, øker utilitaristen velferdsnivået ved å gi henne oksygen. Om begge klatrerne er på fjellet, og utilitaristen bare kan hjelpe den

[60] Prinsippet omtales oftest som prioritarianisme, og det er denne betegnelsen jeg selv bruker i teksten. Tidligere har for eksempel Larry Temkin kalt det for «utvidet humanitærisme» (Temkin, 1993, s. 245). Derek Parfit, som først beskrev prinsippet i sin nåværende form under sitt Lindley-foredrag i 1991, omtaler det gjerne som prioritets-synet.

[61] Jeg har utvidet analogien noe, for å tydeliggjøre forskjellen på de ulike prinsippene. Det underliggende poenget forblir det samme.

ene, vil hen prioritere den som får mest velferd ut av oksygenet. Siden ondene (som gjenspeiles i klatrerens forventede velferdsnivå) ved å klatre ikke øker lineært, men blir betydelig større i det luften er tynn nok til å være livsfarlig, vil også en utilitarist prioritere klatreren som er høyest på fjellet. Derfor skilles ikke utilitarismen og prioritarianismen i sin vurdering av hva som er riktig i dette tilfellet.

Parfit gjenkjenner at utilitarismen ofte har lignende implikasjoner som hans foretrukne prinsipp, siden de dårligst stilte som regel er enklere å hjelpe til økt velferd enn de som har det bedre. Derfor poengterer han at det ifølge prioritarianismen er viktigere å hjelpe de dårlig stilte, også når det er vanskeligere å hjelpe disse – med andre ord, også når de bedre stilte kunne fått mer velferd ut av hjelpen (Parfit, 1991, s. 19). Grunnen til dette er at den moralske viktigheten av velferd ifølge Parfit er avtagende (Parfit, 1991, s. 24). Dette gjør for eksempel at utilitaristen og prioritarianisten ville vært uenige i følgende dilemma: [62]

Onkel Donald har kun mulighet til å hjelpe en av sine nevøer til mer velferd. Han kan enten hjelpe Ole, som er svært dårlig stilt. Eller så kan han hjelpe Dole, som er veldig godt stilt, til en litt større økning i velferdsnivå enn det han kan hjelpe Ole med.

Tabell 2

	Utfall A		Utfall B	
Navn	Ole	Dole	Ole	Dole
Velferdsnivå	20	80	10	91

For utilitaristen er Utfall B det foretrukne, da dette maksimerer den totale lykken i situasjonen. Akkurat hvor mye ekstra vekt velferdsnivået til dårlig stilte personer får, varierer fra prioritarianist til prioritarianist, men i dette tilfellet ville nok de aller fleste sagt at Utfall A er det bedre, siden Oles vektede velferdsøkning veier opp for det lille tapet i total velferd.

[62] Her illustrerer jeg et eksempel Parfit beskriver i sin artikkel «Another Defence of the Priority View» (Parfit, 2012, s. 401–402). Tabellen og tallverdiene er det jeg som har lagt til, men slik jeg ser det er disse i tråd med Parfits beskrivelse. Parfit bruker andre navn, uten at det endrer noe av betydning.

Jeg mener at prioritarianismen bommer med sitt råd til Onkel Donald i denne situasjonen, og at dets begrunnelse er feil. Etter å ha presentert tilstrekkelighetsprinsippet, beskriver jeg hvorfor i del 4.6.

4.5 Tilstrekkelighetsprinsippet

I korte drag sier tilstrekkelighetsprinsippet at det er viktig at alle har *nok* av et gode (Huseby, 2019, s. 1). Det finnes en rekke forskjellige teorier om hva det relevante godet er, og hvor mye av det som bør regnes som nok – og tilstrekkelighetsteorier skilles i tillegg fra hverandre i andre aspekter også. I de neste sidene tar jeg primært for meg en spesifikk versjon av prinsippet, beskrevet av Robert Huseby i artikkelen *Sufficiency, Restated and Defended*. Dette fordi han fremmer et telisk prinsipp, med fokus på fordelingen av velferd (Huseby, 2010, s. 179).[63] Husebys teori egner seg derfor godt til å kontrastere med utilitarismen.

4.5.1 Husebys tilstrekkelighetsprinsipp

Huseby tar utgangspunkt i følgende definisjon på tilstrekkelighetsprinsippet:

> Det er ille i seg selv om en person ikke har det tilstrekkelig bra. Jo lenger unna terskelen for tilstrekkelighet personen er, desto verre er det (og det er spesielt ille om en persons grunnleggende behov ikke er dekket), og jo flere personer som ikke har det tilstrekkelig bra, desto verre er det (Huseby, 2010, p. 180, min oversettelse).

Basert på dette trekker han opp to terskler. Som den nedre grensen peker Huseby på alt som er nødvendig for å dekke menneskers grunnleggende behov – for eksempel klær, ernæring og husly. Den øvre terskelen er nivået av velferd en person er fornøyd med (Huseby, 2010, s. 181).

Huseby forsvarer både en negativ og en positiv tese for tilstrekkelighetsprinsippet. Den positive betyr i denne sammenheng at det er et mål å få alle på et nivå av velferd der de er fornøyde, altså den øvre terskelen, mens den negative tesen sier at hverken prioritering av de «dårligst stilte» eller hensyn til likhet spiller inn i fordelinger over dette nivået. Tiltak for å hjelpe personer under den maksimale (øvre) grensen har absolutt prioritet sammenlignet med personer over den. Mellom de to tersklene

[63] Huseby er riktignok agnostisk angående hva «velferd» står for, og det er mulig at han innlemmer objektiv liste-teorier under begrepet (Huseby, 2010, p. 181, fotnote 10).

vektes fordeler konkavt tyngre, jo lenger unna den øvre terskelen mottakeren er. Personer under den nedre terskelen har sterk-, men ikke absolutt prioritet over folk som befinner seg mellom de to grensene (Huseby, 2010, s. 185). Kort oppsummert er det aller viktigste ifølge prinsippet å sikre alle menneskers grunnleggende behov, for deretter å løfte dem opp på et nivå av velferd de er fornøyde med.

Husebys prinsipp fremstår umiddelbart plausibelt. Samtidig er det iøynefallende hvor nært beslektet teorien er med utilitarisme. Selv om Husebys prinsipp generelt er velferdsbasert, er minimumsterskelen (den nedre) hans kjennetegnet av ressurser. Også utilitarister prioriterer å sikre så mange som mulig sine grunnleggende behov, da dette trygt kan antas å bidra til spesielt mye velferd.[64] Det å oppnå at mennesker er fornøyde med velferdsnivået sitt fremstår også som et godt mål for en utilitarist. «Fornøyd med egen velferd» er riktignok et litt komplisert begrep fra et hedonistisk eller preferansetilfredstillelses-synspunkt, da fornøydheten allerede er innbakt i personers velferd ifølge disse tilnærmingene. Men denne tvetydigheten skyldes trolig primært at Huseby ikke tar stilling til hva velferd består av i denne artikkelen. Så fra de nevnte velferdsteorienes perspektiver er det naturlig å anta at den øvre terskelen for tilstrekkelig velferd er «ganske høyt».[65]

Når alt det er sagt, er det disse tre måtene de to prinsippene mest markant skiller lag:

1. Prioriteten av de dårligere stilte mellom de to tersklene
2. Den absolutte prioriteten av personer under den øvre tilstrekkelighetsgrensen av velferd fremfor personer over den.
3. Den sterke prioriteten av personer som ikke har sine grunnleggende behov dekket fremfor personer over den.

De tre poengene kan altså knyttes til en form for prioritarianisme, og jeg diskuterer slike intuisjoner nøyere i 4.6. Før det er det verdt å merke seg at alle de tre punktene innebærer å godta mindre aggregert velferd enn det det ville vært mulig å oppnå, altså *sløseri*. Dette er en utfordring også Huseby erkjenner at tilstrekkelighetsprinsippet har (Huseby, 2010, s. 186–187). Han vurderer derimot

[64] Huseby lar spørsmålet om prioriteringer under den nedre terskelen stå (halv)åpent, men foreslår utilitarisme i fotnote 21 (Huseby, 2010, p. 185).
[65] Som er en svært upresis betegnelse fra min side, men likevel tilstrekkelig i denne omgang.

dette å være en mindre kamel å svelge enn det utilitarismen kan møte på av implikasjoner – for eksempel i form av at en persons langvarige og uutholdelige smerte kan overskygges av at mange nok personer unngår en mindre skuffelse. Huseby viser til henholdsvis Tim Scanlon og Paula Casal, som skisserer et slikt scenario (Scanlon, 2000, s. 235; Casal, 2007, s. 319–329).[66] Jeg har riktignok tatt for meg situasjoner som ligner på dette i del 4.2. Siden dette eksempelet ikke er helt likt som de tidligere nevnte, og det knytter seg til en annen betydningsfull kritikk av utilitarisme, er det imidlertid på sin plass å gjøre noen refleksjoner rundt Scanlon, Casal og Husebys innvending.

4.5.2 Verdensmesterskapet i fare – en motbydelig konklusjon

I tankeeksperimentet har et barn vandret inn i senderommet til en TV-stasjon som står bak produksjonen for VM-finalen i fotball. Med et uhell velter hun tungt utstyr på seg selv, som knuser hånden hennes, og gir henne svært vonde elektriske støt. Hun kan heldigvis reddes ut av situasjonen med det samme, men det vil nødvendigvis innebære stans i sendingen. Siden manglende TV-signaler ville skuffet millioner av seere, er den riktige tilnærmingen ifølge utilitarismen å vente med å hjelpe barnet til kampen er over. Dette fremstår avskyelig.

Det er flere måter å besvare denne innvendingen på. For det første, så kan man vise til Derek Parfits skille mellom hva som er objektivt og subjektivt riktig eller galt (Parfit, 1984, s. 25). Om en utilitarist kommer løpende inn i TV-stasjonens senderom og ser et barn med knust hånd som får elektrosjokk, har hen solide belegg for å tro at barnet opplever en uutholdelig smerte. På den andre siden av dilemmaet finner hen et langt større antall mennesker, hvor den presise reaksjonen på en eventuell stopp i TV-signalet er mindre kjent, og vanskelig å kalkulere fort. Fra vedkommendes standpunkt er det derfor naturlig å tro at å hjelpe barnet er det som fører til mest aggregert velferd i verden. Siden hen har mest grunn til å tro dette med ufullstendig informasjon *ex ante*, er det riktig, også ifølge en utilitaristisk logikk, å redde barnet, selv om det *ex post* viser seg å være objektivt feil. Å skulle dømme vedkommende,

[66] Paula Casal bygger videre på tankeeksperimentet Scanlon beskriver først. I teksten som følger tar jeg for meg førstnevntes versjon, da hun setter situasjonen mest på spissen. Jeg har i tillegg gjort noen ørsmå endringer selv, uten at det har noe å si for meningsinnholdet.

til tross for at hen handlet på måten hen hadde mest grunn til å handle på der og da, ville vært ren etterpåklokskap.

For det andre, så er det svært vanskelig å se for seg at det noen gang faktisk skulle vært riktig, selv i etterpåklokskapens navn, å ikke redde barnet så fort som mulig. Umiddelbart ville familien (og andre i nær relasjon) til barnet blitt sint og forferdet. Det er altså ikke kun barnet selv som ville tapt velferd på avgjørelsen. I tillegg kommer de mer indirekte konsekvensene. Om det ble kjent at man ikke blir reddet av utilitarister om vår egen lidelse gikk utover noen små gleder for mange personer, ville det skapt en uro i samfunnet generelt. Man ville for eksempel aldri turt å gå i nærheten av senderommene til TV-stasjoner, eller andre steder der lignende situasjoner kan oppstå. Dette ville ført til veldig mange små tap av velferd, mer enn en stopp i TV-sendingen under VM-finalen.

Den tredje måten å avvise innvendingen på knytter seg til en annen indirekte konsekvens: Å redde barnet ville svekket noen normer i samfunnet som er viktige å ha for å maksimere den aggregerte velferden over tid.

I de fleste situasjoner er det vanskelig eller umulig å kalkulere en handlings konsekvenser presist på forhånd. Å alltid regne på hva som er riktigst i en enkelt situasjon ville kostet mye energi, i tillegg til at menneskelige feil i resonnementene kunne ført til svært suboptimale valg. På sikt kan det derfor lønne seg for samfunnet å heller slutte seg til noen normer, motivasjoner og regler som jevnt over fører til de beste konsekvensene (Smart, 1973, Kapittel 7). Gitt at å følge disse fører til de beste konsekvensene over tid, vil utilitarismen fordre at man slutter seg til normene, motivasjonene og reglene, heller enn til en «ren utilitaristisk tankegang», der man må regne lenge på konsekvensene av hver handling på forhånd (Parfit, 1984, s. 29; Sidgwick, 1981b, s. 413).

Dette kan som nevnt være fordi at å slutte seg til disse fører til at vi selv gjør færre handlinger med dårlige konsekvenser enn om vi alltid prøvde å maksimere nytte, men noen ganger feilberegner, for eksempel fordi å alltid resonnere lenge over valg er slitsomt. Og selv om det er gitt at vi selv ikke feilberegner, ville det svekket nyttige normer i samfunnet å ikke redde barnet (f.eks. en norm om å redde lidende barn så fort som mulig, eller mer generelt å opptre empatisk). Gitt at ikke alle andre mennesker også velger å tenke rent utilitaristisk i enhver situasjon, er det totalt sett

mer nyttig å støtte opp om gode normer i samfunnet, enn å maksimere den direkte nytten i egne handlinger, samtidig som resten av samfunnet gradvis går over til å følge mindre nyttige normer. Det ville altså trolig vært riktig å redde barnet fort, både for å ikke svekke egne intuisjoner om at å redde barn er riktig, og for at samfunnet generelt skal beholde normen om å redde barn så fort det lar seg gjøre. På sikt maksimerer dette den aggregerte velferden, og er derfor et utilitaristisk gode.

På dette punktet kan det virke som at jeg forsvarer en form for regelutilitarisme, heller en handlingsutilitarisme. Jeg nevner regler i teksten, og viser til at slike kan være riktig å følge. Begrunnelsen for dette, derimot, er fortsatt handlingsutilitaristisk. Handlingsutilitarisme er riktig, men mennesket er for utsatt for å gjøre feil med en slik tankegang. De fleste aller fleste avgjørelser tar mennesket basert på de automatiske tankeprosessene, uten å ta i bruk den grundigste resonneringen – som er svært energikrevende (Greene, 2014, s. 696; Kahneman, 2012, s. 24). Derfor er en del av en handlings konsekvenser hvilke regler, normer og motivasjoner det bygger opp i aktøren selv og i samfunnet rundt, altså hva den automatiske reaksjonen vil bli når avgjørelser skal tas fremtiden. Siden det fører til økning i lykke over tid om intuisjoner som stort sett maksimerer lykke er vanlige, er hvilke intuisjoner en handling styrker eller svekker viktige deler av et handlingsutilitaristisk regnestykke.

