UiT
NORGES ARKTISKE UNIVERSITET

Institutt for filosofi og førstesemesterstudier

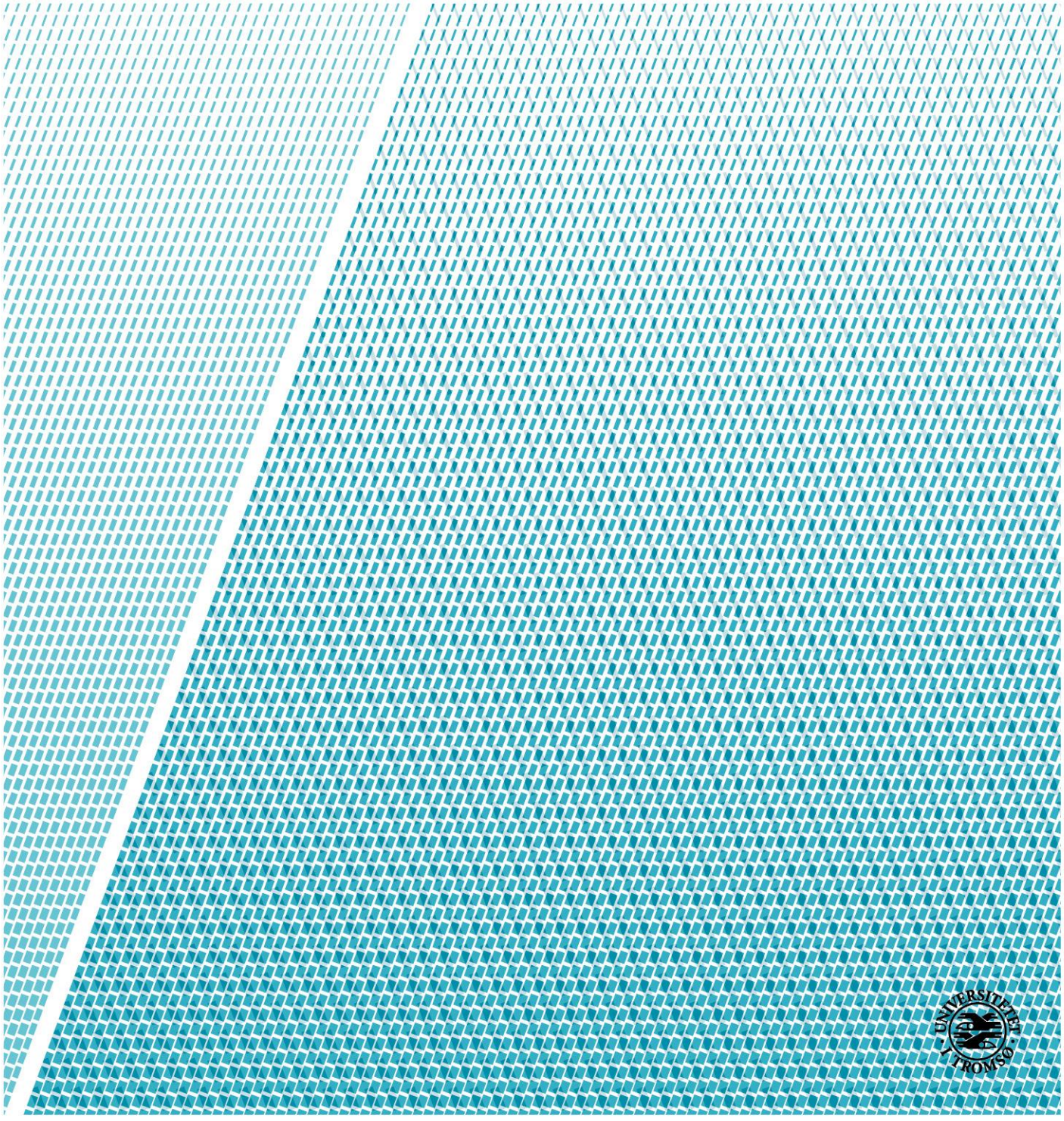
Vitenskap og moralfilosofi

Et forsvar av Joshua Greenes kritikk av deontologisk etikk

—

Torjus Lorentsen

Masteroppgave i filosofi – mai 2018



Forord

Takk til venner og familie for støtte under arbeidet med denne oppgaven.

En særlig takk til min veileder, Ivar Russøy Labukt, for uvurderlig veiledning gjennom denne prosessen.

Torjus Lorentsen

Vannareid, 11.5.2018

Innhold

Kapittel 1: Innledning

Hovedmålet med denne oppgaven vil være å argumentere for at Joshua Greenes prosjekt har potensial til å representere et sårt tiltrengt framskritt innen normativ etikk som disiplin. Gjennom en serie publikasjoner de siste to tiårene har Greene utviklet et høyst originalt argument hvor han ved hjelp av empiriske metoder kritiserer deontologisk etikk ved å vise til svakheter ved de nevrologiske prosessene han mener ligger bak den deontologien som er operativ innen normativ etikk i dag. Sånn sett inngår denne oppgaven i to forskjellige debatter. For det første den klassiske debatten mellom deontologi og konsekvensialisme, og for det andre den mer spesifikke debatten rundt Greenes prosjekt. Før jeg går nærmere inn på detaljene i Greenes argument, og deretter de utfordringene det står overfor, skal jeg derimot si noe om hvorfor jeg karakteriserer normativ etikk som en disiplin preget av stagnasjon, og hvordan Greenes prosjekt kan bøte på denne problematikken. Dette vil aktualisere enkelte kritiske momenter ved Greenes prosjekt som jeg vil undersøke nærmere i denne oppgaven. Jeg avslutter denne innledningen med å presentere disse samt redegjøre for oppbygningen av resten av oppgaven.

1.1. Stillstand

Normativ etikk handler om å gi en teoretisk begrunnelse på spørsmålet om hvordan vi bør handle i et moralsk henseende. Når det skal gis en teoretisk begrunnelse på hvordan man bør handle er konsekvensialisme og deontologi de uten sidestykke vanligste forklaringsmodellene. Ikke bare er denne dikotomien hovedinnholdet i etikk-kurs på universiteter over hele verden, men det er også bakgrunnen for mange av de distinksjonene som diskuteres innen normativ etikk, når man ikke diskuterer dikotomien i seg selv. Kjært barn har mange navn, og termer som konsekvensialisme og deontologi, og pliktetikk og utilitarisme, brukes ofte om hverandre. Mer presist forstås termene slik at konsekvensialisme og deontologi inneholder henholdsvis klassisk utilitarisme og klassisk pliktetikk – men ikke omvendt. En praktisk måte å forstå deontologi og konsekvensialisme, og især deres forhold til hverandre, er å forstå konsekvensialismen som å hevde at den riktige handlingen alltid er den som maksimerer det som defineres som å være godt. Deontologi kan da forstås som benektelsen av denne påstanden[1]: Det er ikke slik at det alltid er riktig å gjøre det som

[1] Kamm 2016: 11

maksimerer det som defineres som å være godt. Det vil si at det for deontologien må finnes minimum ett unntak til regelen om nyttemaksimering, men ikke at konsekvensene ikke er relevante i det hele tatt. Man kan derfor si at en teori som hevder at det er riktig å gjøre det som maksimerer det som defineres som godt, bortsett fra når det er fullmåne, er en deontologisk teori, selv om det vil være en utrolig dårlig teori. De deontologiske unntakene til nyttemaksimering formuleres ofte i form av plikter og rettigheter, uten referanse til månesyklusen. Dette er fortsatt helt i tråd med den definisjonen jeg opererer med her, siden det fremhever at deontologien har andre verdier enn bare det å maksimere det som defineres som godt. Samtidig er denne definisjonen gunstig fordi den på en åpenbar måte inkluderer andre ikke-konsekvensialistiske teorier som tradisjonelt sett ikke ville blitt sett på som deontologiske. Dette gjør at den opprinnelige dikotomien mellom konsekvensialisme og deontologi blir total. Hvis normativ etikk dreier seg om å gi en teoretisk begrunnelse på spørsmålet om hvordan vi bør handle, så finnes det to alternativer: deontologi og konsekvensialisme.

Med denne definisjonen blir vannskille veldig klart, og det finnes ingen mellomvei. Deontologi utelukker konsekvensialisme og konsekvensialisme utelukker deontologi. Siden deontologi og konsekvensialisme slik jeg definerte det i foregående avsnitt *er* normativ etikk, så betyr det at framskritt innen normativ etikk som disiplin også må innebære framgang innen disputten mellom deontologi og konsekvensialisme.