I et fjerde svar på innvendingen kan man akseptere premisset om at det å vente med å redde barnet faktisk maksimerer velferd, selv alle de indirekte konsekvensene tatt i betraktning. Man kan også postulere at man har tilstrekkelig med kunnskap om de ulike konsekvensene på forhånd, og dermed vet at riktig handling (for en utilitarist) er å la barnet lide. Det kan fremstå svært avskrekkende at en utilitarist *kan* vurdere det slik. Jeg mener derimot dette ikke svekker utilitarismen. Det er, som i eksemplene i del 4.2, grunn til å sette spørsmålstegn ved våre intuisjoner angående dette scenarioet. Vi vet hvordan det er å være et menneske som har det vondt. Mange av oss vet også hvordan det er å ha det veldig vondt, for eksempel på grunn av elektriske støt. De som ikke har erfart slik, har i det minste et godt utgangspunkt til å forestille seg «uutholdelig smerte», barnet i eksemplet opplever. Vi vet i tillegg hvordan det er å være litt/middels skuffet, for eksempel fordi TV-en plutselig ikke spiller på lag. Vi har derimot, logisk nok, ikke noe forhold til hvordan det er *å være millioner av skuffede mennesker*. Følgelig er det svært vanskelig å på stående fot kunne gjøre seg opp en mening om hvor ille det egentlig er om så mange må

gjennom en (litt) negativ opplevelse. Vi klarer ikke å addere de små smertene presist nok. Intuisjonene trekker oss fort mot at det er best å unngå den store smerten vi kjenner til i barnets tilfelle.

4.5.3 Den motbydelige konklusjonen

Den siste besvarelsen tydeliggjør koblingen til et annet kjent tankeeksperiment som brukes i argumentasjon mot utilitarismen: Derek Parfits «Den motbydelige konklusjonen». Parfit peker på at utilitarister foretrekker en verden med svært mange mennesker som har liv som så vidt er verdt å leve, fremfor en verden med få mennesker som har veldig gode liv (Parfit, 1984, s. 389–390). Dette mener han er en motbydelig konklusjon på dilemmaet, og en grunn til å avvise visse former for utilitarisme. Men Parfit baserer seg på en antakelse om at mennesket faktisk kan forestille seg de to nevnte scenarioene, og kan stole på intuisjonene sine i slike scenarioer. Det virker derimot lite sannsynlig. At mennesker ikke kan forestille seg ordentlig høye tall er godt dokumentert, og å se for seg to slike eksempler krever mye abstraksjon. Tar man et steg tilbake er det heller ikke åpenbart at utilitarismens svar på dilemmaet er så motbydelig heller. Parfit postulerer tross alt at livene er *verdt* å leve, selv om det er så vidt. Er det virkelig motbydelig å gi mange mulighet til dette, fremfor at noen få skal få veldig gode liv?[67] Nok en gang virker problemet å være at mennesker, i hvert fall intuitivt, sliter med å addere veldig små, men svært mange goder.

4.5.4 Motbydelige fartsgrenser?

Enda et moment angående innvendingen om barnet og VM-finalen, er at lignende avveininger egentlig er ganske vanlige, uten at vi umiddelbart reagerer like kraftig på disse. Barbara H. Fried peker på motorveiers fartsgrenser som en illustrasjon (Fried, 2020, s. 14). Nesten uansett hva den aktuelle fartsgrensen er, så kunne enda flere ulykker vært unngått om den ble senket. Å la biler kjøre relativt fort, er å akseptere at noen få ulykker skjer, i bytte mot at et stort flertall slipper irritasjonen det ville vært å bruke lenger tid på å komme frem. Siden en alvorlig skade ville tatt den uheldige

[67] Som nevnt i del 3.2.2, så påpeker Parfit selv at intuisjonene våre ikke er til å stole på i Nozicks lykkemonster-eksempel. Parfits tankeeksperiment er et forsøk på å skape et eksempel som unngår problemet – men jeg mener han mislykkes. For et detaljerte motargument mot Parfits innvending, se for eksempel *Why Derek Parfit had reasons to accept the Repugnant Conclusion* (Tännsjö, 2020).

bilpassasjeren under den øvre grensen for velferd, nødvendiggjør tilstrekkelighetsprinsippet at fartsgrensen senkes.[68] Å senke grensen fra 110 km/t til 90 km/t for å oppnå en større sikkerhet fremstår kanskje rimelig. Men ulykker skjer også i denne farten. Jo lenger ned fartsgrensen flyttes, jo mindre riktig fremstår det å gi de som kommer under den øvre velferdsterskelen absolutt prioritet.[69]

4.5.5 Utilstrekkelige sko

For et rent «fordelingseksempel», kan man se for seg at alle ressurser i Andeby er delt ut med å følge tilstrekkelighetsprinsippet, utenom et par sko som er igjen uten eier.[70] Skoene er i størrelse 47, og Fetter Anton er den eneste med føtter de passer til. Samtidig er han akkurat over den øvre grensen for tilstrekkelighet, mens Petter Smart, med skostørrelse 42, er et lite stykke under. Fetter Anton har sko fra før, og trenger ikke et nytt par, men om han fikk det, ville de vært til nytte. Petter Smart er heller ikke i mangel på sko, og ville satt dem ut på en tom hylle som pynt. Det vil ikke gjøre at han når nivået der han er fornøyd, men å få en gratis ting vil gi akkurat nok glede et øyeblikk til at det teller som en økning i velferd. Fetter Anton har ingenting han kunne byttet bort med Petter Smart mot skoene, som ville tatt sistnevnte over tilstrekkelighetsgrensen, uten at fetter Anton faller under. Det fremstår feil å gi sko som er fem størrelser for store til Petter Smart, når han får mindre velferd ut av ressursene enn Fetter Anton, som kunne brukt skoene på bena. Den absolutte prioriteringen av personer under den øvre terskelen gjør likevel at vi må det, om vi tar i bruk tilstrekkelighetsprinsippet. Heller ikke dette fremstår som en bedre teori enn utilitarismen.

4.6 Intuisjoner om prioritet og tilstrekkelighet

I de to siste kapitlene har jeg argumentert imot henholdsvis tilstrekkelighetsprinsippet og prioritarianisme. Likevel er det unektelig slik at mange av våre moralske intuisjoner peker mot de to nevnte prinsippene, så vel som egalitarianisme og maximin-kriteriet, heller enn utilitarismen. Disse intuisjonene må forkastes for å

[68] Samme kritikk treffer også maximin-prinsippet, og prioritarianisme mer generelt, avhengig av hvor tungt den aktuelle versjonen prioriterer de dårlig stilte.

[69] Her kan en forsvarer av tilstrekkelighetsprinsippet skyte inn at en senket fartsgrense også går utover mennesker under tilstrekkelighetsterskelen, eller innføre et unntak for den absolutte prioriteten hvis mange nok mennesker påvirkes av en handling. Men begge løsninger virker å skyve tilstrekkelighetsprinsippet veldig nært utilitarismen.

[70] Eksemplet er så klart rent hypotetisk – men sammenlignet med mange av innvendingene utilitarismen møter, virker det ikke urealistisk.

kunne opprettholde reflektert likevekt med utilitarismen som (fordelings)prinsipp. For å kunne ignorere de, er det nødvendig med en forklaring på hvorfor disse intuisjonene er feil.

En sannsynlig forklaring Joshua Greene har pekt på, er at å tenke riktig om noe så abstrakt som velferd er svært vanskelig, og at det ofte blandes med ressurser (Greene, 2013, s. 279–284). I en studie testet han og Jonathan Baron menneskers vurdering av henholdsvis ressurser og velferd (Greene & Baron, 2001).

Deltakerne i eksperimentet besvarte først hvor mye lyst de hadde til å bo i land med ulike fordelinger av ressurser, der de hadde samme muligheten til å havne blant de med få, middels eller mye ressurser. I andre steg gav de alle de ulike mengdene ressurser hver sin verdi på en skala fra 0 til 100, der 0 var den minste mengden ressurser og 100 var det meste. Skalaen var ment å være lineær, og vise hvor ønskelig det var å bli tildelt ulike mengder av den aktuelle ressursen. Respondentenes svar på disse to spørsmålene var konsistente med hverandre, og i tråd med at ressursers nytte er avtakende. Like store økninger i absolutte termer ble vurdert som mer ønskelige, jo lavere utgangspunktet lå. Med andre ord ble for eksempel en lønnsøkning fra 15.000 til 25.000 dollar vurdert som nyttigere enn en lønnsøkning fra 40.000 til 50.000.

I eksperimentets tredje steg ble respondentene bedt om å vurdere land med ulike fordelinger av ressurser igjen. Denne gangen fikk de imidlertid ikke oppgitt konkrete ressursfordelinger, men ressursfordelinger «konvertert» til velferd, basert på steg to av eksperimentet. Heller enn å se ulike inntekter, fikk de tall på hvor nyttig de selv hadde vurdert den aktuelle mengden med inntekt.[71] Likevel vurderte respondentene fordelingene på akkurat samme måte som om det var inntekt, og vektla økninger fra lavt nivå av velferd tyngre enn like store absolutte økninger fra et høyere utgangspunkt. Med andre ord fant deltakerne en økning fra 0 til 25 i «ønskelighet» som mer ønskelig enn en økning fra 75 til 100 – til tross for at de selv hadde laget den lineære skalaen, og at slike økninger derfor burde være likeverdige.

[71] I eksperimentet fikk de strengt tatt ikke oppgitt at det var dem selv som hadde gjort konverteringen over til velferd, men «noen som er akkurat som deg» (Greene, 2013, s. 282; Greene & Baron, 2001, s. 247–248). Dette endrer ikke resonnementet.

Resultatene tolker Baron og Greene – korrekt, mener jeg – dit hen, at mennesker blander velferd med ressurser (Greene & Baron, 2001, s. 252). Dette fremstår som en plausibel forklaring på en del av kritikken utilitarisme har møtt, og intuisjoner som trekker mot andre prinsipper. Som allerede nevnt, er en vanlig innvending mot utilitarisme at det kan innebære at svært fattige mennesker må «ofre seg» så andre får det bedre. Dette er veldig lite sannsynlig, den avtakende nytten gjør at en utilitarist vil foretrekke å gi ressurser til de dårligst stilte. Om man derimot blander ressurser og velferd, og derfor tenker på utilitarismen som maksimering av sistnevnte, fremstår de mulige implikasjonene brått mindre appellerende. John Rawls tar for eksempel feil i at slaveri potensielt kan maksimere velferden i et samfunn. At en slik ordning kan maksimere de totale ressursene, er imidlertid ikke utelukket (Greene, 2013, s. 280).

Samme forvirring kan forklare den intuitive tiltrekningskraft i andre fordelingsprinsipper. Like store hopp i *velferd* er likeverdige, og nettopp derfor er det naturlig å vekte økninger i *ressurser* hos de dårligst stilte tyngre, og passe på at alle har nok. Dette er i tråd med utilitarisme som fordelingsprinsipp. Prioritarianistiske eller tilstrekkelighetsteoretiske intuisjoner angående ressurser sammenfaller altså med en maksimerende tankegang angående velferd, eller utilitarisme, med andre ord. Siden disse prinsippene er ment å omhandle velferd, fremstår utilitarismen som en bedre teori om fordelingsrettferdighet enn de andre jeg har vurdert.

4.7 Konklusjon så langt

Jeg har til nå vurdert fire konkurrerende prinsipper, og sammenlignet dem med utilitarismen. Basert på diskusjonen over, mener jeg man på en mer akseptabel måte kan opprettholde reflektert likevekt med utilitarismen enn både Rawls' teori, egalitarianisme, prioritarianisme og tilstrekkelighetsteori. Utilitarisme fremstår som et mer rasjonelt valg for representanter som skal enes om samfunnskontrakt bak uvitenhetens slør enn maximin-kriteriet. Det treffes ikke av *Innvendingen om utjevning nedover*, slik som egalitarianismen. Det gjør riktignok ikke prioritarianismen eller tilstrekkelighets-teorier heller. Disse prinsippene fremstår umiddelbart appellerende. Deres beste sider er imidlertid også til stede i utilitarismen. I tilfellene der prioritarianistiske eller tilstrekkelighetsteorietiske løsninger er gode, sammenfaller de med utilitarismen. I situasjoner der de skilles virker utilitarismen ofte kontraintuitiv. Som beskrevet over, skyldes imidlertid dette menneskets vansker med

å tenke klart om velferd, heller enn ressurser. Tatt dette i betraktning, bør de alternative prinsippene forkastes til fordel for utilitarismen, i søken etter riktig teori om fordelingsrettferdighet.

I neste del av oppgaven vurderer jeg om den nyttemaksimerende tilnærmingen bør justeres, for å ta høyde for gjengjeldelse eller fortjeneste i samfunnet den anvendes på – altså holde mennesker ansvarlige for deres handlinger. Tanken om moralsk ansvar krever noen forutsetninger angående *problemet fri vilje («the problem of free will»)*. Selv om denne metafysiske problemstillingen er såpass nært knyttet til spørsmål angående rettferdighet, gjøres koplingen relativt sjelden i litteraturen. Teoretikere innenfor fordelingsrettferdighets-debatten tar gjerne en viss grad av menneskelig fri vilje for gitt, eller begrenser sine prinsipper til å være gjeldende mellom aktører som har oppnådd samme grad av fortjeneste. Som blant annet moralsk flaks-diskusjonen viser derimot, er dette å utelate en svært sentral del av teoretiseringen angående rettferdighet. I delen som følger oppsummerer jeg blant annet hvordan Thomas Nagel gjenkjenner koblingen mellom moralsk ansvar og fri vilje-spørsmålet. Deretter undersøker jeg hvilke premisser angående den problemstillingen som er rimelige å legge til grunn når man diskuterer temaer det har implikasjoner for – slik som moralsk ansvar, og følgelig også (fordelings)rettferdighet.

5 Moralsk ansvar

> «Det anses allment rettferdig at hver person oppnår det godet eller ondet som han fortjener, og urettferdig om han skulle skaffe seg et gode, eller bli tvunget til å gjennomgå et onde, som han ikke fortjener. Dette er kanskje den klareste og mest ettertrykkelige formen ideen om rettferdighet blir oppfattet på i befolkningen.» (Mill, 1863, s. 218, min oversettelse)

Som Mill påpeker, er et av de intuitivt mest tiltrekkende konseptene i forbindelse med rettferdighet noe jeg enda ikke har diskutert, nemlig fortjeneste. Det oppfattes umiddelbart riktig om det går bra med noen som man vet jobber hardt og gjør mye godt. I motsatt tilfelle, om vedkommende generelt oppfattes lat eller lite sympatisk, føler man ikke nødvendigvis på samme empatien. Dette er spesielt relevant i forbindelse med staters fordeling av goder, som finansieres av samfunnets kollektive ressurser. Det virker rimeligere å bruke disse midlene på personer som bidrar mye både til egen og andres velferd, enn til noen som sløser bort de samme pengene.

Når man lar personers atferd være en avgjørende faktor for hvilke goder eller onder de får, holder man dem ansvarlige for sine handlinger. Det naturlige spørsmålet å stille i den sammenheng fra et normativt perspektiv, er *hva mennesket er moralsk ansvarlig for*.[72]

Relaterte begrep til dette er gjensidighet og gjengjeldelse. Å straffe noen for usedelighet eller belønne vedkommende for gode gjerninger skiller seg noe fra tanken om fortjeneste, og det er kun det siste som er direkte relevant for problemstillingen min. I teksten som følger behandler jeg likevel alle disse som et samlet begrep. «Straff» kan altså både bety pengebøter for forbrytelser eller ugunstige ressursfordelinger i form av lite fordelaktige skattesatser for visse grupper. Selv om motivasjonen er ulik, innebærer begge disse konseptene å tildele personer flere (eller færre) goder (eller onder) enn det de ellers ville fått, basert på handlinger de har utført. I dette kapittelet undersøker jeg om mennesket kan holdes moralsk ansvarlig, siden dette er en forutsetning tanken om fortjeneste hviler på. I forbindelse med moralsk ansvar er det ofte nyttig å kunne snakke om straff, og eksemplifisere med forbrytelser. Men selv om jeg tidvis bruker begreper som er mer relevante for juridisk rettferdighet enn fordelingsrettferdighet, er formålet med diskusjonen å belyse sistnevnte.[73]

I noen tilfeller fremstår det klart at mennesket ikke kan holdes ansvarlige for sine handlinger eller deres konsekvenser. Man er automatisk mildere i møte med barn eller mentalt utviklingshemmede som gjør forbrytelser, enn om de samme handlingene begås av voksne, friske folk. At noen som ikke har nok mat til å overleve raner en butikk, er også enklere å akseptere enn om en rik investor stjeler til seg mat. En person som med et uhell tråkker noen på foten i en folkemengde, tilgis fortere enn en som gjør det samme med fullt belegg. Intuitivt virker det feil å

[72] Jeg diskuterer ikke teorier om sjanselikhet innenfor i min oppgave, da de kan anses å være en kombinasjon av ulike konsepter (egalitarianisme og moralsk ansvar) jeg allerede beskriver. Det er imidlertid verdt å nevne at resonnementet jeg gjør her også tilbakeviser disse. Teorier om sjanselikhet fordrer (grovt sett) å kompensere mennesker for ren flaks, altså aspekter helt utenfor menneskets kontroll som påvirker livet deres (Dworkin, 1981, s. 293). Det er dårlig ren flaks å plutselig bli kreftsyk, til tross for å ha levd et sunt liv. Valgflaks, altså utfallet av sjanser mennesker selv tar, skal derimot ikke kompenseres (dette eksemplifiseres gjerne med gambling). Siden jeg konkluderer dette kapittelet med at mennesket ikke er moralsk ansvarlig, bør alle utfall anses som ren flaks. Dette gjør teorier om sjanselikhet til ren egalitarianisme, noe jeg allerede har diskutert.

[73] Som beskrevet i del 2.1.3, er det nærliggende å tro at diskusjonen angående moralsk ansvar også har implikasjoner for juridisk rettferdighet (i forbindelse med straff og gjengjeldelse). Jeg undersøker imidlertid kun hva mine refleksjoner innebærer for fordelingsrettferdighet (i forbindelse med fortjeneste).

dømme mennesker basert på hvor heldige de er. «Flaks» betyr i denne sammenheng alt av eksterne faktorer som påvirker en handling og dets konsekvenser.[74] Både et menneske som er nødt til å bruke vold i selvforsvar, eller en intetanende person som tilfeldigvis tråkker på foten til noen kan sies å ha vært uheldige, heller enn å ha tatt dårlige valg. På samme måte er det uflaks i den dårlige oppdragelsen til et barn, eller livssituasjonen til mange fattige, som gjør at de for eksempel velger å rane en butikk – derfor feller man ikke en streng moralsk dom over disse.

Eksemplene gitt er riktignok ulike, men de har en fellesnevner: For å tilegne noen moralsk ansvar for en handlings konsekvenser, må det være resultatet av valg aktøren selv har tatt, der det er rimelig å forvente av aktøren at hen skaffet seg oversikt over de forskjellige alternativene for handling. Dana Nelkin oppsummerer dette med det hun kaller *Kontrollprinsippet*:

> *«Vi er kun moralsk vurderbare i den grad det vi vurderes for avhenger av faktorer under vår kontroll»* (Nelkin, 2019, min oversettelse).

5.1 Moralsk flaks

Å si at moralsk ansvar krever kontroll, byr derimot på nye problemer. Som Tomas Nagel påpeker, tilegner vi aktører moralsk ansvar langt oftere enn det kontrollprinsippet tilsier. Undersøker man ulike handlinger nærmere, virker menneskelig kontroll sjelden å være til stede på måten prinsippet fordrer.

I artikkelen *Moral Luck* identifiserer Nagel fire ulike former for moralsk flaks - altså eksterne hensyn man vanligvis lar påvirke vår moralske vurdering av personer, til tross for at dette går imot *Kontrollprinsippet* (Nagel, 1976, s. 140). Disse er henholdsvis resultatsflaks, omstendighetsflaks, konstituerende flaks og årsaksflaks.

5.1.1 Resultatsflaks

Resultatsflaks viser til hvordan forskjellig utfall av samme handling fører til ulik moralsk vurdering av aktører.

[74] I teksten bruker jeg «eksterne» i vid forstand. Det trenger ikke peke på forhold utenfor mennesket, men forhold som til syvende og sist er utenfor menneskets kontroll.

Kriminelle handlinger som lykkes, straffes for eksempel adskillig hardere, både juridisk og moralsk, enn tilsvarende handlinger som går skeis.[75] Om to personer tar hver sin pistol og skyter mot et menneske, vil den ene bli vurdert betydelig mildere enn den andre, om vedkommende tilfeldigvis hadde et uladet våpen. Mennesker som kjører i sterkt alkoholpåvirket tilstand møtes med hard kritikk, men om en ender opp med å kjøre på noen i trafikken, dømmes hen ekstra hardt, til tross for at enhver fyllekjører har like lite kontroll. Våre moralske vurderinger gjøres tilsynelatende med hjelp av etterpåklokskap ex post, selv om alle avgjørelser tas under usikkerhet ex ante.[76]

5.1.2 Omstendighetsflaks

Også omstendigheter påvirker hvilken moralsk dom man feller over en person, selv om dette ikke er noe hen selv har valgt. Innbyggere i Tyskland under nazismen fikk muligheten til å vise enten heltemot, ved å stå opp mot regimet, eller tvert om å være en del av det (Nagel, 1976, s. 145–146). I et mye omtalt eksperiment undersøkte Stanley Milgram menneskers lydighet overfor autoriteter (Milgram, 1963). Han fant en uventet stor velvilje til å påføre mennesker betydelig ubehag på oppfordring fra autoritetspersoner. Dette er blitt tolket i sammenheng med de grusomme handlingene under nazismen. Om en såpass vanlig egenskap som autoritetstro er hovedforklaringen på hvorfor så mange bidro til lidelse i Nazi-Tyskland, kan det tenkes at et betydelig antall mennesker ville handlet på tilsvarende måte, om de befant seg i samme situasjon.[77]

5.1.3 Konstituerende flaks

Hvordan et menneske er som person kan også sies å være et uttrykk for flaks. Tvillingstudier tyder på at politiske holdninger kan være ca. 40-50% genetisk arvelige (Churchland, 2019, s. 111). Det vil si at gener forklarer nesten halvparten av

[75] Dette poengteres av en rekke ulike filosofer. For en detaljert gjennomgang av hvorfor denne praksisen fremstår urettferdig, se spesielt side 53-57 i David Lewis' artikkel *The Punishment that Leaves Something to Chance* (Lewis, 1989, s. 53–57).

[76] At alle avgjørelser tas under usikkerhet, er et viktig poeng. Det er riktignok tilfeller dette er mer tydelig enn andre, men uansett hva man gjør finnes det en sjanse for at det fører til noe annet enn man forventer. Selv rutinerte og lovlydige sjåfører ender noen ganger opp med å forårsake en ulykke, selv om det nok skjer sjeldnere med dem enn med råkjørere. For en grundig beskrivelse av etterpåklokskap i moralske vurderinger, se side 33-37 i Facing Up to Scarcity (Fried, 2020, s. 33–37).

[77] Eksperimentet er blitt grundig kritisert. Det overordnede poenget er likevel gyldig.

variasjonen i politiske holdninger mellom mennesker. Miljø, slik som oppdragelse, er en annen viktig komponent.[78] Siden barn hverken velger sine gener eller oppdragelse, virker holdningene de vokser opp til å ha å være uttrykk for flaks. Men mens ingen mener at en persons høyde, som også i stor grad styres av gener, er vedkommendes ansvar, vurderes mennesker stadig basert på deres holdninger, personlighetstrekk og hvordan de handler basert på disse.

5.1.4 Årsaksflaks

Med årsaksflaks peker Nagel på at tidligere hendelser og omstendigheter avgjør hva som skjer i fremtiden. Selv om Nagel beskriver det som en egen kategori, fremstår dette som en samlebetegnelse på de to foregående punktene, siden «tidligere omstendigheter» er det som konstituerer et menneskes personlighet og avgjør hvilke situasjoner hen havner i. Utfordringen årsaksflaks utgjør for moralsk ansvar er, som Nagel påpeker, parallell med *problemet fri vilje* (Nagel, 1976, s. 146–148). Dersom det stemmer at tidligere hendelser og omstendigheter avgjør hvordan en person handler, har vedkommende aldri reelt sett mulighet til å opptre på en alternativ måte – hens handlinger er alltid en direkte konsekvens av (eksterne) årsaker. Dette er et deterministisk syn på menneskelig handling, som ofte anses å være problematisk for tanken om fri vilje og moralsk ansvar.

Samtidig har vi en sterk intuisjon om at de fleste aktører har et moralsk ansvar. Man ender derfor opp med følgende holdninger:

1. Mennesker er ansvarlige for (i hvert fall noen av) sine handlinger
2. Mennesker kan kun holdes ansvarlige for faktorer under deres egen kontroll
3. Menneskelig handling styres av faktorer utenfor deres egen kontroll

Kombinasjonen av 1 og 2 kolliderer med kombinasjonen av 2 og 3. Dersom eksterne faktorer avgjør hvordan et menneske handler, kan hen aldri gjøre noe annerledes enn måten hen handler på – hen har aldri reell kontroll. Følgelig kan de aldri holdes moralsk ansvarlige. Samtidig er de fleste overbevist om at mennesker kan holdes

[78] 40-50% kan riktignok være upresist, å tallfeste genetisk arvelighet er utfordrende, siden å skille miljø- og geneffekter er svært vanskelig (Plomin et al., 2016, s. 10–11). Poenget består uansett – personlig atferd forklares i stor grad av gener.

moralsk ansvarlige. En slik dissonans bryter med den reflekterte likevekten. Dette er en utfordring Nagel selv ikke ser en løsning på (Nagel, 1976, s. 150, 1986, s. 137).

5.2 Determinisme

I de neste sidene undersøker jeg foreslåtte løsninger på utfordringen Nagel skisserer. Som nevnt over er det et inntrykk av determinisme som utgjør «problemet». Det er derfor på sin plass å først beskrive hva determinisme står for.

Benedict de Spinoza formulerer det slik:

> «Fra en gitt årsak følger nødvendigvis en effekt; og motsatt, hvis det ikke er gitt en årsak, kan det umulig følge en effekt.» (Spinoza, 1954a, p. 42, oversatt til norsk fra den engelske oversettelsen av meg).

Som sitatet tyder på, knytter determinismen seg til troen på at de kjente fysiske lovene alltid er gjeldende – enhver hendelse i verden kan forklares som en effekt av en årsak. Dette innebærer at om ting er på en spesifikk måte i tidspunkt T, er måten ting går etter T fastsatt av nevnte mekanismer (Hoefer, 2016). Med den umiddelbart plausible antakelsen om at mennesket består av materie (og kun materie), og fysiske lover også gjelder mennesket, betyr det at ingen noen gang kunne opptrådd annerledes enn det de gjorde. Menneskelig handling er, ifølge dette synet, determinert av hvordan verden er i forkant av handlingen.

5.3 Mulige løsninger på *problemet fri vilje*

Som jeg skriver over, utgjør dette et problem for tanken om fri vilje. Vi kjenner mennesket som en del av naturen, og vet at fysikkens lover er gjeldende der. Samtidig virker det ikke å stemme overens med oppfatningen om at mennesket har kontroll over (noen av) sine handlinger. I sidene som følger vurderer jeg noen foreslåtte svar på problemet.

Det første alternativet jeg vurderer er fremmet av Harry Frankfurt. Han mener at determinisme er kompatibelt med fri vilje og moralsk ansvar. Deretter undersøker jeg to teorier som holder på oppfatningen om fri vilje med å avvise at all menneskelig handling er determinert, før jeg til slutt undersøker motsatte påstand nærmere, og ser hva en verden uten fri vilje innebærer. Litteraturen om *problemet fri vilje* er omfattende, og det er følgelig mye som kunne vært nevnt. Teoriene og teoretikerne

jeg diskuterer representerer imidlertid de viktigste motpolene i debatten (for en oversikt, se f.eks. O'Connor & Franklin, 2021, Kapitler 2, 3).

5.3.1 Kompatibilisme

Kombatibilisme er et syn som erkjenner at menneskelig handling er (eller kan være) determinert, men likevel ser rom for fri vilje og moralsk ansvar.