Det er særlig én ting som taler til fordel for konsekvensialistisk teori, nemlig dens enkelhet. Ut fra enkle antakelser om hva som er godt og at det som er godt bør maksimeres så kan konsekvensialismen i teorien, og med det menes det at alle relevante ikke-moralske fakta er kjent, avgi en konkret moralsk dom i ethvert tenkelig moralsk dilemma. Om man som de klassiske utilitaristene antar at det som er godt er fraværet av smerte og tilstedeværelsen av nytelse,[2] så vet man at det som produserer mest mulig netto nytelse til enhver tid er riktig å gjøre. Denne metoden har selvfølgelig blitt kritisert, og det er særlig én ting som fremfor noe har vist seg å være problematisk. I mange tilfeller – både tenkelige og mer abstrakte – viser det seg at konsekvensialismen anbefaler handlinger som virker å være *fryktelig, fryktelig gale*. Selv om konsekvensialismen selvfølgelig kan gjøre grep for å unngå mange av de

[2] Bentham 2000: 14

kontraintuitive resultatene, så betyr det igjen å gi opp noe av den enkelheten som gjør teorien så appellerende i utgangspunktet. Gjennom den nyere filosofihistorien har derfor anklagen om kontraintuitivitet vært en konstant torn i siden for konsekvensialismen.

Ifølge Peter Singer står Fjodor Dostojevskij i *Brødrene Karamasov* for en av de tidligste anklagene om kontraintuitivitet[3]:

> «*tenk deg at du skulle oppføre en bygning som kunne bringe menneskeheten endelig lykke, som endelig kunne gi menneskene fred og ro, men at vilkåret var at en liten ubetydelig skapning måtte utsettes for tortur, for eksempel dette lille barnet som slo seg med neven for brystet, for at denne bygningen skulle bygges på dette barnets uhevnede tårer, ville du da påta deg oppgaven som byggmester? Jeg vil ha et oppriktig svar!*»[4]

Robert Nozicks *Utility Monster* er et annet tankeeksperiment som gyver løs på konsekvensialismen: Om man ser for seg et vesen som er oss overlegen i å konvertere ressurser om til nytte, så får dette kontraintuitive konsekvenser. Jeg oppnår kanskje bare én nytteenhet av å spise et kakestykke, mens dette vesenet oppnår hundre nytteenheter av det samme kakestykket. Om det dernest legges en konsekvensialistisk kost-/nytteanalyse til grunn så virker det klart at alle ressurser bør brukes på Nozicks *utility monster*.[5] På et lignende vis har John Rawls vist til et hypotetisk samfunn hvor den totale nytten maksimeres ved at en majoritet slavebinder en minoritet.[6] I sin introduksjonsbok til etikk, *Justice*, presenterer Michael Sandel under kapitlet om utilitarisme flere eksempler på anklagen om kontraintuitivitet. Et av eksemplene har en lignende struktur som Rawls eksempel: I det antikke Rom ofret de til massenes store begeistring kristne til løvene på Coloseum. Visst er det enorme smerter involvert for den kristne som rives i stykker av sultne løver. Men igjen må man tenke på den kollektive gleden den ekstatiske folkemengden opplever. Om dette maksimerer den totale mengden nytelse så har ikke utilitarismen noe den skulle ha sagt, og er tvunget til å bifalle handlingen. Og som Sandel påpeker, det må det vel være noe fryktelig galt med?[7]

[3] Singer 2005: 343
[4] Dostojevskij 1993: 313
[5] Nozick 1974: 41
[6] Rawls 1999: 137
[7] Sandel 2009: 24

Dette er kun noen få eksempel på forskjellige scenarier som tilsynelatende vil bifalles av konsekvensialistisk normativ teori, men samtidig virker å være forferdelig galt. På dette tidspunktet er ikke poenget å vurdere om dette er gode eller dårlige argument. Poenget er snarere å fremheve det repetitive i det hele. Dostojevskijs roman utkom i 1880. Rawls og Nozicks verker ble publisert i henholdsvis 1971 og 1974. Sandels bok kom ut i 2009. Det er med andre ord en klar tendens at denne typen moteksempel går igjen, helt fra slutten av attenhundretallet og frem til i dag. Denne type moteksempel fremsettes med like stor selvfølgelighet i 2009 som i 1971 – og denne perioden er utvilsomt den mest produktive perioden i filosofihistorien, også hva normativ etikk angår. Dette tyder altså på at den mest sentrale debatten innen normativ etikk er en debatt blottet for utvikling og innovasjon.

Slike moteksempel er selvfølgelig ment å rokke ved plausibiliteten til konsekvensialistiske teorier, og på den måten framheve deontologisk teori. De er i så måte bare halve historien når det gjelder debatten mellom deontologi og konsekvensialisme. En mer helhetlig representasjon av denne debatten er det som i dag er kjent som *the trolley problem*. *The trolley problem* fikk sitt navn av Judith Jarvis Thomson i 1976, som da brukte det som navn på to konkrete scenarier som ble presentert av Phillipa Foot i hennes tekst fra 1967, *The problem of abortion and the doctrine of double effect*. Foot ser for seg en person som styrer en jernbanevogn ute av kontroll. Selv om personen ikke kan stoppe vognen, så kan han styre den over fra et spor til et annet. På det ene sporet er det fem arbeidere som vil bli drept, mens på det andre sporet er det bare en arbeider som vil bli drept. I et annet scenario ser hun for seg en dommer i en liten by, hvor en forarget folkemengde krever at en angivelig forbryter stilles til ansvar for sine ugjerninger. Problemet er at dommeren ikke vet hvem forbryteren er, og den eneste måten å forhindre at den sinte folkemengden gjør opprør og dreper fem personer, er ved å dømme og følgelig henrette en uskyldig person for ugjerningene. Spørsmålet er, ifølge Foot, hvorfor vi uten å nøle skal tillate føreren av jernbanevognen å drepe én person heller enn fem personer, men ikke tillate dommeren å gjøre et tilsvarende offeret.[8]

[8] Foot 1967: 3

Et annet eksempel som også tas opp av Foot er legen som kan redde fem dødelig syke pasienter, men bare ved å drepe en annen person for så å brygge et serum av levningene hans.[9] Dette eksempelet, samt det om jernbanevognen ute av kontroll, var det Thomson hadde i tankene når hun første gang komponerte det hun kalte *the trolley problem*:

> «*David is a great transplant surgeon. Five of his patients need new parts – one needs a heart, the others need, respectively, liver, stomach, spleen, and spinal cord – but all are of the same, relatively rare, blood-type. By chance, David learns of a healthy specimen with that very blood-type. David can take the healthy specimen's parts, killing him, and install them in his patients, saving them. Or he can refrain from taking the healthy specimen's parts, letting his patients die.*»[10]

> «*Edward is the driver of a trolley, whose brakes have just failed. On the track ahead of him are five people; the banks are so steep that they will not be able to get off the track in time. The track has a spur leading off to the right, and Edward can turn the trolley onto it. Unfortunately there is one person on the right-hand track. Edward can turn the trolley, killing the one; or he can refrain from turning the trolley, killing the five.*»[11]

I en senere artikkel gjør hun likevel noen revisjoner, og presenterer et nytt scenario, som er det de fleste i dag kjenner som *the trolley problem*. Her er eksempelet med Edward endret til førsteperson, slik at det er du som må ta et valg. Du er en tilfeldig tilskuer til det som holder på å utspille seg, og du må ta valget mellom å la jernbanevognen treffe de fem arbeiderne, eller du kan dra i en spake, slik at jernbanevognen skifter spor og bare treffer én person. I tillegg til dette scenariet, som i dag ofte omtales som *switch,* introduserer hun det som ofte omtales som *footbridge*:

> «*you are standing on a footbridge over the trolley track. You can see a trolley hurtling down the track, out of control. You turn around to see where the trolley is headed, and there are five workmen on the track where it exits from under the footbridge. What to do? Being an expert on trolleys, you know of one certain way to stop an out-of-control trolley: Drop a really heavy weight in its path. But where to find one? It just so happens that standing next to you on the footbridge is a fat man, a really fat man. He*

[9] Foot 1967: 4
[10] Thomson 1976: 206
[11] Thomson 1976: 206

is leaning over the railing, watching the trolley; all you have to do is to give him a little shove, and over the railing he will go, onto the track in the path of the trolley»[12]

Selv om det finnes mange variasjoner og beslektede problem kan man si at dette er kjernen av *the trolley problem*: Hvorfor det er tillatelig å ofre den ene for de fem i *switch*, men ikke i *footbridge*? Sagt på en annen måte: Hvorfor er det slik at vi når vi blir konfrontert med *switch* er kalkulerende konsekvensialister, men straks vi blir konfrontert med *footbridge* er troende deontologer? Her er det verdt å merke at til tross for at dette mønsteret i utgangspunktet var stipulasjoner som ble gjort fra armkroken, så har det også blitt verifisert gjennom empiriske undersøkelser.[13] En potensiell løsning på *the trolley problem* er en som peker tilbake på konflikten mellom konsekvensialisme og deontologi. Enten har konsekvensialismen rett, og det er tillatelig å ofre den ene for de fem både i *footbridge* og i *switch*, eller så tar konsekvensialismen feil. Om konsekvensialismen tar feil så har nødvendigvis deontologien rett. På grunn av måten jeg har definert konsekvensialisme og deontologi på betyr det i teorien at alle kombinasjonene av tillatelig og utillatelig er mulige. Kanskje finnes det veldig sterke begrensninger for når vi kan maksimere nytte slik at det ikke er tillatelig å ofre den ene for de fem i noen av tilfellene. Kanskje finnes det særdeles få begrensninger for når vi kan maksimere nytte, slik at det faktisk er tillatelig å ofre den ene for de fem i begge tilfellene. Eller så er en mellomting sant. Det finnes visse begrensninger på når vi kan maksimere nytte, slik at det bare er tillatelig å ofre den ene i *switch*, men ikke i *footbridge.* I teorien er det selvfølgelig også mulig at det er omvendt: Det finnes begrensninger på når vi kan maksimere nytte slik at det er tillatelig å ofre den ene i *footbridge*, men ikke i *switch.*