En som fremmer et slikt syn, det vil si som argumenterer for at determinisme ikke utgjør et problem for fri vilje eller moralsk ansvar, er Harry Frankfurt. Ifølge Frankfurt er muligheten til å opptre annerledes enn det man gjør ikke er avgjørende for om man er moralsk vurderbar eller ikke (Frankfurt, 1969, s. 829–830). Mangelen på alternativ vil si at en handling er determinert. Om muligheten for å handle annerledes ikke er en forutsetning for moralsk ansvar, er ikke en eventuell aksept av determinisme problematisk for konseptet. Følgende tankeeksperiment illustrerer Frankfurts poeng:[79]

> Sam er misfornøyd med politikken i byen sin, og bestemmer seg for å myrde borgemesteren. Han stoler på sin venn John, og forteller ham sine planer. John er like misfornøyd med tingenes tilstand, og støtter Sams prosjekt. Samtidig frykter han at Sam får kalde føtter. I all hemmelighet installerer han derfor et apparat i hjernen til John, som kan overstyre noen av valgene han tar. Hvis John virker å trekke seg fra sin opprinnelige plan om å drepe ordføreren, kan Sam overstyre dette, slik at John likevel gjennomfører mordet. De to vennene går sammen til rådhuset. Han viker aldri fra sin opprinnelige plan, og myrder borgemesteren uten at Sam trenger å aktivere apparatet. Man kan også anta at John hverken handler på impuls, er hypnotisert, hjernevasket, lurt eller lignende, og er en voksen, frisk person.

Med slike eksempler mener Frankfurt å vise at tilgangen til alternativer for handling ikke er avgjørende for hvorvidt man tilegner en aktør moralsk ansvar. John valgte selv å ta livet av borgemesteren, og ville gjort det også uten Sams apparat i hodet. Sam var ikke avgjørende for Johns atferd. Det fremstår intuitivt riktig å felle en

[79] Frankfurts originale tankeeksperiment beskrives på side 835-836 av *Alternate Possibilities and Moral Responsibility* (Frankfurt, 1969, s. 835–836). I etterkant er det blitt formulert flere eksempler med samme struktur for å illustrere Frankfurts poeng mer presist. I teksten over oppsummerer jeg en variant av slike tankeeksperiment beskrevet av John Martin Fischer og Mark Ravizza (Fischer & Ravizza, 1998, s. 29–30).

moralsk dom over John for hans handling. Dette til tross for at John ikke hadde tilgang til alternativer. Hadde han trukket seg i siste liten, og for eksempel innsett at det å myrde noen på grunna av politisk uenighet er å gå for langt, ville Sams apparat overstyrt disse argumentene. John hadde uansett endt opp med å utføre mordet. Til sammen virker altså Frankfurts eksempel å inneholde disse to forholdene:

1. John hadde ingen mulighet til å handle annerledes enn han gjorde.
2. John er moralsk ansvarlig for sin handling.

Likevel lykkes ikke Frankfurt i å forsone determinisme med fri vilje. I eksemplet fremstår John riktignok ansvarlig for sine handlinger. Dette er imidlertid fordi man antar at han befinner seg i en ikke-determinert verden, der han tilfeldigvis, uten å vite om det selv, er uten reelle alternativ for handling. Sam og hans apparats tilstedeværelse påvirker derimot ikke situasjonen. John står overfor et fritt valg mellom å myrde borgermesteren eller å ikke myrde hen, og velger førstnevnte. Om han velger sistnevnte, vil Sam overstyre beslutningen og borgemesteren blir myrdet likevel. I dette tilfellet ville det imidlertid ikke vært like naturlig å holde John ansvarlig for handlingen, selv om det er han som gjennomfører mordet. Heller enn å ansvarliggjøre determinerte aktører, viser Frankfurts tankeeksperiment at moralsk ansvar ikke er kompatibel med en determinert verden.

En videreskriving av eksemplet kan være illustrerende:

> Sam trengte riktignok aldri bruke apparatet sitt for å overtale John. Det viser seg imidlertid at en annen venn av John, David, hadde tenkt akkurat som Sam. Han hadde også implantert et apparat i hjernen til John. David ser en usikkerhet hos John før Sam rekker å legge merke til noe, og aktiviserer derfor apparatet sitt. På grunn av dette gjennomfører John mordet.

I et slikt tilfelle vil det igjen være unaturlig å holde John moralsk ansvarlig. Som nevnt i det første eksemplet, kan man anta at John hverken handler på impuls, er hypnotisert, hjernevasket, lurt, tvunget eller lignende. Derfor er det naturlig å forutsette at han ikke blir styrt av en til nå ukjent persons apparat i hjernen. Poenget er likevel at man i Frankfurts (og Fischers) eksempel antar at aktøren hadde hatt et valg om Sam ikke var til stede. I en determinert verden er det imidlertid ikke slik. Hvis verden er determinert, har John bare ett alternativ uansett. I så fall er det ikke Sam eller andre med hjerneapparater, men mindre synlige faktorer, slik som genene hans

eller påvirkning fra miljøet rundt, som styrer han. Det vanlig å gjenkjenne denne determinismen i barn, psykisk utviklingshemmede aktører, eller i situasjoner med synlige eksterne påvirkninger, slik som en folkemengde der det er vanskelig å ikke tråkke på hverandre, eller, som i dette eksemplet, en ond venn med apparat til å styre hjernen. I slike tilfeller skjønner vi umiddelbart at aktørene ikke kunne handlet annerledes enn de gjorde.[80] Samme holdninger faller imidlertid ikke like naturlig i tilfeller der de eksterne forholdene som avgjør aktørens handlinger ikke er like synlige – og man ender opp med å holde mennesker moralsk ansvarlig.

Frankfurts tankeeksperiment er en illustrasjon på at disse moralske skillene faller naturlig for mennesket, men lykkes ikke som et argument for å forsvare intuisjonene. Determinisme fremstår ikke kompatibelt med moralsk ansvar eller fri vilje.

5.3.2 Libertariansk fri vilje

Harry Frankfurt klarer altså ikke å forklare de tre holdningene Nagel peker på, beskrevet i del 5.1.4, samtidig. En annen fremgangsmåte er å akseptere at moralsk ansvar er inkompatibelt med full determinisme, men avvise at enhver menneskelig handling er resultat av eksterne faktorer. Libertariansk fri vilje er et indeterministisk syn, som kan oppsummeres med et sitat fra Aristoteles' *Fysikken*:

> Dermed kan staven (som brukes som en spak) som forskyver en stein selv bli beveget av hånden, som i sin tur blir flyttet av mannen hvis hånd det er. *Men mannen blir ikke skjøvet av noe annet enn seg selv* (Aristotle, 1934, p. 319, oversatt fra den engelske oversettelsen av meg, kursiv er lagt til av meg).

Ifølge et slikt syn kan altså (i hvert fall noe) menneskelig handling begrunnes med mennesket selv. Robert Kane deler opp teorier om libertariansk fri vilje i to underkategorier, henholdsvis *teorier om aktørforårsakelse* og *teorier om teleologisk forståelighet*. Hovedforskjellen mellom de to er at teorier om aktørforårsakelse baserer seg på en forestilling om menneskelig handling som skjer utenfor en årsakskjede, mens teorier om teleologisk forståelighet ikke gjør det (Kane, 1989, s. 221–222). Jeg tar for meg begge i tur.

[80] Skillet mellom situasjoner der moralsk ansvarliggjøring faller naturlig, og situasjoner der det ikke gjør det, beskriver godt av P. F. Strawson (P. F. Strawson, 2008). I førstnevnte tilfeller dominerer ifølge Strawson våre reagerende holdninger, mens det er med våre objektive holdninger vi vurderer visse handlinger uten å tilegne aktøren moralsk ansvar.

5.3.3 Teorier om aktørforårsakelse

Av slike teorier tar jeg for meg en versjon Richard Taylor beskriver i boken *Metaphysics* (Taylor, 1992, Kapittel 5). Taylor bygger sitt syn på to oppfatninger (som han selv kaller for «datapunkter»):[81]

> 1. Opplevelsen av at han (og andre mennesker) noen ganger overveier hva han skal gjøre, og
>
> 2. at det noen ganger er opp til han selv (eller den aktuelle handlende aktøren) hva han gjør (Taylor, 1992, s. 39).

Han erkjenner selv at verden ved første øyekast virker determinert, der alt som skjer er nødvendiggjort av tidligere hendelser. Han ser at denne determinismen ikke er kompatibelt med opplevelsene hans om eget aktørskap, men mener disse opplevelsene fremstår enda vanskeligere å forkaste enn determinismens metafysikk (Taylor, 1992, s. 36, 39, 42). Han kan imidlertid ikke løse dissonansen med det han kaller *simpel indeterminisme*, altså en verden der fri menneskelig handling skjer helt uten grunn. Taylor illustrerer absurditeten i en slik teori ved å beskrive en person hvis høyre hånd er fri i *simpel indeterministisk* forstand. Personens hånd ville beveget seg helt vilkårlig, med tilfeldige vink, slag og klapp, uten at vedkommende ville hatt mulighet til å kontrollere den – da dette ville vært en grunn, noe som ifølge teorien er utelukket.

Taylors libertarianske indeterminisme går ut på at mennesker noen ganger er selvbestemmende vesen (Taylor, 1992, p. 51). Det vil si at vi noen ganger er årsakene til vår egen oppførsel, fordi det er vi som gjennomfører handlingen – men ikke kun i form av å være del av en årsakskjede (Taylor, 1992, s. 52).

Som nevnt erkjenner Taylor at frie, ansvarsfulle handlinger ikke kan være helt tilfeldige. Ifølge han gjøres slike handlinger av en grunn. Det kan imidlertid ikke være en tilstrekkelig grunn, da dette ville innebære at handlingen er determinert. Som han

[81] Taylors datapunkter, kombinert med en vanlig oppfatning av at verden fremstår determinert, svarer i stor grad til de tre punktene Nagel viser til, beskrevet i del 5.1.4. Hovedforskjellen er at Taylor går et steg videre – mens Nagel ser en uløselig dissonans, forkaster Taylor oppfatningen om determinisme, og oppnår dermed koherens i sine oppfatninger. I tillegg er det verdt å nevne at Taylor ikke eksplisitt nevner moralsk ansvar blant sine punkter. Videre i teksten referer jeg som regel til Taylors punkter, da disse i større grad inngår i «fri vilje-sjargongen» resten av litteraturen er hentet fra. Men den overordnede diskusjonen forblir akkurat den samme.

selv skriver, er slike handlinger gjort av en grunn, men grunnen er ikke årsak til handlingen (Taylor, 1992, s. 51).

Siste påstand, at en handling gjøres av en grunn som ikke er årsak til handlingen, fremstår svært forvirrende. Når Taylor beskriver menneskers frie handlinger og deres rasjonale, viser han kun til det David Lewis kaller for ordinære grunner, ikke kontrasterende grunner (Lewis, 1986, s. 177, 230). Ordinære grunner gir riktignok en forklaring på hvorfor en person gjør det hen gjør, men begrunner ikke hvorfor hen gjør det ene fremfor det andre. Det er altså ikke en tilstrekkelig forklaring. At hendelser skjer uten en tilstrekkelig forklaring, bryter med de grunnleggende antakelsene beskrevet i del 5.2. At alle hendelser i verden nødvendiggjøres av en årsak, innebærer at disse har en tilstrekkelig grunn – det må alltid være en grunn til at ting er på en bestemt måte heller enn en annen, selv i tilfeller der disse grunnene ikke er kjent for mennesket (se f.eks. Leibniz, 1902, pargr. 32). Når Taylor i motsetning til dette postulerer at visse menneskelige handlinger kan skje kun av én ordinær grunn, beskriver han hendelser vi ellers antar at ikke skjer i verden. Siden ordinære grunner i seg selv ikke er nok til at noe skjer fremfor noe annet, må menneskets «frie» innsats være en årsak på toppen av den ordinære grunnen til at handlingen skal skje uten å være tilfeldig. Denne frie innsatsen kan ikke være forårsaket av noe, da handlingen i så fall ville vært determinert av denne årsaken, i kombinasjon med den ordinære grunnen nevnt i sted. Til en viss grad må menneskelig handling være *casa sui*, altså forårsaket av seg selv. Dette stemmer ikke overens med mekanismene vi kjenner fra verden ellers.

Taylor erkjenner selv at hans teori krever at man aksepterer to uvanlige metafysiske forestillinger, delvis beskrevet over. For det første, så kan ikke mennesket styres av de samme reglene som resten av den fysiske verden, gitt at hans teori er korrekt. Mennesket må kunne være selvgående, og må derfor «bestå i noe mer enn ting og hendelser» (Taylor, 1992, p. 52, min oversettelse). Med andre ord kan ikke mennesket kun være materie, siden det da bare ville bestått i «ting og hendelser».

For det andre, så må mennesket kunne forårsake ting av seg selv, selv om mennesket er en ting, ikke en hendelse. Når man ser en fotball knuse et vindu, så er det ikke fotballen selv som er grunnen til at vinduet knuser, men det faktum at den krasjet i vinduet i høy nok fart, og med solid nok masse. Det er altså hendelsen som

er årsaken til det knusende vinduet, ikke ballen. Årsaken til den hendelse er blant annet at noen sparket den (altså krasjet en fot i ballen med tilstrekkelig masse og fart til å dytte ballen). Mennesket må kunne være en årsak uten å være en hendelse – bare slik kan den være den første i en årsakslenke, og ikke forklares av tidligere hendelser (Taylor, 1992, s. 52).

Grunnen til at Taylor aksepterer disse uvanlige metafysiske fenomenene, er at de forklarer hans to faste datapunkter nevnt i starten av dette underkapittelet: (1) Opplevelsen av å noen ganger overveie forskjellige muligheter, for å så å bestemme seg; og (2) oppfatningen om at noen avgjørelse er opp til han selv, altså noe han har moralsk ansvar for.

Som han skriver, så «gir det mening å måtte tenke gjennom en handling som er min egen og hvis utfall er opp til meg selv, og ikke bare på noe mer eller mindre esoterisk jeg er nært assosiert med, slik som min tanker, viljer, valg eller lignende.» (Taylor, 1992, p. 52, sitat oversatt av meg). Taylors teori om aktørforårsakelse forklarer riktignok hans to faste datapunkter, men betaler en svært høy pris for dette. Å bryte med ellers etablerte fysiske regler for å kunne forklare disse oppfatningene er svært ad hoc.

Teorien han anser som det nest beste alternativet, determinisme, avviser riktignok innholdet i de to datapunktene. I en deterministisk verden er ingen avgjørelser dypest sett opp til den menneskelige aktøren, da de ikke på samme måte kan være *causa sui*, men kun et ledd i en årsakskjede. Følgelig er det heller ikke slik at mennesker reelt sett overveier valgalternativer for å velge selv – utfallet er allerede gitt (dog ikke kjent).[82] At de to datapunktene finnes, altså at mennesket har opplevelsen av både moralsk ansvar og overveielse før valg, har imidlertid også en forklaring innenfor deterministiske rammer. Jeg utdyper grunnen til at et determinert menneske kan ha opplevelsen av å være *causa sui* i del *5.4*. I korte trekk går det ut på å holde hverandre moralsk ansvarlige er gunstig for menneskelig samhandling – dette er altså en intuisjon mennesket evolusjonært sett har vært tjent med.

5.3.4 Opplevelsen av å overveie

[82] Denne påstanden avhenger av hva man legger i uttrykket *overveier*. Jeg bruker det som et uttrykk for prosessen som fører til valg mellom udeterminerte alternativer, og skiller det fra *opplevelsen av overveielse*.

Heller ikke opplevelsen av overveielse er inkompatibelt med determinisme. Det er naturlig at mennesket grubler når hen selv enda ikke er bevisst på hva hen kommer til å foreta seg. Taylor avviser riktignok at dette er en god nok forklaring, og mener manglende informasjon ikke kan være grunnlag for overveielse (Taylor, 1992, s. 49). Dette illustrerer han med at en fengselsfange som ikke vet når hen blir henrettet ikke overveier mulighetene, men spekulerer og/eller gjetter på hvilket øyeblikk det skjer.

Det er kanskje riktig at fengselsfangen ikke overveier muligheter, men det er i så fall ikke på grunn av at mangel på informasjon utelukker overveielse. Taylor poengterer selv at man kun kan gjøre slikt i forbindelse med egen oppførsel, aldri andres (Taylor, 1992, s. 39). Når fengselsfangen spekulerer i når hen kommer til å dø, så gjetter hen på når ansatte ved fengselet bestemmer seg for å gjennomføre henrettelsen, og ikke noe ved hens eget sinn eller atferd. Eksemplet Taylor gir er diskvalifisert fra å kunne inneha (opplevelsen av) overveielse, uansett hvilken rolle tilgangen til eller mangelen på informasjon spiller for fenomenet.

Om det Taylor mener er at opplevelsen av overveielse ikke kan skje i total mangel på informasjon, har han rett. I slike tilfeller har en aktør ingenting å basere sine refleksjoner på.

En deterministisk opplevelse av overveielse krever imidlertid ikke total mangel på informasjon. Spinoza beskriver menneskets opplevelse av fri vilje som en bevissthet av egne handlinger, kombinert med uvitenhet rundt handlingens grunner (Spinoza, 1954b, s. 108).[83] I den sammenheng kan overveielse sees på som situasjoner der aktøren kjenner til noen ordinære grunner til å opptre på ulike måter, uten å selv være klar over et fullt nok sett med deterministiske årsaker til at de er tilstrekkelige for noen av alternativene. Opplevelsen av overveielse og et fritt valg er altså menneskets erfaring av å kjenne til noen, men ikke tilstrekkelig mange av årsakene som til sammen utgjør en kontrasterende (og følgelig determinerende) grunn til handlingen aktøren ender opp med. Taylor tar altså feil i at determinisme ikke gir en adekvat forklaring på denne opplevelsen.

Det er også verdt å merke seg at disse to oppfatningene – altså den deterministiske versjonen av opplevelsen av å overveie ting, og oppfatningen om at ingen handling

[83] For en beskrivelse av hvordan bevissthet kan passe inn i et deterministisk verdenssyn, se for eksempel Daniel M. Wegners oppsummering av sin egen bok *The illusion of conscious will* (Wegner, 2004).

dypest sett er opp til mennesket – gir hverandre gjensidig støtte. Hvis det er på måten jeg beskriver at opplevelsen av overveielse faktisk finner sted, stemmer det at mennesket ikke er den «endelige» årsaken til noen hendelser (og hen kan derfor ikke holdes moralsk ansvarlig). Og hvis det er slik at mennesket ikke er *causa sui*, men en bit av en årsakskjede, så gir det mening at opplevelsen av overveielse oppstår på den skisserte måten. Det er også mulig å gi en adekvat forklaring på hvorfor man har «feil i dataen» innenfor determinismens rammer, altså hvor menneskelige oppfatninger som virker å motsi determinisme stammer fra. Dette utdyper jeg i del 5.4.

Man kan altså bygge et koherent verdensbilde – opprettholde reflektert likevekt – med å forkaste oppfatningen om at mennesket er selvhandlende. Richard Taylor syns riktignok det er en for høy pris å betale. For meg virker det imidlertid langt mer akseptabelt enn å gi mennesket unntak fra ellers gjeldende fysiske lover, og innføre en hittil ukjent form for kausalitet. Determinisme er derfor en bedre kandidat til å være sann enn Taylors libertarianske teori om aktørforårsakelse.

5.3.5 Teorier om teleologisk forståelighet

Robert Kane, som er en av de mest fremtredende filosofene innenfor den libertarianske fri vilje-tradisjon, erkjenner at teorier om aktørforårsakelse innebærer alt for usannsynlige metafysiske forutsetninger (Kane, 1989, s. 248). Samtidig deler han Taylors (og Nagels) sterke intuisjon om at noe menneskelig handling gjøres av fri vilje, og avviser følgelig full determinisme. Han setter seg derfor fore å forklare de nevnte datapunktene på en mer plausibel måte enn Taylor. Dette gjør han med å beskrive en teori om teleologisk forståelighet.

En av hovedgrunnene til at formuleringen av en teori som forklarer moralsk ansvar er vanskelig, er det også Richard Taylor berører i sin diskusjon av *simpel indeterminisme*. Ikke bare er moralsk ansvar inkompatibelt med determinisme. Fullstendig indeterminisme hjelper heller ikke på saken. Om handling ikke er forårsaket av forutgående hendelser, virker det å være totalt overlatt til tilfeldigheter – og følgelig også utenfor aktørens kontroll. Mennesket kan ikke holdes ansvarlig for

handlinger som er tilfeldige.[84] Kanes teori om teleologisk forståelighet er et forsøk på å lage et system der mennesket kan holdes moralsk ansvarlig, til tross for truslene både determinisme og indeterminisme utgjør mot dette.

Kane anerkjenner at mange valg er determinerte av for eksempel aktørens personlighet. Han mener imidlertid at det finnes enkelte valg som ikke er avgjort på forhånd, og følgelig kan anses å være frie. Utfallet av disse valgene former aktørens personlighet, derfor kan mange av de determinerte valgene også spores tilbake til et fritt valg den aktuelle aktøren har foretatt. Følgelig er det, i fravær av ytre faktorer som for eksempel tvang, på sin plass å holde mennesket moralsk ansvarlig for sine handlinger (Kane, 1989, s. 252, 1999, s. 224).

De nevnte *selvdannende handlingene*, som Kane kaller de, finner sted når en aktør rives mellom de ulike grunnene til å gjøre det ene eller det andre (Kane, 1989, s. 236, 241, 1999, s. 224). Dette kan for eksempel være fordi vedkommende veier om hen skal opptre uselvisk opp mot det å fremme sine egne interesser. Et annet eksempel er valget mellom kortsiktige og langsiktige mål. Om de deterministiske grunnene – de forskjellige tilgjengelige valgmulighetene, aktørens personlighet og sinnstilstand – taler akkurat like mye for begge alternativene, er det, før valget skjer, ubestemt hva utfallet blir – altså er verden indeterministisk.[85] På det fenomenologiske plan går aktøren i slike tilfeller gjennom usikker sjelsøking og overveielse. Dette skjer mens hjernen beveger seg vekk fra termodynamisk likevekt, til en kaotisk tilstand som er følsom for utfallet av tilfeldige utfall på kvantenivå (Kane, 1999, s. 224–225). Med andre ord: Når en person står overfor to alternativer vedkommende har akkurat like gode grunner til å velge, avgjøres hvilke grunner som vinner frem med en indeterministisk endring på kvantenivå i hjernen.[86]

[84] Mange ulike filosofer har beskrevet denne utfordringen for indeterminisme og moralsk ansvar. For spesielt klare formuler og diskusjoner av problemet, se side 19-20 *The Impossibility of Moral Responsibility*, samt side 371-374 av *The Mystery of Metaphysical Freedom* (G. Strawson, 1994, s. 19–20; van Inwagen, 1998, s. 371–374). Også Kane selv beskriver innvendingen grundig, før han fremmer sine motargumenter (Kane, 1999, s. 229).

[85] I teksten beskriver jeg *selvdannende valg* som ikke-determinerte valg mellom to alternativer. Dette fordi Kane gjennomgående omtaler det slik. Når det er sagt, ser jeg ingen grunn til at samme modell ikke skal kunne anvendes for valg med flere alternativer enn 2. Dette endrer uansett ikke substansielt på resonnementet.

[86] Når jeg skriver «like gode grunner», mener jeg her «like betydningsfulle deterministiske årsaker». Dette utgjør en kombinasjon av hva valgalternativene er, hvordan aktøren er som person, vedkommendes sinnstilstand osv.

Ifølge Kane oppfyller slike selvdannende handlinger de nødvendige kriteriene for at et system av moralsk ansvar skal kunne bygges på disse (Kane, 1989, s. 254, 1999, s. 240). Handlingen aktøren gjør er ikke forhåndsbestemt – før valget skjer har vedkommende flere alternativer åpne. Samtidig skjer valgene av en årsak, og disse årsakene er aktørens egne grunner. Siden slike valg former aktørens personlighet, som i tur er med på å determinere senere handling, kan mennesket generelt holdes moralsk ansvarlig, ifølge Kane.

Det er verdt å merke seg at denne teorien gir en forklaring på Taylors tidligere beskrevede datapunkter.[87] Kane nevner selv at aktøren opplever overveielse mens hjernen går vekk fra den termodynamiske likevekten. Utfallet av slik overveielse er i tillegg valg som er aktørens *egne*, da de ikke kan spores tilbake til noen tidligere hendelser, kun personen selv.[88] I motsetning til Taylor kombinerer Kane indeterminisme og valg uten å ty til nye «metafysiske oppfinnelser». Med endringer på kvantenivå viser han til den eneste kjente indeterminismen i det fysiske universet.[89] Hvorvidt bevegelser på kvantenivå faktisk har slike indeterministiske utslag på makronivå er et empirisk spørsmål, som foreløpig ikke er besvart.[90] Robert Kane lykkes med å beskrive en indeterministisk teori som er langt mer plausibel enn for eksempel Richard Taylors system. I de siste to setningene av *Two Kinds of Incompatibilism* foreslår Kane å kalle den formen for selvdannende handlinger han beskriver for *fri vilje* (Kane, 1989, s. 254). Aksepterer man denne betegnelsen, noe jeg mener man kan, bør man ikke utelukke at mennesket har fri vilje. Min påstand er imidlertid at det ikke rettferdiggjør å holde mennesket moralsk ansvarlig, selv om det

[87] Denne diskusjonen tydeliggjør en forskjell mellom datapunktene til henholdsvis Taylor og Nagel. Kanes teori viser til handlinger som kun kan knyttes til mennesket, og svarer dermed til Taylors punkter. Som jeg beskriver i de neste sidene innebærer derimot dette ikke moralsk ansvar, og går dermed imot Nagels oppfatninger.
[88] At disse situasjonene oppstår, altså at personen er indeterminert (eller «like determinert») mellom gitte alternativer, kan riktignok spores tilbake til tidligere, «eksterne» hendelser. Hvilket av alternativene for handling vedkommende ender opp med, kan derimot ikke spores tilbake til noe slikt. Dette utdyper jeg de neste sidene.
[89] Kane er ikke den første til å foreslå en slik type løsning på fri vilje-problemet. Så tidlig som i antikkens Hellas skal Epikur ha vist til atomenes uforutsigbare svingninger som en kilde til at den deterministiske årsakslenken brytes, en forklaring blant annet den romerske epikureeren Lucretius videreutviklet (Huby, 1967, s. 358; Lucretius, 1924, s. 113–115).
[90] For ordens skyld: Det er heller ikke full konsensus om at bevegelser på kvantenivå faktisk er udeterminerte, og ikke heller skyldes hittil ukjente variabler. All den tid det er mulig (og sannsynlig) at slike tilfeldigheter forekommer, endrer dette uansett ikke på poenget i diskusjonen over.

skulle vise seg at man faktisk har en slik form for fri vilje. For å utdype dette, bruker jeg et av Kanes egne illustrasjoner på hans teori (Kane, 1999, s. 225–239).

I Kanes eksempel er en forretningsperson på vei til et møte som er viktig for karrieren hennes, da hun ser noen bli angrepet i en bakgate. Hun er da i en typisk situasjon som kommer til å ende i en selvdannende handling. Hun rives mellom sitt sterke og selvsentrerte ønske om å rekke møtet, og det tilsvarende sterke uselviske ønsket om å hjelpe personen som angripes. Siden årsakene som fordrer å gjøre det ene eller det andre er like sterke, er handlingen ikke determinert. Frem til valget skjer, er det usikkert om forretningsperson skynder seg videre, eller stopper i bakgaten. Forskjellen på en verden der hun velger møtet og en mulig verden der hun hjelper offeret, er derfor kun den tilfeldige bevegelsen i på kvantenivå i hjernen hennes. Absolutt alle hendelser frem til handlingen er ellers de samme.