Som allerede påpekt mener de fleste at det er tillatelig å ofre den ene i *switch*, men ikke i *footbridge.* De fleste forsøkene på å løse *the trolley problem* har også forsøkt å gjøre det ved å forklare nettopp dette mønsteret. Et av de vanligste forsøkene har vært å referere til doktrinen om dobbeleffekten; altså at det er en signifikant moralsk forskjell på om man dreper noen som et middel eller kun som en forutsigbar konsekvens av en handling. En annen har vært å påstå at negative plikter utveier positive plikter, mens et beslektet forsøk har vært å påstå at det er forskjell på å aktivt drepe noen og kun å lø noen dø. Jeg skal ikke gå i dybden på disse

[12] Thomson 1985: 1409
[13] Greene 2008: 42

forklaringsmodellene, men nøyer meg med å bemerke at de har en fellesnevner: De er deontologiske distinksjoner. Det betyr at dersom konsekvensialismen skulle vise seg å være den korrekte normative modellen faller uansett disse foreslåtte forklaringsmodellene bort.

Om en slik triumf skulle være på horisonten for konsekvensialismen vil *the trolley problem* opphøre å eksistere siden det da vil være tillatelig å ofre den ene både i *switch* og i *footbridge.* På lignende vis har Thomson påstått at *the trolley problem* ikke lenger eksisterer fordi det ikke er tillatelig å ofre den ene verken i *footbridge* eller i *switch.*[14] Frances Kamm har foreslått at vi burde se på *the trolley problem* som et vitenskapelig eksperiment, konstruert spesifikt for å teste konkrete teorier og prinsipper.[15] Deontologi og konsekvensialisme er selvfølgelig blant disse teoriene. Likevel har det ofte, som jeg allerede har vært inne på, også blitt brukt som et tankeeksperiment for å teste deontologiske distinksjoner. Videre i denne teksten skal jeg derimot se på det som et eksperiment designet for å studere kontrastene mellom deontologisk- og konsekvensialistisk normativ teori.

Det kan være fristende å spørre hva som egentlig er problemet med at den samme debatten og de samme argumentene gjentas på denne måten, uten at det tilsynelatende leder noen vei. Alle disipliner har sine utfordringer som det tar tid å jobbe seg gjennom og finne ut av. I den sammenhengen er det passende å trekke fram J.L. Mackies *Argument From Relativity.* Selv om Mackies *Argument From Relativity* i utgangspunktet et argument myntet på den metaetiske debatten rundt moralsk realisme så bør det være illustrerende også i denne sammenhengen. Som Mackie har påpekt er det, om man forutsetter eksistensen av objektive moralske fakta, besynderlig at det tilsynelatende ikke eksisterer noen konsensus rundt hva disse fakta består i. Denne uenigheten akkomoderes enklest av å forstå de forskjellige moralske kodene som et resultat av forskjeller i levesett og kultur, heller enn en eller annen form for imperfekt persepsjon av objektive verdier, vil kritikerne hevde.[16] Et vanlig tilsvar til denne anklagen er å vise til at sekulær normativ etikk er en relativt ung disiplin. Med unntak av Buddha, Konfutse og noen antikke grekere og -romere har de aller fleste som har arbeidet med å utvikle normativ teori gjort det med utgangspunkt i en religiøs autoritet. De siste

[14] Thomson 2008
[15] Kamm 2016: 13
[16] Mackie 1977: 37

århundrene har det selvfølgelig vært flere, men sammenlignet med naturvitenskapene er det forsvinnende få årsverk som er investert i sekulær normativ etikk.[17] Tanken med dette er å appellere til en viss tålmodighet: Om vi bare gir det tid, så vil normativ etikk som disiplin utvikle seg. Vi vil løse flere av problemene og utfordringene som eksisterer innen disiplinen, og vi vil i større grad oppleve enighet rundt de kontroversielle dilemmaene.

Men når det er sagt så er det også rimelig å snu på det. Til tross for at ressursene som er lagt ned i utviklingen av normativ etikk er marginale sammenlignet med de ressursene som er lagt ned i mange av naturvitenskapens disipliner, så fritar ikke dette normativ etikk fra kravet om en viss framgang. Om utviklingsnivået til de mest framgangsrike naturvitenskapene er ti ganger så høyt som utviklingsnivået til sekulær normativ etikk så bør det fortsatt være rimelig å forvente at sistnevnte kan vise til en tidel av framgangen til førstnevnte. Som de foregående sidene har vist er den mest sentrale debatten innen normativ etikk en debatt preget av stagnasjon, hvor det samme argumentet resirkuleres om og om igjen. Dette er definitivt ikke framgang i stil med det man bør kunne forvente, og det er her Greene og hans prosjekt kommer inn i bildet.

1.2. Joshua Greene

Jeg har argumentert for at den mest sentrale debatten innen normativ etikk de siste hundre årene har vært preget av manglende utvikling. Et prov på dette er den repetitive debatten mellom deontologi og konsekvensialisme. Anklagen om kontraintuitivitet, som jeg har kalt det, har med jevne mellomrom – fra slutten av attenhundretallet fram til i dag – blitt presentert i forskjellige former; uten at dette har tilført debatten noe nytt. Slik jeg forstår *the trolley problem*, som en representasjon av den overnevnte konflikten, vil med andre ord et fremskritt her representere et enormt fremskritt for normativ etikk som disiplin. Som det vil framkomme av neste kapittel er det nettopp et slikt fremskritt Joshua Greenes prosjekt har potensial til å utgjøre. Greenes innovative prosjekt henter sine ressurser fra psykologi og nevrovitenskap, og forsøker med utgangspunkt i *the trolley problem* å vise hvorfor deontologi er uegnet som normativ teori. Om Greenes prosjekt er vellykket representerer det ikke bare et stort fremskritt

[17] Parfit 1987: 453

for normativ etikk som disiplin, men det er også et godt eksempel på hvordan de empiriske vitenskapene kan ha en vellykket innflytelse på filosofien, og især normativ etikk.

Jeg har allerede referert til Mackies *argument from relativity*. I den forbindelse er det viktig å påpeke at selv om Greenes prosjekt bærer prov om framskritt innen normativ etikk så bærer det nødvendigvis ikke prov om en frifinnelse av moralsk realisme. Greene er selv klar på at til tross for at hans prosjekt viser hvorfor deontologi er uegnet som normativ teori, og han selv favoriserer konsekvensialismen som normativ teori, så mener han ikke at vi nødvendigvis skal se på konsekvensialismen som *sann*.[18] Jeg kommer tilbake til dette i kapittel fire, og enn så lenge nøyer jeg meg med å si at dette er en tekst om et potensielt framskritt innen normativ etikk – ikke nødvendigvis innen metaetikk.

I neste kapittel, kapittel to, vil jeg presentere Greenes prosjekt i sin helhet. Som det kapitlet også vil illustrere er noe av det mest fascinerende ved Greenes argument at det i motsetning til mange beslektede argument søker å avsløre kun deler av moralen, altså deontologien, og ikke moralen i sin helhet. Det er derfor jeg har valgt å presentere Greenes argument i den konteksten som denne innledningen representerer, nettopp fordi det bærer prov om fremskritt innen en disiplin preget av stagnasjon.

Siden jeg ønsker å argumentere for at Greenes prosjekt kan være berikende for normativ etikk som disiplin er det noen kritiske momenter som vil være særlig viktig for meg å forfølge for å nå dette målet. For det første: For at Greene skal lykkes i å diskreditere deontologisk etikk er det viktig at den deontologien han kritiserer gjennom sine eksperiment er en god representasjon av den deontologien som er operativ når det bedrives normativ etikk. At normative teorier på en god måte kan reduseres til laboratorieformat er tross alt ingen selvfølge. Som neste kapittel vil avsløre gjør Greene dette ved hjelp av det han kaller *karakteristisk deontologiske dommer*, som er de dommene som mest naturlig følger av deontologisk teori. Men hva om Greenes *karakteristiske deontologiske dommer* kun er dårlige

[18] Greene 2008: 77

deontologiske dommer? Det følgende spørsmålet vil derfor være utgangspunktet for kapittel tre:

- Kritiserer Greene deontologien i sin beste form, og kan den eventuelt tenkes på en måte som gjør den mindre sårbar for Greenes kritikk?