Ifølge Kane bør hun holdes moralsk ansvarlig for sin handling, uansett hvilke av de to alternativene hun velger. Dette fordi det i begge tilfellene er hennes egne grunner som gjør at hun handler slik hun gjør. Hvis hun stopper og hjelper personen som blir angrepet, er de moralske grunnene og hennes samvittighet det som får henne til å gjøre det. Og hvis hun kjører videre mot jobb, er det på grunn av hennes selvsentrerte ønske om en blomstrende karriere. Som beskrevet i del 5.3.3, antar man som regel at mennesket ikke er moralsk ansvarlig for en handling som er tilfeldig, siden aktøren ikke kan ha kontroll over dette. Som Kane imidlertid påpeker, kommer indeterminismen i dette tilfellet ikke utenfra. Grunnen til at det er usikkert om forretningspersonen lykkes med å følge sitt moralske kompass og dermed stoppe, er hennes eget ønske om å følge sin selvsentrerte motivasjon og kjøre til jobb. På samme måte er det på grunn av hennes eget ønske om å følge sitt moralske kompass at det er usikkert om hun lykkes i å kjøre videre. Så selv om tilfeldighet spiller en rolle, er denne tilfeldigheten ikke noe eksternt til forretningspersonens grunner og valg, men en del av de (Kane, 1999, s. 231). Som Kane skriver lykkes forretningspersonen med noe hun prøvde på, uansett hvilket alternativ hun velger, og bør derfor holdes moralsk ansvarlig for sin handling (Kane, 1999, s. 233)

For å se hvorfor denne begrunnelsen av moralsk ansvar ikke er god nok, kan man undersøke forretningspersoners grunner til å velge å kjøre videre, og igjen ta i bruk

David Lewis' typologi.[91] Vedkommende har riktignok sine grunner til å kjøre videre. Disse er derimot kun ordinære grunner, ikke kontrasterende grunner. De gir riktignok en forklaring på hvorfor hun gjør det hun gjør, men begrunner ikke hvorfor hun gjør det ene fremfor det andre. Man kommer derfor ikke unna at utfallet av en tilfeldighet (en kvantebevegelse i forretningspersonens hjerne) er en INUS-forutsetning for at aktøren gjorde som hen gjorde.[92] For at forretningskvinnen skal kjøre videre til jobb i akkurat den situasjonen, er det nødvendig at nevronene beveger seg på en gitt måte i hjernen hennes i det avgjørende øyeblikket. Hvorvidt de gjør det er rent tilfeldig, og ikke noe aktøren har kontroll over. Denne indeterminismen kan ikke være grunn til å tilegne personer moralsk ansvar.

Det hjelper imidlertid ikke å ta vekk denne usikkerheten heller. For at kvanteprosessen i forretningspersonens hjerne ikke skal være avgjørende for at hun velger møtet, må årsakene for å kjøre videre være sterkere enn de for å stoppe. For eksempel må hun ha en sterkere egoistisk- enn moralsk motivasjon, eller så må hun rett og slett ikke legge merke til det pågående angrepet i bakgaten. I slike situasjoner er riktig nevronbevegelse i hjernen ikke en nødvendig forutsetning for å kjøre videre. Å unngå dette kommer riktignok med en pris. Om forretningspersonen kjører videre uten overveielse, er det fordi handlingen er determinert av tidligere hendelser. Slike determinerte handlingene kan heller ikke i seg selv være grunn til å holde aktøren moralsk ansvarlig - om noen skal være ansvarlige for en forutbestemt handling i Kanes system, må disse være determinert av et fritt valg de tok tidligere. Man kommer imidlertid enten til nye determinerte hendelser, eller til selvdannende handlinger, uansett hvor langt tilbake man følger årsakslenken.

Som beskrevet, så er selvdannende handlinger ikke mer enn to determinerte sett med ordinære grunner, i tillegg til en ren tilfeldighet som en nødvendig forutsetning for å skille de determinerte alternativene. Ingen av disse ingrediensene kan gi

[91] Resonnementet er akkurat like passende for det motsatte valget, altså det å stoppe for å hjelpe.

[92] En INUS-forutsetning er en nødvendig, men ikke tilstrekkelig del av en årsakslenke som er tilstrekkelig, men ikke nødvendig (Mackie, 1965, s. 245). I eksempelet over er det nødvendig at nevronene i forretningskvinnens hjerne gjør et gitt utslag, slik at grunnene hennes til å kjøre videre vinner frem fremfor grunnene til å stoppe. Denne prosessen i hjernen hadde imidlertid ikke funnet plass om hun ikke allerede hadde akkurat like gode grunner til å kjøre videre som til å stoppe. Altså er utfallet av tilfeldigheten en nødvendig, men ikke tilstrekkelig del av et sett med grunner som til sammen er tilstrekkelig. I en annen situasjon kunne hun derimot kjørt videre uten en slik prosess i hjernen – for eksempel fordi hun ikke så den angrepede personen, eller fordi hun selv hadde endret holdninger, og var mer egoistisk. Altså er det fulle, tilstrekkelige settet med grunner i det første eksemplet ikke nødvendig for å stoppe.

grunnlag for å tilegne mennesket moralsk ansvar. Robert Kane beskriver riktignok et plausibelt alternativ til determinisme, og viser at mennesket potensielt har fri vilje (i hvert fall med hans definisjon av sistnevnte begrep). Men heller ikke hans alternativ innebærer at mennesket er moralsk ansvarlig. Ønsker man en teori om menneskelig handling som muliggjør å holde aktørene moralsk ansvarlige, trenger man å akseptere spesielle metafysiske forutsetninger, som i Richard Taylors rammeverk. Da fremstår det bedre å la være.

5.4 Determinismens plausibilitet

Som Robert Kanes indeterministiske verden, virker også determinisme som en plausibel metafysisk posisjon. Det stemmer overens med det vi vet om fysiske regler og en rimelig antakelse om at mennesket består av kun materie. Som beskrevet i del 5.3.4, finner man også en god forklaring på hvorfor mennesket opplever å overveie sine valg i en determinert verden. Et siste datapunkt gjenstår likevel. Som Richard Taylor poengterer, har vi menneske ren intuitiv oppfatning av at noen handlinger er *bare våre* (Taylor, 1992, s. 32). I en determinert verden, der alle nødvendigvis følger av noe som skjedde tidligere, er ikke slike handlinger mulige. For å opprettholde den reflekterte likevekten med et deterministisk standpunkt, er det derfor nødvendig med en forklaring på hvorfor disse, ifølge determinismen, feilaktige, oppfatningene finner sted.

Min hypotese er at disse intuisjonene kan forklares i fire steg. Det er, i seg selv, ingen grunn til å ha en oppfatning om at en opplevelse er *bare min egen*. Det er derimot evolusjonært nyttig å holde andre mennesker (moralsk) ansvarlige for sine handlinger, fordi det er gunstig å kunne straffe dem. Intuitivt er det ikke riktig å tilegne mennesker moralsk ansvar for atferd som styres av eksterne faktorer. Derfor er det, for å kunne holde mennesker ansvarlige, nyttig å anse handling som aktørenes egen, om man ikke ser umiddelbare grunner til at dette er feil. Samtidig ville det vært vanskelig å opprettholde et bilde av at andre handler ut av egen fri vilje, mens man selv ikke gjør det. Følgelig tilegner man også seg selv denne egenskapen. Punktene er altså som følger:[93]

[93] Merk at steget mellom første og andre punkt også kan være relevant for Kanes teori om teleologisk forståelighet. Hans teori forklarer riktignok Taylors datapunkter som fordrer at noen handling kun stammer fra mennesket selv, men ikke Nagels datapunkt om at mennesket noen ganger kan holdes moralsk ansvarlig. Som Kane selv er et eksempel på, er det imidlertid mulig å (feilaktig, ifølge diskusjonen i del 5.3.5) slutte seg til en tanke om moralsk ansvar fra Kanes teori på andre måter enn den jeg skisserer her.

1. Det er evolusjonært nyttig å straffe eller belønne andre aktører
2. Kun mennesker med moralsk ansvar kan straffes
3. Mennesker er kun ansvarlige for handling som er deres egen
4. Jeg er som andre mennesker

Det er verdt å merke seg at stegene over også er et svar på oppfatningene Nagel peker på, beskrevet i del 5.1.4. Med en god forklaring på hvorfor man har en feilaktig intuisjon om at mennesker er moralsk ansvarlige, opprettholder man reflektert likevekt innenfor determinismens rammer. For å sannsynliggjøre stegene, utdyper jeg her spesielt første punkt.

5.4.1 Den evolusjonære fordelen ved moralsk ansvarliggjøring

For å illustrere hvordan intuisjoner om straff og belønning kan ha oppstått, tar jeg i bruk det spillteoretiske eksemplet *fangenes dilemma*. I teksten som følger fokuserer jeg primært på straff, da min oppfatning er at denne intuisjonen er enda sterkere enn den om belønning.

I fangenes dilemma mistenkes to personer, henholdsvis Spøkelseskladden og Svarte-Petter, for å ha begått noe kriminelt. Begge har muligheten til å tyste på den andre eller å holde tett. Fangene får ulik grad av straff, basert på hva de to velger å gjøre i dilemmaet. Aktørenes valg gjøres simultant, følgelig vet ingen av de to hva den andre foretar seg (Hovi, 2008, s. 37). Jeg illustrerer de mulige utfallene i tabell 3. Tallene beskriver antall år i fengsel, der Spøkelseskladdens utmålte straff står før Svarte-Petters.

Tabell 3

		Svarte-Petter	
		Holde tett	Tyste
Spøkelseskladden	Holde tett	1, 1	3, 0
	Tyste	0, 3	2, 2

Man kan se for seg dilemmaet fra Spøkelseskladdens perspektiv. Han ønsker så kort tid i fengsel som mulig. Han vet ikke hva Svarte-Petter gjør, men uansett Svarte-Petters valg, fører tysting til kortest mulig straff for Spøkelseskladden. Om Svarte-Petter holder tett, får Spøkelseskladden gå fri ved å tyste, fremfor å måtte sone i ett år. Og om Svarte-Petter tyster, får Spøkelseskladden 2 år fengsel ved å tyste, heller en 3, som ville vært konsekvensen av å holde tett. På den andre siden av dilemmaet gjør Svarte-Petter akkurat det samme resonnementet. Å tyste er altså dominant strategi for begge parter.[94] Følgelig er det naturlige utfallet at Spøkelseskladden og Svarte-Petter tyster på hverandre, og begge må sone i 2 år.

Dette kan sies å være et lite paradoks. Både Spøkelseskladden og Svarte-Petter foretrekker ett års fengselsstraff fremfor to, og de kunne oppnådd det resultatet med å holde tett om forbrytelsen sammen.[95] En løsning kunne vært at de to inngikk en avtale om å tie. Problemet er at løfter kan brytes. Om Spøkelseskladden holder avtalen, men Svarte-Petter likevel velger å tyste på Spøkelseskladden, ender Spøkelseskladden opp med å bli utnyttet, og å sone i 3 år mens Svarte-Petter går fri. Om dilemmaet er et enkelttilfelle, er det dømt til å ende i at begge velger å tyste, da ingen av de to kan stole på den andre.

Situasjonen endres imidlertid om Spøkelseskladden og Svarte-Petter planlegger å fortsette å begå forbrytelser sammen.[96] I slike tilfeller vil de to kunne lære hverandre å kjenne, og finne ut om medsammensvoren er til å stole på eller ikke. Igjen kan spillet sees fra Spøkelseskladdens perspektiv. Det er to måter Spøkelseskladden kan oppnå et bedre utfall på. Den ene er å selv gå fra å holde tett til å tyste. Den andre måten er at Svarte-Petter går fra å tyste til å holde tett. Om Spøkelseskladden selv tyster, får han enten 0 eller 2 års fengsel, avhengig av hva Svarte-Petter velger å gjøre. Og om det er gitt at Svarte-Petter holder tett, kan Spøkelseskladden få enten 0 eller 1 års fengsel, avhengig av hva han selv gjør. Det forventede resultatet

[94] Dominant strategi defineres slik: «En strategi er dominant (eller dominerende) hvis den leder til minst like godt resultat som enhver annen strategi, uansett hvilke strategier de andre spillerne måtte velge» (Hovi, 2008, s. 39).

[95] Situasjonen der ingen parter kan få det bedre uten at minst en av aktørene får det dårligere, kalles pareto-optimale (Hovi, 2008, s. 40). I denne situasjonen ville det vært pareto-optimalt om både Spøkelseskladden og Svarte-Petter holdt tett om forbrytelsen.

[96] For at resonnementet som følger skal fungere, må det være snakk om en ubestemt tidshorisont, altså må aktørene kunne anta at det er «uendelig» slike dilemmaer igjen (Hovi, 2008, pp. 81-82). Siden dette er gjeldende i sosiale relasjoner, som eksempelet er ment å illustrere, kan ubestemt tidshorisont legges til grunn.

av å ikke bli tystet på er altså høyere enn det forventede resultatet av å tyste. Med andre ord er det viktigere for Spøkelseskladden å bli stolt på av Svarte-Petter enn hva han selv velger å gjøre.[97] Måten å bygge tillitt på med Svarte-Petter er å ikke tyste på ham. Om situasjonen er gjentakende, er det derfor best for Svarte-Petter å bygge tillitt med å holde tett i slike dilemmaer – og akkurat samme resonnement gjelder fra Spøkelseskladdens perspektiv også.

Løfter man blikket fra Svarte-Petter og Spøkelseskladden, står mennesket stadig overfor situasjoner der den underliggende logikken er den samme: Man kan velge mellom å gjøre det som er best for felleskapet, eller maksimere egen vinning. Men om også alle andre gjør sistnevnte, gir det et enda dårligere resultat enn om man sammen med de andre prioriterer felleskapet. Det er for eksempel, fra et egoistisk perspektiv, optimalt om man kan ta maten til alle mennesker rundt seg, mens en selv ikke blir tatt fra. Motpolen til dette er å ikke stjele fra andre selv, mens folk rundt tar maten man eier. De to mellomposisjonene er enten at alle prøver å stjele fra alle, eller at alle holder seg til sine egne porsjoner – hvorav det siste er å foretrekke.

Disse eksemplene tydeliggjør at det er gunstig med samarbeid. Men for at samarbeid skal fungere, må deltakerne gjøre enkeltvalg som er suboptimale fra et egoistisk perspektiv. Spøkelseskladden og Svarte-Petter kommer til å havne i mange tilsvarende situasjoner sammen. Følgelig er det enkelt å forsikre seg om at de kommer til å samarbeide: I det den ene velger egoistisk, brytes tillitten, og i senere dilemma vil det være naturlig å tyste på hverandre. Det ønsker hverken Spøkelseskladden eller Svarte-Petter. I det daglige er det ikke alltid de samme personene man samhandler med, og det er ikke nødvendigvis identiske varianter av fangenes dilemma man befinner seg i. Men lignende logikk kan likevel fremme det felles beste. Om de som velger samarbeider konsekvent straffer de som opptrer egoistisk, vil det alltid lønne seg å jobbe for felleskapet. Straff har altså en viktig sosial funksjon. Det er også nyttig fra et egoistisk perspektiv: Ved å straffe alle som velger egoistisk, unngår man å bli utnyttet (altså at en annen aktør velger selvsentrert, mens man selv prioriterer felleskapet).

[97] Det finnes forskjellige måter å komme frem til ca. samme konklusjon på, og å vise til «tillit» er kun en av disse. For en beskrivelse av ulike strategier for gjentakende spill med ubestemt tidshorisont, se kapittel 7 av *Spillteori: En innføring* (Hovi, 2008, Kapittel 7). For mitt resonnement er forskjellen mellom disse ikke avgjørende, da alle kan forklare den intuitive appellen til fortjeneste og straff.

Dette viser også hvorfor mennesker kan ha utviklet seg en intuisjon om å straffe folk. Å være en som straffer de som bryter med felleskapets regler, er riktig strategi fra et selvsentrert perspektiv. På sikt vil mennesker som straffer få det bedre enn de som ikke straffer – for eksempel med å ikke bli frastjålet maten, eller å ikke måtte gjøre jobben til andre. Mennesker med bedre (lengre, sunnere, mindre konfliktfullte osv.) liv vil ha større mulighet til å videreføre genene sine enn andre. Følgelig vil gener som predisponerer til å være en som straffer oftere bli videreført. Evolusjonen går sin gang, og den intuitive reaksjonen av å ville straffe folk som bryter med felleskapet (eller forskjellige normer og regler som er ment for felleskapets beste) blir normalen.

5.4.2 Fra ris og ros til fri vilje

Når intuisjonen om straff er på plass, er de resterende stegene til opplevelsen av å noen ganger handle *helt på egenhånd* rimelig rett frem. Av grunnene beskrevet over vil man for eksempel ikke straffe folk for annet enn handlinger som fremstår som deres egen. Om noen blir tvunget til å gjøre en forbrytelse (for eksempel truet til å stjele mat), trenger ikke dette bryte med tillitten til vedkommende. Siden hen åpenbart ble påvirket utenfra da hen tok det egoistiske valget, kan man anta at hen vil prioritere felleskapet igjen i neste dilemma - gitt at hen ikke lenger blir tvunget. Siden strategien bygger på en gjensidighet, er det også en annen selvsentrert interesse for å unngå straff av slike handlinger: Man vil ikke risikere å selv bli straffet for handlinger man var tvunget til å utføre. Som Joshua Greene påpeker, gjelder samme mekanisme åpenbare uhell (Greene, 2013, s. 270). Utilsiktede «forbrytelser» vil man kun straffe om de åpenbart stammer fra uaktsomhet – siden dette tyder på at vedkommende kan være uaktsom igjen (og man kan selv unngå straff med å ikke opptre uaktsomt. Å unngå alt av uflaks ville derimot vært vanskelig).

Man ser riktignok ofte indikasjoner på at menneskelig handling til syvende og sist kan forklares av forhold utenfor vår kontroll. Men, på grunn av dets nytte har man også en sterk intuisjon av å ville straffe folk, og at straff må følge av kontrollert handling. Man kan ikke holde på de to oppfatningene samtidig, og ender opp med å beholde intuisjonen om selvstendig handling, både hos andre aktører og seg selv.

Det er altså mulig å gi mer plausible forklaringer på Richard Taylors to datapunkter enn hans egen teori. Både determinisme og Kanes variant av fri vilje fremstår som et mer akseptabelt alternativ. Ingen av de to synene innebærer at aktørene er moralsk

ansvarlige, dypest sett. Følgelig bør fortjeneste ikke spille en rolle i en rettferdig ressursfordeling, til tross for sin intuitive appell. Konseptet hviler på svært usannsynlige forutsetninger, og bør derfor ikke inngå i en teori om fordelingsrettferdighet.

Diskusjonen over viser likevel at belønning (eller straff) kan ha en viktig funksjon, da det kan få mennesker til å opptre på nyttige måter for samfunnet. Det finnes altså et konsekvensialistisk, gjerne utilitaristisk, rasjonale for både ris og ros, selv i en verden der man er klar over at mennesker ikke er moralsk ansvarlige. Å stimulere arbeid med skattenivåer, pensjonsordninger og kan altså være på sin plass. Dette gjelder imidlertid kun i den grad det bidrar til mer aggregert velferd. I slike tilfeller fordrer utilitarismen denne fordelingen. At mennesker får godt (eller dårlig) betalt som følge av valgene de tar, må altså være fordi en slik ordning har gode konsekvenser. Det er urettferdig å gjøre en tilsvarende fordeling basert på en tanke om fortjeneste.

6 Konklusjon

I litteraturen om fordelingsrettferdighet blir utilitarismen ofte oversett, eller avvist som en lite plausibel tilnærming. I denne oppgaven har jeg argumentert for at teorien bør tilbake i varmen. Utilitarismen samsvarer for det første med mange av våre oppfatninger om fordelingsrettferdighet. Oppfatningene som tilsynelatende utgjør problemer for utilitarismen er enten forenlige med en mer sofistikert utilitarisme, eller kan forkastes som lite troverdige.

Som jeg forsvarer i kapittel 4, er utilitarismen mer plausibel enn andre veletablerte teorier om fordelingsrettferdighet. Den er et mer rasjonelt valg fra bak John Rawls' uvitenhetslør enn Rawls' foretrukne maximin-kriterium, som er i overkant risikominimerende. Egalitarianismen rammes av *innvendingen om utjevning nedover* og bør derfor avvises. Egalitarianismen tilsier at det er positivt, i alle fall i én forstand, om de godt stilte får det dårligere, uten at noen får det bedre. Det kan forsvares om det er penger eller andre ressurser som fordeles. Når det er velferd som fordeles, er ikke dette akseptabelt for en teori om rettferdig fordeling. Prioritarianismen og teorier om tilstrekkelighet unngår denne innvendingen. I mange tilfeller er deres implikasjoner de samme som utilitarismens. De gangene prinsippene gir ulike svar, stemmer imidlertid utilitarismen bedre overens med våre oppfatninger enn konkurrentene. Prioritarianismen og tilstrekkelighetsteoriens intuitive appell kan i

tillegg forklares av menneskets tendens til å blande velferd og ressurser. Utilitarismen er derfor å foretrekke.

Rawls lar ikke fortjeneste spille en rolle i sin rettferdighetsteori. De andre prinsippene adresserer ikke temaet. Det til tross for at fortjeneste, som Mill påpeker, er blant de vanligste konseptene i befolkningens rettferdighetsoppfatning. Både skoleansatte og helsearbeidere som streiker for høyere lønn, samt næringslivsledere som legitimerer sin store formue, peker på at de fortjener visse goder, basert på riktige avgjørelser, hardt arbeid og samfunnsnytten de bidrar til.

I kapittel 5 undersøkte jeg derfor om fortjeneste bør være en faktor i en teori om fordelingsrettferdighet. Jeg viser at fortjeneste forutsetter moralsk ansvar og at moralsk ansvar hviler på lite sannsynlige forutsetninger om menneskers frie vilje. Følgelig bør man ikke få mer eller mindre ressurser basert på fortjeneste, utover det som er nødvendig for å fremme velferd.

Reflektert likevekt-metoden brukes sjelden for å understøtte utilitarismen. Det er til og med blitt påstått at teorien ikke kan oppnå berettigelse på denne måten (Tersman, 1991, s. 398). I denne oppgaven har jeg imidlertid argumentert for at det er mulig. Videre forskning bør undersøke hvilke andre problemstillinger enn fordelingsrettferdighet reflektert likevekt vil tilsi at utilitarismen er den foretrukne teorien. I forsvaret av utilitarismen har jeg tatt for meg mange av de hyppigst brukte innvendingene mot teorien. Flere av resonnementene i teksten kan være relevante på andre arenaer, for eksempel for utilitarismens holdbarhet som prinsipp for enkeltmenneskers handlinger. Tilsvarende er også kapittelet om fortjeneste og moralsk ansvar relevant for flere problemstillinger enn de som angår fordelingsrettferdighet. Et spesielt nærliggende tema er juridisk rettferdighet. Om konseptet gjengjeldelse mister sin appell i fraværet av moralsk ansvar, kan dette ha viktige implikasjoner for vårt syn på straff og ansvar innenfor den juridiske sfæren.

Samtidig er det også begrensninger ved min oppgave. Avvisningen av konseptet fortjeneste hviler på den metafysiske diskusjonen angående moralsk ansvar og fri vilje. Jeg har ikke kunnet ta for meg alle teorier innenfor fri vilje-debatten. Om konklusjonen i denne delen skulle vise seg å være feilaktig, kan dette true utilitarismens posisjon sammenlignet med prinsipper som baserer seg på fortjeneste. I tillegg har jeg ikke vurdert om utilitarismen bør fordre gjennomsnittlig eller total

velferdsmaksimering. Siden fordeling av goder trolig påvirker fremtidige befolkningstall, er dette et relevant område for fordelingsrettferdighet som er utelatt fra oppgaven, og som senere forskning bør ta for seg.

En annen begrensning ved oppgaven er at diskusjonen utelukkende har foregått på et prinsipielt nivå. Utilitarismens implikasjoner og praktiske anvendbarhet som fordelingsprinsipp bør undersøkes nærmere. Dette gjøres ved å kartlegge hvilke valgmuligheter som er åpne foran politikere (og andre som fordeler ressurser mellom folk), altså f.eks. undersøke hvor betydelige omfordelinger velgere er åpne for. I tillegg bør det gjøres mer forskning på hvordan vi kan rangere alternative fordelinger. Vi har ikke direkte tilgang til folks velferdsnivå og vi må derfor bruke mer eller mindre rimelige proxyer for å måle velferd. Jo sikrere antakelser man har angående mennesker velferdsnivåer, desto mer anvendbar er utilitarismen. Å gjøre teorien praktisk implementerbar er viktig. Bare slik får man et verktøy til å fordele ressurser på en rettferdig måte. Et fruktbart spor er den gryende forskningen på hvordan kostnytte-analyser kan ta mer direkte hensyn til folks velferdsnivå, blant annet ved hjelp av sosiale velferdsfunksjoner (for foreløpig forskning på dette, se f.eks. Adler, 2019).

Jeg har sammenliknet utilitarismen med de mest plausible alternativene, og tatt for meg de viktigste innvendingene mot prinsippet. Utilitarismen fremstår som den beste teorien om fordelingsrettferdighet. Ressurser bør derfor fordeles på måten som antas å maksimere mengden velferd. Alt annet ville vært urettferdig.

7 Litteraturliste

Adcock, R., & Collier, D. (2001). Measurement Validity: A Shared Standard for Qualitative and Quantitative Research. *American Political Science Review*, *95*(3), 529–546. https://doi.org/10.1017/S0003055401003100

Adler, M. D. (2019). *Measuring social welfare: An introduction*. Oxford University Press.

Anderson, E. S. (1999). What Is the Point of Equality? *Ethics*, *109*(2), 287–337. https://doi.org/10.1086/233897

Aristotle. (1934). *Physics* (P. H. Wicksteed & F. M. Cornford, Overs.) [Data set]. Harvard University Press. https://doi.org/10.4159/DLCL.aristotle-physics.1957

Arneson, R. J. (2013). Egalitarianism. I E. N. Zalta (Red.), *The Stanford Encyclopedia of Philosophy (Summer 2013 Edition)*. https://plato.stanford.edu/archives/sum2013/entries/egalitarianism/

Barry, B. (2012). Equal Opportunity and Moral Arbitrariness. I J. Lamont (Red.), *Distributive Justice*. Ashgate.

Bentham, J. (1789). *An introduction to the principles of morals and legislation. Printed in the year 1780, and now first published. By Jeremy Bentham, of Lincoln's Inn, Esquire*. printed for T. Payne, and Son, at the Mews Gate; Eighteenth Century Collections Online. https://link-gale-com.ezproxy.uio.no/apps/doc/CW0104763964/ECCO?u=oslo&sid=ECCO&xid=9f3c67e1&pg=1

Bentham, J. (1966). Appendix IV. I D. Baumgardt, *Bentham and the Ethics of Today* (s. 554–566). Octagon Books Inc.

Bentham, J. (1977). A Fragment on Government. I J. H. Burns & H. L. A. Hart (Red.), *A Comment on the Commentaries and A Fragment on Government* (s. 393–501). The Athlone Press.

Berlin, I. (2002). Two Concepts of Liberty. I H. Hardy & I. Harris (Red.), *Liberty: Incorporating four essays on liberty* (s. 166–217). Oxford University Press.

Broome, J. (1991). *Weighing Goods*. Blackwell.

Broome, J. (2002). *Respects and Levelling Down*. http://users.ox.ac.uk/~sfop0060/pdf/respects%20and%20levelling%20down.pdf

Burns, J. H. (2005). Happiness and Utility: Jeremy Bentham's Equation. *Utilitas*, *17*(1), 46–61. https://doi.org/10.1017/S0953820804001396

Casal, P. (2007). Why Sufficiency Is Not Enough. *Ethics*, *117*(2), 296–326. https://doi.org/10.1086/510692

Churchland, P. S. (2019). *Conscience: The origins of moral intuition* (First edition). W. W. Norton & Company.

Cohen, G. A. (1989). On the Currency of Egalitarian Justice. *Ethics*, *99*(1), 906–944.

Cohen, G. A. (2001). *Karl Marx's theory of history: A defence* (7. print., 1. expanded ed). Princeton Univ. Press.

Daniels, N. (1979). Wide Reflective Equilibrium and Theory Acceptance in Ethics. *The Journal of Philosophy*, *76*(5), 256–282. https://doi.org/10.2307/2025881

Daniels, N. (1980). On Some Methods of Ethics and Linguistics. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, *37*(1), 21–36.

de Lazari-Radek, K., & Singer, P. (2010). Secrecy in consequentialism: A defence of esoteric morality. *Ratio*, *23*(1), 34–58. https://doi.org/10.1111/j.1467-9329.2009.00449.x

Descartes, R. (2003). Discourse on method. I E. S. Haldane & G. R. T. Ross (Overs.), *Discourse on method and Meditations* (s. 1–52). Dover Publications.

Dworkin, R. (1981). What is Equality? Part 2: Equality of Resources. *Philosophy & Public Affairs*, *10*(4), 283–345. https://doi.org/10.4324/9781315199795-7

Elgin, C. (1996). *Considered Judgement*. Princeton University Press.

Eyal, N. M., Norheim, O. F., Hurst, S. A., & Wikler, D. (Red.). (2013). *Inequalities in health: Concepts, measures, and ethics*. Oxford University Press.

Fellner, W. (1965). *Probability and Profit: A Study of Economic Behavior Along Bayesian Lines*. Richard D. Irwin, Inc.

Fischer, J. M., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge University Press.