Her vil jeg i utgangspunktet argumentere for at Greene nødvendigvis ikke kritiserer deontologien i sin beste form. Både fordi det er mulig å se for seg at det finnes faktorer som er i stand til å forbedre de deontologiske dommene, og fordi forbindelsen mellom de karakteristiske dommene og deontologiens essens kan sies å være svak og preget av metodiske utfordringer. På dette punktet er jeg i overenstemmelse med en av de vanligste kritikkene Greenes prosjekt er blitt utsatt for. For eksempel har, som jeg vil komme nærmere inn på, både Frances Kamm, så vel som Guy Kahane og Nicholas Shackel, reist lignende spørsmål vedrørende Greenes mangelfulle metode. Der de nøyer seg kun med å bemerke at Greene har disse utfordringene går jeg derimot ett steg lenger, og spør om det finnes en bedre måte å forklare de eksperimentelle resultatene Greene tross alt kan vise til på enn ved å forutsette en sterk relasjon mellom de karakteristiske dommene og deontologiens essens. Jeg argumenterer så for at dette kan gjøres ved å forutsette en distinksjon mellom det jeg kaller rigide- og fleksible teorier, men at distinksjonen kun lykkes med å gjøre den jobben dersom man forutsetter at det ikke finnes relevante forskjeller mellom deontologiens grunnleggende intuisjoner og konsekvensialismens tilsvarende intuisjoner.

For at Greenes konklusjon skal kunne være et positivt bidrag til normativ etikk slik jeg har fremstilt det er det viktig at kritikken hans ikke ender opp med også å ramme konsekvensialistisk etikk og dens grunnleggende intuisjoner. For som jeg vil argumentere i kapittel fire er også konsekvensialismen avhengig av å gjøre noen grunnleggende antagelser, og dersom disse grunnleggende intuisjonene rammes av den samme kritikken Greene retter mot deontologiens intuisjoner så bidrar det kun til å sannsynliggjøre moralsk skeptisisme, og det kan definitivt ikke regnes som framgang innen normativ etikk. I tillegg til at dette er en selvstendig grunn til å ta opp denne problematikken så følger den også naturlig av diskusjonen i kapittel tre. Der vil jeg nemlig lansere en utradisjonell type deontologi hvor den mest naturlige innvendingen er at den er avhengig av å gjøre mange kunstige normative antagelser, og for at denne innvendingen skal være tilgjengelig for konsekvensialisten må det finnes grunner til å anta at konsekvensialismens grunnleggende intuisjoner er mer robuste enn

deontologiens tilsvarende intuisjoner. Det følgende spørsmålet vil derfor være utgangspunktet for kapittel fire:

- Finnes det relevante forskjeller mellom konsekvensialismens grunnleggende intuisjoner og deontologiens tilsvarende intuisjoner, som gjør førstnevnte bedre egnet til å bygge normative teorier på?

Her vil jeg ta utgangspunkt i Henry Sidgwicks distinksjon mellom dogmatiske- og filosofiske intuisjoner, som han mente begrunnet et sett med intuisjoner som til syvende og sist resulterte i hans foretrukne form for utilitarisme. Jeg argumenterer derimot for at denne distinksjonen i seg selv ikke er tilstrekkelig til å forklare hvorfor konsekvensialismens grunnleggende intuisjoner er å foretrekke over deontologiens tilsvarende intuisjoner. Jeg vil derimot argumentere for at en syntese mellom Sidgwicks argument og det siste tilskuddet til Greenes prosjekt, distinksjonen mellom modellfrie- læring- og beslutningsprosesser og modellbaserte-læring- og beslutningsprosesser, er i stand til å gjøre denne jobben. Denne syntesen vil også vise seg å være helt avgjørende siden heller ikke Greenes argument alene virker å være tilstrekkelig til å gjøre den jobben han er avhengig av.

Greenes prosjekt har lenge vært populært og derfor også mye kritisert. Det vil selvfølgelig være naivt å tro at Greenes prosjekt er en endelig suksess selv om det overlever mine to kritiske innvendinger. Men mine kritiske innvendinger og det perspektivet jeg tar til Greenes prosjekt er så vidt jeg har oversikt over ikke særlig diskutert i litteraturen. Mine kritiske innvendinger er også en forutsetning for at hans prosjekt skal kunne forstås som et fremskritt innen normativ etikk. Det vil si at dersom hans prosjekt består mine to kritiske prøver, representert med de to problemstillingene så langt presentert, så vil man i det aller minste kunne si at Greene prosjekt har potensial til å kunne representere et etterlengtet framskritt innen normativ etikk som disiplin. Og om så er tilfelle bør dette gjøre Greenes prosjekt fortjent til en dose tålmodighet hva angår andre former for kritikk hans prosjekt måtte møte. Å argumentere for dette er det som er det endelige målet for denne oppgaven.

I det avsluttende kapittel fem vil jeg samle trådene og oppsummere de viktigste momentene fra oppgaven. Siden Greenes prosjekt overlever mine to kritiske prøver vil jeg konkludere

med at hans prosjekt faktisk har potensial til å representere etterlengtet framskritt for normativ etikk som disiplin.

Kapittel 2: Greenes prosjekt

2.1. Dual-processing theory

I kjernen av Greenes teori finner man *dual-processing theory*. Ifølge denne teorien har den menneskelige hjernen to distinkte måter den kan fungere på. Greene sammenligner dette med måten et moderne digitalkamera fungerer på. Akkurat som den menneskelige hjernen er et digitalkamera designet for å finne den beste balansen mellom effektivitet og fleksibilitet. Av og til er jeg på farta, og kommer over et nydelig landskap eller en sjelden fugl jeg ønsker å forevige. Siden jeg har dårlig tid er jeg avhengig a å være effektiv. Derfor benytter jeg meg av en av de forhåndsprogrammerte fotograferingsmodusene som allerede finnes på kameraet mitt. Om jeg for eksempel velger modusen «landskap» vil kameraet mitt automatisk benytte seg av de innstillingene som vanligvis vil gjøre den beste jobben med å fremheve kvalitetene i et vakkert landskap. Dette landskapet jeg skuer ut over nå er dog ikke et ordinært landskap, og lysforholdene gjør også at et godt fotografi av dette landskapet krever at jeg avviker noe fra de standardiserte innstillingene som vanligvis gir de beste landskapsbildene. Derfor har fotografiapparatet mitt også en manual-modus som tillater meg å velge innstillingene helt selv. Dette gir meg ofte et bedre resultat, siden innstillingene kan skreddersys hvert enkelt motiv. Men samtidig er også dette mer ressurskrevende siden det tar mye lenger tid å velge de riktige innstillingene for hvert enkelt motiv. Derfor benytter jeg kun denne modusen når jeg kommer over et helt ekstraordinært motiv, og benytter ellers noen av de forhåndsprogrammerte innstillingene som følger med kameraet.[19]

Ifølge dual processing-theory er det på akkurat samme måte med menneskehjernen. Den har en automatmodus som består av automatiske reflekser, impulser og intuisjoner. Samtidig har den også en manualmodus: «*a general-purpose reasoning system, specialized for enabling behaviors that serve (…) goals that are not automatically activated by current environmental stimuli*»[20]. Denne dualiteten kan beskrives på den følgende måte: Av og til er vi avhengige av å være effektive. Skoleeksempelet på dette er når du holder på å tråkke på en slange. I det du oppdager at du holder på å tråkke på en slange vil det som oftest være en dårlig ide å reflektere over de forskjellige handlingsalternativene du har i den gitte situasjonen du befinner

[19] Greene 2014: 696
[20] Greene 2014: 696-697

deg i. Da kan det ofte være for sent før du får bestemt deg. I et slikt tilfelle vil det være bedre å ha en refleks som gjør at man trekker seg brått tilbake straks man blir oppmerksom på slangen. I andre tilfeller vil lignende reflekser være katastrofale. Når jeg kjører bil er jeg av og til avhengig av å nærme meg andre biler, for eksempel når jeg står i kø eller skal kjøre inn i en rundkjøring. Som oftest er det nødvendig å holde en viss avstand, men om «bil» skulle utløse en lignende refleks som «slange» ville konsekvensene naturligvis vært katastrofale.[21] I enkelte situasjoner er vi med andre ord best tjent med å reflektere over de forskjellige handlingsalternativene, og være fleksible med tanke på hvordan vi tilnærmer oss forskjellige situasjoner, mens det i andre situasjoner er stikk motsatt, og vi er best tjent med å stole på automatiske intuisjoner og reflekser.

Greene har beskrevet disse to systemene som henholdsvis «kognitivt» og «affektivt». Tanken bak dette er at kognitive representasjoner er nøytrale representasjoner, mens affektive representasjoner er representasjoner som utløser konkrete handlingsmønster eller disposisjoner. «Bil» og «slange» fra forrige avsnitt er gode eksempler på dette. Alle prosesser i hjernen må til en viss grad være kognitive, men Greene ønsker altså å peke ut de arketypiske «kogntitive» prosessene, de som er sentrale for bevisst deliberasjon, planlegging, aktiv hukommelse, impulskontroll og «*higher executive functions*».[22] Dette er funksjoner som gjerne assosieres til konkrete deler av hjernen, slik som «*the dorsolateral surfaces of the prefrontal cortex and parietal lobes*»[23]. Til sammenligning identifiseres de arketypiske «affektive» prosessene til «*the amygdala and the medial surfaces of the frontal and parietal lobes*». Med dette ønsker Greene først og fremst å henvise til prosesser som er hurtige, automatiske og ofte ubevisste.[24] Man kan også, med Daniel Kahneman, si det på denne måten: Når vi tenker på oss selv, så identifiserer vi oss med det kognitive systemet; det bevisste og resonerende selvet som handler og tar valg.[25]

Likevel er det verdt å påpeke at denne måten å karakterisere de to prosessene på ikke er helt uproblematisk. Fiery Cushman har påpekt at begge prosesser involverer et visst «affektivt»

[21] Greene 2008: 40
[22] Greene 2008: 40
[23] Greene 2008: 40
[24] Greene 2008: 41
[25] Kahneman 2011: 21

innhold siden de ikke kun bearbeider informasjon, men også produserer motivasjon for sine foretrukne handlingsalternativer. På samme måte involverer de begge et visst kognitivt innhold i den forstand at begge faktisk bearbeider informasjon.[26] Dette er noe Greene er klar over, og det rammer uansett kun navnet Greene velger å assosiere de to prosessene med – ikke deres respektive innhold. Til tross for dette er det verdt å påpeke at «kognitiv» og «affektiv» er imperfekte plassholdere for det mange kun omtaler som *system 1* og *system 2*.[27] Det er også denne terminologien jeg vil forholde meg til i denne oppgaven, hvor system 1 tilsvarer det Greene beskriver som det affektive systemet, mens system 2 tilsvarer det Greene beskriver som det kognitive systemet.

Dual-processing theory er ikke en teori som primært knyttes til normativ etikk og moralfilosofi, men er først og fremst en psykologisk teori. Derfor er det heller ikke gitt hva som blir resultatet om en forsøker å anvende dual-processing theory på moralfilosofien. Eksempelvis har Jonathan Haidt i sin moderne klassiker *The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment* prøvd å vise at våre normative vurderinger i det store og det hele beror på system 1. Haidt konfronterte en rekke testsubjekter med det følgende dilemmaet: På sommerferie i Frankrike bestemmer to søsken seg for at det ville vært interessant om de prøvde å ha sex. De brukte flere former for prevensjon, og selv om de var enige i at det var en hyggelig opplevelse som brakte dem nærmere hverandre, så bestemte de seg for ikke å gjøre det igjen. Spørsmålet er så, var det OK for dem å ha sex? De fleste som ble spurt mente at det ikke var OK. Men når de ble konfrontert med hvorfor de mente det, fikk de problemer. Mange søkte til grunner som ble ekskludert allerede i det opprinnelige spørsmålet, slik som muligheten for innavl eller emosjonell forringelse av søskenforholdet. Til slutt ble derfor mange drevet til siste skanse: «*I dont know, I cant explain it, I just know its wrong*»[28]. Haidts poeng er at våre normative vurderinger først og fremst styres av system 1: De er hurtige og effektive reflekser. System 2 kommer først inn i bildet som en post-hoc rasjonalisering som forklarer hvorfor den opprinnelige mageføelelsen må stemme.[29]

[26] Cushman 2013: 274
[27] Cushman 2013: 275
[28] Haidt 2001: 814
[29] Haidt 2001: 817

Dette er selvfølgelig kortversjonen av Haidts prosjekt. Poenget er at dette er en plausibel konsekvens av å forstå moralfilosofien i lys av dual-proccecing theory. Greenes prosjekt forsøker derimot å identifisere system 1 men kun en del av de normative vurderingene vi gjør, mens en annen del av de normative vurderingene vi gjør identifiseres med system 2. Mer konkret identifiserer Greene de vurderingene som i forrige kapittel konstituerte anklagen om kontraintuitivitet med system 1. Når vi vurderer det som moralsk utillatelig å dytte personen i *footbridge* ned på jernbanesporet for å bremse jernbanevognen og redde de fem personene er det med andre ord et resultat av de hurtige intuisjonene og refleksene som utgjør system 1. Når vi derimot dømmer det som moralsk tillatelig å dra i spaken i *switch*, og dermed endre sporvalget for jernbanevognen, slik at den heller treffer den ene enn de fem, så er det derimot et resultat av den langsomme og fleksible prosessen som utgjør system 2.

2.2. Deontologi og konsekvensialisme

Greene kaller dette, å vurdere offeret i *footbridge* som utillatelig og offeret i *swtich* som tillatelig, for henholdsvis *karakteristiske deontologiske dommer* og *karakteristiske konsekvensialistiske dommer*. Dette innebærer ikke at det er umulig å produsere en deontologisk dom også i *switch*, eller at en deontologisk dom må være negativ i *switch*. Det betyr snarere at en negativ dom i *footbridge*, som er den vurdering de fleste gjør, mest naturlig lar seg begrunne i form av deontologiske termer, for eksempel med referanse til rettigheter og plikter vi har som rasjonelle skapninger. På lignende vis er en positiv dom i *switch*, som også er den vurderingen de fleste gjør, enklest å rettferdiggjøre med henvisning til konsekvensialistisk teori, for eksempel en kost-/nytteanalyse hvor fem liv veies opp mot ett liv.[30] Denne ukonvensjonelle bruken av de to respektive normative teoriene er ikke ukontroversiell, og dette er noe jeg kommer tilbake til i neste kapittel. For nå nøyer jeg meg med å si at Greene forsvarer denne bruken ved å påpeke at han er interessert i prosessene som ligger bak, og er opphavet til, de respektive normative teoriene, og at det derfor er et nødvendig metodisk grep å vise til deres karakteristiske dommer, siden han ikke kan studere teoriene i seg selv. Hans hypotese er at det er system 1 som ligger bak deontologisk teori, mens det er system 2 som ligger bak konsekvensialistisk teori.[31] Når man allerede har identifisert system 1 og system 2 med forskjellige deler av hjernen kan man ved å redusere deontologisk- og konsekvensialistisk teori til deres respektive karakteristiske dommer studere

[30] Greene 2014: 699
[31] Greene 2008: 39

hva som skjer i hjernen når vi er deontologer og hva som skjer i hjernen når vi er konsekvensialister.

I et av de første eksperimentene Greene og hans kolleger utførte i forbindelse med sin teori ble ni testsubjekter konfrontert med til sammen 60 forskjellige praktiske dilemma samtidig som hjerneaktiviteten deres ble registrert av en fMRI-maskin. Settet av forskjellige dilemma ble først delt opp i moralske- og ikke-moralske dilemma, før de førstnevnte igjen ble delt opp i de som var strukturelt like *switch* og de som var strukturelt like *footbridge;*[32] de som vanligvis fører til karakteristisk konsekvensialistiske dommer og de so vanligvis fører til karakteristisk deontologiske dommer. Jeg kommer tilbake til kriteriene for akkurat denne inndelingen.

Dette er eksempler på noen av dilemmaene testsubjektene ble konfrontert med, i tillegg til *switch* og *footbridge*:
Ikke-moralske dilemma;

> «*You are beginning your senior year of college. In order to fulfill your graduation requirements you need to take a history class and a science class by the end of the year. During the fall term, the history class you want to take is scheduled at the same time as the science class you want to take. During the spring term the same history class is offered, but the science class is not. Is it appropriate for you to take the history class during the fall term in order to help you fulfill your graduation requirements?»*[33]

> *«You have decided to make a batch of brownies for yourself. You open your recipe book and find a recipe for brownies. The recipe calls for a cup of chopped walnuts. You don't like walnuts, but you do like macadamia nuts. As it happens, you have both kinds of nuts available to you. Is it appropriate for you to substitute macadamia nuts for walnuts in order to avoid eating walnuts?»*[34]

Moralske dilemma som er strukturelt like *switch;*

> *«You are at home one day when the mail arrives. You receive a letter from a reputable international aid organization. The letter asks you to make a donation of two hundred dollars to their organization. The letter explains that a two hundred-dollar donation will allow this organization to provide needed medical attention to some poor people in another part of the world. Is it appropriate for you to not make a donation to this organization in order to save money?»*[35]

[32] Greene et al. 2001: 2106
[33] Greene et al. 2001: Supplerende materiell
[34] Greene et al. 2001: Supplerende materiell
[35] Greene et al. 2001: Supplerende materiell

> «*While on vacation on a remote island, you are fishing from a seaside dock. You observe a group of tourists board a small boat and set sail for a nearby island. Soon after their departure you hear over the radio that there is a violent storm brewing, a storm that is sure to intercept them. The only way that you can ensure their safety is to warn them by borrowing a nearby speedboat. The speedboat belongs to a miserly tycoon who has hired a fiercely loyal guard to make sure that no one uses his boat without permission. To get to the speedboat you will have to lie to the guard. Is it appropriate for you to lie to the guard in order to borrow the speedboat and warn the tourists about the storm?»*[36]

Moralske dilemma som er strukturelt like *footbridge*;

> «*Enemy soldiers have taken over your village. They have orders to kill all remaining civilians. You and some of your townspeople have sought refuge in the cellar of a large house. Outside you hear the voices of soldiers who have come to search the house for valuables. Your baby begins to cry loudly. You cover his mouth to block the sound. If you remove your hand from his mouth his crying will summon the attention of the soldiers who will kill you, your child, and the others hiding out in the cellar. To save yourself and the others you must smother your child to death. Is it appropriate for you to smother your child in order to save yourself and the other townspeople?»*[37]