Folger, R. (1987). Distributive and procedural justice in the workplace. *Social Justice Research*, *1*(2), 143–159. https://doi.org/10.1007/BF01048013

Frankfurt, H. G. (1969). Alternate Possibilities and Moral Responsibility. *The Journal of Philosophy*, *66*(23), 829–839.

Fried, H. B. (2020). *Facing Up to Scarcity: The Logic and Limits of Nonconsequentialist Thought*. Oxford University Press.

Greene, J. (2013). *Moral tribes: Emotion, reason, and the gap between us and them*. Penguin Press.

Greene, J. (2014). Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics. *Ethics*, *124*(4), 695–726. https://doi.org/10.1086/675875

Greene, J. (2017). The rat-a-gorical imperative: Moral intuition and the limits of affective learning. *Cognition*, *167*, 66–77. https://doi.org/10.1016/j.cognition.2017.03.004

Greene, J., & Baron, J. (2001). Intuitions about declining marginal utility. *Journal of Behavioral Decision Making*, *14*(3), 243–255. https://doi.org/10.1002/bdm.375

Haddad, B. M. (2005). Ranking the adaptive capacity of nations to climate change when socio-political goals are explicit. *Global Environmental Change*, *15*(2), 165–176. https://doi.org/10.1016/j.gloenvcha.2004.10.002

Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment. *Psychological Review*, *108*(4), 814–834.

Hare, R. M. (1979). What is Wrong with Slavery. *Philosophy & Public Affairs*, *8*(2), 103–121.

Hare, R. M. (1982). Ethical theory and utilitarianism. I A. Sen & B. Williams (Red.), *Utilitarianism and Beyond* (1. utg., s. 23–38). Cambridge University Press. https://doi.org/10.1017/CBO9780511611964.003

Hare, R. M. (2003). *Freedom and reason* (Reprint). Clarendon Press.

Harsanyi, J. C. (1953). Cardinal Utility in Welfare Economics and in the Theory of Risk-taking. *Journal of Political Economy*, *61*(5), 434–435.

Harsanyi, J. C. (1955). Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility. *Journal of Political Economy*, *63*(4), 309–321.

Harsanyi, J. C. (1975). Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory. *The American Political Science Review*, *69*(2), 594–606.

Hirose, I. (2015). *Egalitarianism*. Routledge.

Hoefer, C. (2016). Causal Determinism. I E. N. Zalta (Red.), *The Stanford Encyclopedia of Philosophy (Spring 2016 Edition)*. https://plato.stanford.edu/archives/spr2016/entries/determinism-causal/

Hovi, J. (2008). *Spillteori: En Innføring*. Universitetsforlaget.

Huby, P. (1967). The First Discovery of the Freewill Problem. *Philosophy*, *42*(162), 353–362. https://doi.org/10.1017/S0031819100001534

Huemer, M. (2007). Compassionate Phenomenal Conservatism. *Philosophy and Phenomenological Research*, *74*(1), 30–55. https://doi.org/10.1111/j.1933-1592.2007.00002.x

Huseby, R. (2010). Sufficiency: Restated and Defended. *Journal of Political Philosophy*, *18*(2), 178–197. https://doi.org/10.1111/j.1467-9760.2009.00338.x

Huseby, R. (2019). Sufficientarianism. I *Oxford Research Encyclopedia of Politics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228637.013.1382

Kahneman, D. (2012). *Thinking, Fast and Slow*. Penguin Books.

Kane, R. (1989). Two Kinds of Incompatibilism. *Philosophy and Phenomenological Research*, *50*(2), 219–254. https://doi.org/10.2307/2107958

Kane, R. (1999). Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism. *The Journal of Philosophy*, *96*(5), 217–240. https://doi.org/10.2307/2564666

Kant, I. (1970). Grunnlegging til moralens metafysikk. I E. Storheim (Overs.), *Morallov og frihet: Moralfilosofiske skrifter* (s. 3–76). Gyldendal. https://urn.nb.no/URN:NBN:no-nb_digibok_2009062401028

Keynes, J. M. (1952). *A Treatise on Probability*. Macmillan and co., limited.

Kirkeby, E. (2009). *Against Moral Intuitions—Peter Singer's Arguments Against The Use Of Moral Intuitions In Moral Methodology* [Master's thesis, University of Oslo]. https://www.duo.uio.no/bitstream/handle/10852/24997/Against_Moral_Intuitions.pdf?sequence=1&isAllowed=y

Kymlicka, W. (2002). *Contemporary political philosophy: An introduction* (2nd ed). Oxford University Press.

Laplace, P.-S. (1902). *A Philosophical Essay on Probabilities* (F. W. Truscott & F. L. Emory, Overs.). Wiley.

Leibniz, G. W. von. (1902). Monadology. I P. Janet (Red.), & G. Montgomery (Overs.), *Discourse on metaphysics ; Correspondence with Arnauld ; Monadology* (s. 251–272). Open Court.

Lewis, D. K. (1986). *Philosophical papers* (Revised for 2nd edition). Oxford University Press.

Lewis, D. K. (1989). The Punishment that Leaves Something to Chance. *Philosophy & Public Affairs*, *18*(1), 53–67.

Lucretius. (1924). *De Rerum Natura* (M. F. Smith, Red.; W. H. D. Rouse, Overs.) [Data set]. Harvard University Press. https://doi.org/10.4159/DLCL.lucretius-de_rerum_natura.1924

Lycan, W. (2019). *On Evidence in Philosophy*. Oxford University Press.

Mackie, J. L. (1965). Causes and Conditions. *American Philosophical Quarterly*, *2*(4), 245–264.

Marx, K. (1976). Wage-Labour and Capital. I K. Marx, *Wage-labour and capital & Value, price, and profit* (1st paperback (combined) ed, s. 15–48). International Publishers.

McFarlin, D. B., & Sweeney, P. D. (1992). Research Notes. Distributive and Procedural Justice as Predictors of Satisfaction with Personal and Organizational Outcomes. *Academy of Management Journal*, *35*(3), 626–637. https://doi.org/10.5465/256489

Milgram, S. (1963). Behavioral Study of obedience. *The Journal of Abnormal and Social Psychology*, *67*(4), 371–378. https://doi.org/10.1037/h0040525

Mill, J. S. (1863). *Utilitarianism*. Parker, Son and Bourn, West Strand.

Miller, D. (2017). Justice. I E. N. Zalta, *The Stanford Encyclopedia of Philosophy (Fall 2017 Edition)*. https://plato.stanford.edu/archives/fall2017/entries/justice

Mulgan, T. (2011). *Ethics for a Broken World: Imagining Philosophy Aft er Catastrophe*. Acumen Publishing Limited.

Nagel, T. (1976). Moral Luck. *Proceedings of the Aristotelian Society, Supplementary Volumes*, *50*, 115–151.

Nagel, T. (1986). *The view from nowhere*. Oxford University Press.

Nelkin, D. K. (2019). Moral Luck. I E. N. Zalta (Red.), *The Stanford Encyclopedia of Philosophy (Summer 2019 Edition)*. https://plato.stanford.edu/archives/sum2019/entries/moral-luck/

Neuhouser, F. (2008). *Rousseau's Theodicy of Self-Love*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199542673.001.0001

Nozick, R. (1974). *Anarchy, state, and utopia*. Blackwell.

Nussbaum, M. (2003). Capabilities as fundamental entitlements: Sen and social justice. *Feminist Economics*, *9*(2–3), 33–59. https://doi.org/10.1080/1354570022000077926

Nussbaum, M. C. (2000). *Women and human development: The capabilities approach*. Cambridge University Press.

O'Connor, T., & Franklin, C. (2021). Free Will. I E. N. Zalta (Red.), *The Stanford Encyclopedia of Philosophy (Spring 2021 Edition)*. https://plato.stanford.edu/archives/spr2021/entries/freewill/

Parfit, D. (1984). *Reasons and Persons*. Clarendon.

Parfit, D. (1991). *Equality or Priority*. The Lindley Lecture, The university of Kansas, Lawrence, Kansas. http://www.stafforini.com/docs/Parfit%20-%20Equality%20or%20priority.pdf

Parfit, D. (2012). Another Defence of the Priority View. *Utilitas*, *24*(3), 399–440. https://doi.org/10.1017/S095382081200009X

Plomin, R., DeFries, J. C., Knopik, V. S., & Neiderhiser, J. M. (2016). Top 10 Replicated Findings From Behavioral Genetics. *Perspectives on Psychological Science*, *11*(1), 3–23. https://doi.org/10.1177/1745691615617439

Popper, K. R. (2002). *The logic of scientific discovery*. Routledge.

Powers, M., & Faden, R. R. (2006). *Social justice: The moral foundations of public health and health policy*. Oxford University Press.

Quine, W. V. (1981). *Theories and things*. Harvard University Press.

Rawls, J. (1971). *A Theory of Justice* (Online reprint.). The Belknap Press of Harvard University Press.

Rawls, J. (1999). *A Theory of justice—Revised edition*. Oxford University Press.

Raz, J. (1986). *The morality of freedom* (Reprinted). Clarendon Press.

Robeyns, I., & Byskov, M. F. (2020). The Capability Approach. I E. N. Zalta (Red.), *The Stanford Encyclopedia of Philosophy (Winter 2020 Edition)*. https://plato.stanford.edu/entries/capability-approach/

Rousseau, J.-J. (2011a). Discourse on the Origin of Inequality. I D. A. Cress (Red. & Overs.), *The Basic Political Writings* (Second edition, s. 27–92). Hackett Publishing Company.

Rousseau, J.-J. (2011b). On the Social Contract. I D. A. Cress (Overs.), *The Basic Political Writings* (Second edition, s. 153–253). Hackett Gazelle distributor.

Sandel, M. J. (2009a). *Justice: What's the right thing to do?* (1st ed). Farrar, Straus and Giroux.

Sandel, M. J. (2009b). *Justice: What's the Right Thing to Do?* Farrar, Straus and Giroux.

Scanlon, T. M. (2000). *What We Owe to Each Other* (Fourth printing). The Belknap Press of Harvard University Press.

Scheffler, S. (1988). Introduction. I S. Scheffler (Red.), *Consequentialism and its Critics* (s. 1–13). Oxford Univ. Press.

Sen, A. (1970). *Collective Choice and Social Welfare*. Holden-day, Inc.

Sen, A. (1979). *Utilitarianism and Welfarism*. *76*(9), 463–489.

Sen, A. (2006). What Do We Want from a Theory of Justice? *The Journal of Philosophy*, *103*(5), 215–238.

Shaver, R. (2004). The Appeal of Utilitarianism. *Utilitas*, *16*(3), 235–250. https://doi.org/10.1017/S0953820804001153

Sidgwick, H. (1981a). *The Methods of Ethics* (7th edition, Hackett reprint). Hackett Publishing Company.

Sidgwick, H. (1981b). *The Methods of Ethics*. Hackett Publishing Company.

Singer, P. (1974). Sidgwick and Reflective Equilibrium: *The Monist*, *58*(3), 490–517. https://doi.org/10.5840/monist197458330

Singer, P. (1995). *Animal liberation* (2nd ed., with a new preface by the author). Pimlico.

Singer, P. (2005). Ethics and Intuitions. *The Journal of Ethics*, *9*(3–4), 331–352. https://doi.org/10.1007/s10892-005-3508-y

Singer, P. (2011). *Practical Ethics* (Third edition). Cambridge University Press.

Smart, J. J. C. (1973). An outline of a system of utilitarian ethics. I J. J. C. Smart & B. Williams, *Utilitarianism: For and against* (s. 3–74). University Press.

Spinoza, B. D. (1954a). Ethics—Part I. Concerning God. I J. Gutmann (Red.), *Ethics—Preceded by On the Improvement of the Understanding* (s. 41–78). Hafner Publishing Company.

Spinoza, B. D. (1954b). Ethics—Part II. Of the nature and origin of the mind. I J. Gutmann (Red.), *Ethics—Preceded by On the Improvement of the Understanding* (s. 79–126). Hafner Publishing Company.

Strawson, G. (1994). The impossibility of moral responsibility. *Philosophical Studies*, *75*(1–2), 5–24. https://doi.org/10.1007/BF00989879

Strawson, P. F. (2008). *Freedom and resentment and other essays*. Routledge.

Tappolet, C. (2013). Evaluative vs. Deontic Concepts. I H. Lafollette (Red.), *International Encyclopedia of Ethics* (s. 1791-1799 (numbered 1-9)). Blackwell Publishing Ltd. https://doi.org/10.1002/9781444367072.wbiee118

Taylor, R. (1992). *Metaphysics* (4th ed). Prentice Hall.

Temkin, L. S. (1993). *Inequality*. Oxford University Press.

Temkin, L. S. (2000). Equality, Priority, and the Levelling-Down Objection. I M. Clayton & A. Williams (Red.), *The Ideal of Equality* (s. 126–161). Palgrave Macmillan uk. https://www.law.upenn.edu/institutes/cerl/conferences/prioritarianism_papers/Session2Temkin2.pdf

Temkin, L. S. (2003). Egalitarianism Defended. *Ethics*, *113*(4), 764–782. https://doi.org/10.1086/373955

Tersman, F. (1991). Utilitarianism and the idea of reflective equilibrium. *The Southern Journal of Philosophy*, *29*(3), 395–406.

Trochim, W. M. K., Donnelly, J. P., & Arora, K. (2016). *Research methods: The essential knowledge base* (Student ed). Cengage Learning.

Tännsjö, T. (2020). Why Derek Parfit had reasons to accept the Repugnant Conclusion. *Utilitas*, *32*(4), 387–397. https://doi.org/10.1017/S0953820820000102

Valentini, L. (2012). Ideal vs. Non-ideal Theory: A Conceptual Map: Ideal vs Non-ideal Theory. *Philosophy Compass*, *7*(9), 654–664. https://doi.org/10.1111/j.1747-9991.2012.00500.x

van Inwagen, P. (1998). The Mystery of Metaphysical Freedom. I P. van Inwagen & D. W. Zimmerman (Red.), *Metaphysics: The big questions* (s. 365–374). Blackwell Publishers.

Washington, B. T. (1986). *Up from slavery*. Penguin Books.

Wegner, D. M. (2004). Précis of The illusion of conscious will. *Behavioral and Brain Sciences*, *27*(5), 649–659. https://doi.org/10.1017/S0140525X04000159

Wolff, J. (2006). *An introduction to political philosophy* (Rev. ed). Oxford University Press.

Woodard, C. (2019). *Taking utilitarianism seriously*. Oxford University Press.