> *«Your plane has crashed in the Himalayas. The only survivors are yourself, another man, and a young boy. The three of you travel for days, battling extreme cold and wind. Your only chance at survival is to find your way to small a village on the other side of the mountain, several days away. The boy has a broken leg and cannot move very quickly. His chances of surviving the journey are essentially zero. Without food, you and the other man will probably die as well. The other man suggests that you sacrifice the boy and eat his remains over the next few days. Is it appropriate to kill this boy so that you and the other man may survive your journey to safety?»*[38]

Resultatene viste at områder av hjernen som forbindes med system 1 var betraktelig mer aktive når testsubjektene avga karakteristisk deontologiske dommer i forbindelse med moralske dilemma som var strukturelt like *footbridge,* enn når de avga karakteristisk konsekvensialistiske dommer i forbindelse med dilemma som var strukturelt like *switch*. Ingen signifikant forskjell i hjerneaktivitet ble til sammenligning funnet mellom når testsubjektene avga en dom i ikke-moralske dilemma og når de avga en karakteristisk konsekvensialistisk dom i *switch*.[39]

[36] Greene et al. 2001: Supplerende materiell
[37] Greene et al. 2001: Supplerende materiell
[38] Greene et al. 2001: Supplerende materiell
[39] Greene et al. 2001: 2107

2.3. Personlig- og ikke-personlig skade

Det finnes med andre ord et empirisk grunnlag som knytter karakteristisk deontologiske dommer til system 1 og karakteristisk konsekvensialistiske dommer til system 2. Men hvordan kunne Greene og hans kolleger avgjøre om et konkret moralsk dilemma ville fremprovosere en karakteristisk deontologisk dom eller en karakteristisk konsekvensialistisk dom? Deres hypotese var at tanken på å dytte noen i deres visse død på en måte som er «*up close and personal*», slik som i *footbridge*, stimulerer kraftigere emosjonelle impulser enn det som er tilfelle ved tanken på å kun dra i en spake, og dermed sende noen i døden, slik som i *switch*. Med andre ord er det graden av *personlighet* i skaden som gjøres, som avgjør om et dilemma klassifiseres som strukturelt likt *footbridge* eller strukturelt likt *switch*.[40] Den praktiske inndelingen ble gjort av to uavhengige kodere. De ville karakterisere et dilemma som personlig dersom (a) den aktuelle handlingen kunne forventes å forårsake alvorlig legemlig skade (b) på en konkret person eller en konkret gruppe med mennesker, (c) samtidig som skaden ikke var et resultat av en allerede eksisterende trussel. Selv om det siste kriteriet kan virke påfallende er den bakenforliggende tanken at en allerede eksisterende trussel som agenten kun kan regulere på visse måter er mindre emosjonelt stimulerende enn en trussel agenten selv sanksjonerer.[41]

Her er det dog nødvendig å legge til et lite apropos. Akkurat slik som Greenes karakteristikk av system 1 og system 2 som henholdsvis affektivt og kognitivt kan anses som en imperfekt plassholder, så kan det samme sies om distinksjonen mellom personlige og ikke-personlige moralske dilemma som indikator på hvilke dilemma som sannsynligvis vil produsere karakteristisk deontologiske og -konsekvensialistiske dommer. Frances Kamm har produsert et eksempel som illustrerer hvorfor: Jeg holder flere spedbarn i armene mine samtidig som en jernbanevogn ute av kontroll er på vei mot meg slik at de vil treffe og drepe barnene i armene mine. Jeg er ikke truet av vognen, men for å redde barna må jeg vende meg andre vei, noe som vil resultere i at jeg vil treffe en annen person med armene mine, slik at vedkommende vil bli dyttet og falle ned et langt stup og følgelig omkomme.[42] Dette eksempelet tilfredsstiller alle kriteriene for å karakteriseres som et personlig moralsk dilemma. Likevel vil de fleste gi

[40] Greene 2008: 43
[41] Greene et al. 2001: 2107
[42] Kamm 2009: 334

en karakteristisk konsekvensialistisk dom og si at det er tillatelig for meg å ofre den ene personen for å redde spedbarnene i armene mine. Greene har siden bekreftet dette ad empiriske veier.[43]

Det å henvise til personlighet alene er derfor ikke tilstrekkelig for å forklare hvorfor enkelte dilemma, slik som *footbridge*, utløser sterke negative emosjonelle impulser. Men selv om graden av personlighet ikke utgjør hele forklaringen så betyr det ikke at det ikke kan være en del av forklaringen.[44] At graden av personlighet i skaden er en relevant faktor har Greenes eksperimenter tross alt allerede indikert. I *footbridge* sier for eksempel 31% av testsubjektene at det er tillatelig å ofre den ene for å redde de fem. Men når betingelsene blir endret slik at man ikke trenger å dytte den ene personen på gangbroen, men ganske enkelt kan slippe vedkommende ned på skinnene under via en fjernutløst fallem i gangbroen øker antallet som er villige til å ofre den ene til 63%.[45]

I ettertid har Greene og hans team også gjennomført en rekke forsøk for å se på hvordan graden av intensjon i skaden også kan spille inn. Dette er en naturlig antakelse siden doktrinen om dobbelteffekten tross alt har vært en av de aller vanligste måtene å forsøke å forklare *the trolley problem* på. Den tilsier at det av og til kan være tillatelig å forårsake skade når den kun er en forutsett men ikke nødvendig konsekvens av å nå et mål, men ikke når skaden i seg selv er et middel for å oppnå målet. I et eksperiment ble testsubjektene konfrontert med fire forskjellige moralske dilemma. To av dilemmaene var ikke-personlige, mens to av dilemmaene var personlige. Av de ikke-personlige dilemmaene involverte ett skade som et nødvendig middel og ett skade som en forutsett bieffekt. Av de personlige dilemmaene var det også ett som involverte skade som et nødvendig middel, mens det andre kun involverte skade som en forutsett bieffekt.[46] Det Greene og hans team fant ut var at personligheten av skaden kun var relevant når den aktuelle skaden var intensjonal i den forstad at den var et middel for å nå et mål.[47] Dette forklarer også Kamms moteksempel, siden skaden som her må forårsakes

[43] Greene 2008b: 108
[44] Greene 2008b: 114
[45] Greene 2014: 709
[46] Greene er al. 2009: 19
[47] Greene et al 2009: 21

for å redde spedbarnene kun er en forutsett bieffekt og ikke et nødvendig middel for å nå målet.

Det som til nå er blitt sagt bør være nok til å konkludere med at graden av personlighet av den aktuelle skaden er en relevant – om enn ikke alene tilstrekkelig – faktor når det kommer til å forklare hvorfor enkelte scenarier utløser mekanismene i system 1, mens andre ikke gjør det. Forklaringen på hvorfor det er slik er også intuitiv: Vold som er personlig er noe vi mennesker har forholdt oss til så lenge vi har eksistert, og lenge var det også den eneste formen for vold vi kunne utøve. Fra et evolusjonsmessig perspektiv er vår evne til å utøve denne type vold mye eldre enn vår evne til kompleks og abstrakt tenkning, og derfor er det ikke overraskende at vi har iboende responser til denne type personlig vold som er særdeles kraftige, om enn primitive. Det er ikke vanskelig å se for seg hvorfor utviklingen av en respons-mekanisme av denne typen er gunstig. Din sjanse til i første omgang å overleve og i andre omgang å sørge for at «dine» gener i størst mulig grad blir brakt videre øker dersom du er en del av ev en gruppe som kan samarbeide. Uten her å gå i dybden i dybden av et slikt argument er det rimelig å påstå at du, *ceteris paribus*, vil ha mindre sjanse for å bli drept og ha større tilgang på ressurser dersom du er en del av et samarbeidene kollektiv, og at disse faktorene igjen vil bedre dine muligheter til å forplante deg. Å leve i en gruppe er likevel ikke kostnadsfritt. Det forutsetter for eksempel at vi løser konflikter på en passende måte og ikke slår hverandre i hjel ved første anledning. Det er her de sterke negative impulsene som ifølge Greene ligger bak de *karakteristisk deontologiske dommene* kommer inn i bildet, fordi de nettopp promoterer vår evne til samarbeid.[48] Siden ikke-personlig vold er en utradisjonell form for vold som mennesket ikke har vært i stand til å utøve like lenge som personlig vold har vi heller ikke utviklet de samme mekanismene for å begrense denne type vold. Derfor utløses ikke de samme system 1-impulsene av *switch* og lignende scenarier, og de blir heller vurdert ved hjelp av en nøytral kost-/nytteanalyse.

Kjernen i Greenes teori så langt er at enkelte faktorer utløser sterke negative impulser som får enkelte handlinger til å virke helt uholdbare. Dette er for eksempel tilfelle i *footbridge*, men ikke i *switch*. Likevel er det ikke alle som følger mønsteret Greene ønsker å belyse. Enkelte

[48] Greene 2013: 23

konkluderer eksempelvis med at offeret i *footbridge* faktisk er tillatelig. Likevel må man, med enkelte unntak, anta at også disse opplever den samme negative impulsen som får de aller fleste til å steile i et scenario som *footbridge*. En tilleggshypotese for Greene og hans team var derfor at deltakere som gikk mot trenden på denne måten ville bruke lengre tid på å avgi sin dom. Grunnen til dette er at disse subjektene først måtte overkomme de negative impulsene produsert av system 1, før de i kraft av system 2 kunne iverksette en mer nøytral kost-/nytteanalyse. Nok en gang bekreftet de eksperimentelle resultatene deres hypotese.[49]

Når det er sagt må det også legges til at ikke absolutt alle opplever slike negative impulser i forbindelse med intensjonal og personlig vold. Pasienter med frontotemporal demens har for eksempel tre ganger så høy sannsynlighet for å gi karakteristiske konsekvensialistiske dommer på den type dilemma som subjektene i Greenes eksperiment ble konfrontert med.[50] Videre har pasienter med skade på ventromedial prefrontal korteks (VMPFC) fem ganger så høy sannsynlighet for å gi konsekvensialistiske dommer.[51] Psykopater har også mye høyere sannsynlighet enn ikke-psykopater for å gi karakteristisk konsekvensialistiske dommer.[52] Alle disse bemerkningene bidrar selvfølgelig til å bygge opp under Greenes teori, siden fellesnevneren for alle disse diagnosene er at de medfører defekter eller redusert kapasitet i de delene av hjernen som er antatt å stå bak de negative emosjonelle impulsene bak karakteristisk deontologiske dommer. Dette er kun et knippe av de empiriske bevisene Greene har til sin disposisjon,[53] men for mitt vedkommende vil jeg anse de empiriske bevisene som de nå er framlagt for tilstrekkelig til å anse Greenes empiriske grunnpilarer som bevist, og derfor gå videre med presentasjonen av Greenes prosjekt.

Hjernen vår har altså to moduser. Automatisk- og manuell modus. Eller system 1 og system 2, som jeg velger å kalle det. Greenes fMRI-resultater har koblet system 1 til karakteristisk deontologiske dommer og system 2 til karakteristisk konsekvensialistiske dommer. Grunnen til dette er at system 1 aktiveres av enkelte faktorer som er til stede i enkelte moralske dilemma, slik som *footbridge*, men ikke i andre, slik som *switch*. Det har vært

[49] Greene 2001: 2107
[50] Greene 2014: 701
[51] Greene 2014: 702
[52] Greene 2014: 703
[53] For en mer utførlig liste kan man se side 700-707 i Greenes *Beyond Point-and-Shoot-Morality*

evolusjonsmessig gunstig for oss å utvikle sterke aversjoner mot personlig og intensjonal vold, fordi dette har redusert vår evne til å samarbeide og leve sammen. Derfor kan man forklare *the trolley problem* på den følgende måten: I *switch* er de fleste av oss komfortable med å være konsekvensialister fordi vi anerkjenner at fem liv er mer enn ett liv, og når *switch* heller ikke aktiverer system 1 og dets alarmfunksjoner så later dette til å være den eneste relevante betraktningen. I *footbridge* er vi derimot ikke komfortable med å være konsekvensialister. Siden *footbridge* involverer vold som er personlig og intensjonal så aktiverer det system 1. Da hjelper det for de fleste av oss ikke lenger at fem liv er mer enn ett liv. Det er galt uansett.

Til tross for dette er det ikke helt klart nøyaktig hva som er galt med deontologisk etikk og den påståtte relasjonen til system 1. System 1 virker tross alt som en veldig praktisk mekanisme, og det må jo sies å være en god ting at vi ikke liker å slå hverandre i hjel? Som Kahneman skriver: Selv om det er system 2 vi identifiserer oss med, så er det system 1 som er den virkelige helten.[54] Greene er faktisk ikke uenig i dette, og vedgår at vi i hverdagen bør følge system 1s direktiver. Hans poeng er derimot at det stiller seg annerledes om en ønsker å løse moralske problemer – å bedrive moralfilosofi – og at det er her deontologisk etikk kommer til kort.[5556]

Problemet med de karakteristisk deontologiske dommene er ikke i seg selv at de kan spores tilbake til vår evolusjonshistorie eller at de i større grad er støttet av prosesser som er affektive heller enn «kognitive». Problemet er snarer faktorene som utløser mekanismene som står bak. For graden av personlighet i den aktuelle skaden, slik beskrevet i Greenes eksperiment, kan umulig være en relevant faktor for å avgjøre hvorvidt den aktuelle skaden er moralsk tillatelig eller ei. Se for deg at du holder på å utleve et virkelig *footbridge*-scenario, og skrekkslagen ringer en venn for å be om råd. Om vennen din svarer «Vel, det spørs, må du dytte han over rekkverket selv, eller kan du fjernutløse en fallem i gangbroen?»[57] vil du åpenbart avskrive vennen din som en kompetent moralsk aktør og uansett se bort fra hans råd. Grunnen til at

[54] Kahneman 2011 :21
[55] Greene 2017: 2
[56] Greene 2014: 714
[57] Greene 2014: 713

system 1s influens er et problem når vi ønsker å bedrive moralfilosofi er med andre ord at faktorene som utløser dens mekanismer ikke er moralsk relevante. Om jeg tar livet av deg ved å slå deg i hjel med en stein eller om jeg tar livet av deg ved å dra i en spake som utløser en fallem kan i beste fall sies å være avslørende for min moralske karakter, men kan på ingen måte sies å være relevant for hvorvidt det er moralsk tillatelig for meg å slå deg i hjel. Siden deontologisk etikk på denne måten tar hensyn til moralsk irrelevante faktorer bør vi se på den og dens karakteristiske dommer som en uegnet aktør når vi ønsker å løse moralske problemer.

I senere tid Greene gått bort fra menneskets evolusjonshistorie som den viktigste forklaringen på hvorfor system 1s intuisjoner ofte er villedende. For vår automatmodus er bare velfungerende, sier han, der hvor den har blitt formet av erfaring. Denne erfaringen kan tilegnes på tre måter: Personlig erfaring, kulturell erfaring og evolusjonær erfaring. Vi er for eksempel genetisk predisponert til å frykte slanger. Vi er kulturelt disponert til å være redd for skytevåpen. Når det gjelder personlig erfaring kan det variere mer, men mange av oss vet for eksempel å vokte oss for en varm koketopp. Når vi møter på ukjente problem, der hvor ingen av disse kildene til kunnskap er til stede, vil det altså være et kognitivt mirakel om system 1 viser seg å være en god veiviser.[58]

Dette er dette som er tilfelle når vi møter på ukjente moralske problemer. Der hvor i mangler tilstrekkelig erfaring bør vi altså stole mer på system 2 enn system 1. Én bråte ukjente moralske problem er selvfølgelig de som springer ut av den brå kulturelle- og teknologiske utviklingen vi har opplevd siden den industrielle evolusjon, slik som de forbundet med eksempelvis klimaendringer og bioteknologi. Greene mener også at *footbridge* er et ukjent moralsk problem på grunn av dets besynderlige konstruksjon. Han skriver også at en mer generell indikator for ukjente moralske problem bør være uenighet. Når to personer er uenig om hva de bør gjøre så er det sannsynligvis fordi de har motstridene intuisjoner, noe som betyr at minst en av personenes system 1-intusjoner må være feil. Siden de ikke kan vite hvem som tar feil bør de begge se bort fra sine intuisjoner, og heller stole på system 2.[59]

[58] Greene 2014: 714
[59] Greene 2014: 716

2.4. Modellbaserte- og modellfrie læring- og beslutningsprosesser

I en nylig publikasjon fra 2017 har Greene bygget videre på denne strategien på en veldig interessant måte, og i kapittel fire vil det bli tydelig nøyaktig hvorfor dette tilskuddet er så verdifullt. Her presiserer han hvordan han nå ser for seg at et mer generelt begrep om læring, heller enn en eksplisitt referanse til menneskets evolusjonshistorie, kan forklare hvorfor tanken på å forårsake personlig skade ledsages av kraftige negative intuisjoner:

> «*I have at times (Greene, 2007; Greene & Haidt, 2002) suggested that the negative reaction to causing ‘‘personal’’ harm, as in the footbridge case (Greene et al., 2001, 2009), reflects a domain-specific, innately supported affective response. My view on this question has since shifted. Following Cushman (2013) and Crockett (2013), I’m increasingly convinced that learning plays a dominant role in generating these patterns of judgment and that whatever genetic influences are at work—which may be very important—are likely operating on domain-general cognitive systems (Greene, 2014b; Shenhav & Greene, 2010), rather than on a domain-specific ‘‘module’’ related to morality, or to causing personal harm more specifically.*»[60]

I denne teksten tar Greene utgangspunkt i distinksjonen mellom modell-baserte- og modell-frie lærings- og beslutningsmodeller, som har sin opprinnelse maskinlæring. Modell-basert læring går ut på å akkumulere informasjon om det aktuelle beslutningsmiljøet for å bygge en kausal modell av det aktuelle miljøet. En rotte i en labyrint kan for eksempel lære seg å finne en belønning ved å konstruere et internt kart over labyrinten som inkluderer hvor belønningen finnes. Et kart er en integrert kausal modell fordi det inneholder informasjon om de forventede konsekvensene av å bevege seg i forskjellige retninger fra forskjellige start-punkt. Men det trenger heller ikke være et kart. Man kan for eksempel lære seg å operere en maskin ved å konstruere en eksplisitt forståelse for hva de forskjellige delene gjør og hvordan de påvirker hverandre, og hvordan man bruker de for å oppnå ønsket resultat.[61]

Modell-fri læring og beslutningstaking konstruerer derimot ikke en kausal modell, men identifiserer positive- og negative verdier direkte til konkrete handlinger, basert på hvorvidt handlingene har ført til en form for belønning tidligere. Om en rotte i en labyrint for eksempel

[60] Greene 2017: 3
[61] Greene 2017: 4

snubler over en ostebit etter å ha tatt en høyresving ut av et rødt rom vil den føle en trang til å ta en høyresving neste gang den befinner seg i et rødt rom. Denne type responser kan lenkes sammen ved hjelp av det som er kjent som «*temporal difference reinforcement learning* (TDRL)» slik at rotten kan se på det å ankomme et rødt rom som en belønning i seg selv. Om rotten kommer til det røde rommet ved å svinge til venstre ut av et blått rom kan den så koble en positiv verdi til det å svinge til venstre i et blått rom og i sin tur igjen vurdere det som en belønning å ankomme det blå rommet. Ved å lenke sammen disse responsene kan rotten lære seg å navigere labyrinten for å finne belønningen; ikke fordi den forstår hvor den på enhver tid er på vei hen, men simpelt hen fordi det føles riktig.[62]

Den modell-baserte strategien er mer kostbar fordi den krever at man lagrer mer informasjon om det aktuelle miljøet, og når en beslutning skal tas må også hele beslutnings-treet gås gjennom for å identifisere hvilken sekvens av handlinger som på en mest effektiv måte vil realisere målet. Den modell-frie strategien har derimot ikke disse utgiftene siden den ikke lagrer all informasjonen om miljøet, men kun assosierer positive eller negative verdier direkte med handling-kontekst-par. Fordelen til den modell-baserte strategien er at den er mer fleksibel. Om plasseringen til enten gevinsten eller startpunktet endrer seg vil den modell-frie rotten måtte starte helt forfra, mens den modell-baserte rotten kan tilpasse seg endringene ganske enkelt ved å oppdatere det integrerte kartet.[63] Enda viktigere er det kanskje at den modell-baserte rotten ikke bare kan tilpasse seg et miljø som endrer seg, men også endringer i seg selv og dens verdier. I et eksperiment ble for eksempel en rotte trent opp til å trykke på en bryter for å få en belønning. For å holde rotten motivert gikk den på en streng diett. Men så endrer betingelsene seg. Før en test ble rotten foret helt til den ikke lenger viste interesse for mat. Til tross for at den ikke lenger hadde et ønske om mat fortsatte rotten i noen situasjoner å trykke på bryteren. Denne irrasjonelle handlingen lar seg lett forklare av den modell-frie modellen ved å påpeke at rotten har en positiv representasjon av å trykke på bryteren i seg selv, når den er i treningsapparatet, til tross for at det i utgangspunktet er belønningen som følger av handlingen den egentlig verdsetter.[64]

[62] Greene 2017: 4
[63] Greene 2017: 4
[64] Cushman 2013: 279

Poenget er at denne formen for intuitiv beslutningstaking klarer seg dårlig i en verden som stadig endrer seg. Det er ingen overraskelse at den modell-frie strategien identifiseres med system 1 og den modell-baserte strategien identifiseres med system 2, og man kan nå begynne å ane hvordan dette relaterer seg til moralfilosofien. Når de aller fleste dømmer det som utillatelig å ofre den ene i *footbridge* så er det som kjent på grunn av en automatisk og affektiv respons som utløses av handlinger av den typen som er representert i *footbridge*. Både gjennom vår egen og andres erfaring har vi lært at voldelige handlinger har dårlige konsekvenser. Ikke bare har det negative konsekvenser for offeret, men i de fleste samfunn fører det også til negative sanksjoner for den som står bak. Det *footbridge* gjør er altså å ta en type adferd som i den virkelige verden produserer dårlige konsekvenser og plasserer det i en kunstig kontekst hvor det nødvendigvis produserer de beste konsekvensene, for så å spørre om det nå plutselig er akseptabelt.[65] *Footbridge* snur altså opp-ned på vår moralske labyrint, og akkurat som den modell-frie rotten mangler deontologien fleksibiliteten til å tilpasse oss.

Deontologi svarer altså til en modell-fri strategi fordi den ifølge Greene kobler verdi direkte opp mot handlinger, mens konsekvensialisme svarer til modell-basert strategi fordi den kobler verdi opp mot konsekvenser.[66] Akkurat som rotten som fortsetter å trykke på bryteren selv om bryteren ikke lenger fører til den positive effekten som i utgangspunktet var grunnen til at den begynte å trykke på bryteren, men nå trykker på bryteren fordi den ser på bryteren i seg selv som et gode så dømmer vi volden i *footbridge* som utillatelig fordi konsekvensene av volden vanligvis vil være negative, og vi derfor vurderer handlingen i seg selv som et onde, selv når omstendighetene endrer seg slik at konsekvensene av volden faktisk vil være positiv. Som Greene skriver:

> «*If he's a simple sort of rat, he might shrug and say, ''I don't know. I just feel like pressing." Or perhaps a lame excuse: ''I really need the exercise". A more philosophically minded rat, however, might insist that he presses the lever for a more noble reason. He does this, he explains, not to receive some crass reward, but out of a sense of duty. And if such a rat were inspired by the success of mathematics, he might*

[65] Greene 2017: 9
[66] Cushman 2013: 286

attempt to derive this rat-a-gorical imperative from principles of pure rodent reason»[67]

I kapittel fire skal jeg utfordre oppfatningen om at deontologisk etikk nødvendigvis må svare til modellfrie- læring- og beslutningsmodeller. Men selv om det er mulig å se for seg en form for deontologi som heller baserer seg på en modellbasert strategi så vil jeg også argumentere for at denne formen for deontologi vil være sårbar for Greenes kritikk på andre måter. I dette kapitlet har jeg kun fokusert på å beskrive Greenes kritikk slik det fremstår i dag. I den grad jeg har vist til utfordringer for Greenes argument har det kun dreid seg om mindre utfordringer som har vært en del av utviklingen av Greenes prosjekt. I neste kapittel skal jeg derimot ta for meg innvendinger mot Greenes prosjekt med langt mer skadepotensial. For hva om de karakteristisk deontologiske dommene, som har spilt en viktig rolle i dette kapitlet, kun er dårlige deontologiske dommer: Kritiserer Greene deontologien i sin beste form, og kan den eventuelt tenkes på en måte som gjør den mindre sårbar for Greenes kritikk?

[67] Greene 2017: 9

Kapittel 3: Kritiserer Greene deontologien i sin beste form?

Om man ser for seg Immanuel Kant sitte over skrivebordet sitt i Königsberg i 1700-tallets Preussen og streve med utarbeidelsen av det kategoriske imperativ så kaster det et nesten science fiction-aktig lys over Greenes prosjekt. Ved hjelp av hans empiriske metode kan han skjære inn i den deontologiske ryggraden, grave fram dens essens og plukke den fra hverandre. I dette kapitlet skal jeg argumentere for at denne betegnelsen kan være mer passende enn man skulle tro. For det kan argumenteres for at Greene er forut på en tid på en slik måte at han mangler adekvate metoder for å fullbyrde sitt prosjekt. Med dette mener jeg at han mangler adekvate metoder for (1) å skille mellom genuin system 2-aktivitet og post hoc-rasjonalisering, og (2) å studere normative teorier i seg selv. Greene gjør sistnevnte ved hjelp av karakteristiske deontologiske dommer, men hva om dette kun er dårlige deontologiske dommer? Derfor setter jeg i dette kapitlet spørsmålstegn ved hvorvidt Greene kritiserer deontologien i sin beste form, og om den eventuelt kan tenkes på en måte som gjør den mindre sårbar for kritikk. Innledningsvis i dette kapitlet spør jeg hvem bevisbyrden faktisk tilfaller i denne debatten, for så å vise, med utgangspunkt i en fiktiv versjon av Greene originale eksperiment, hvordan de metodiske problemene gjør bevisbyrden problematisk for begge parter i debatten. Jeg avslutter kapitlet med å introdusere en distinksjon mellom rigide- og fleksible normative teorier, som har potensial til å forklare Greenes eksperimentelle resultater på en måte som ikke innebærer en kausal relasjon mellom deontologisk etikk og system 1.

3.1. Bevisbyrden

Den store antagonisten i Greenes prosjekt er Immanuel Kant. Til tross for at fornuft og rasjonalitet ifølge Kant selv er grunnpilarene i hans moralfilosofi er den faktisk et ubevisst uttrykk for de underliggende prosessene i system 1, ifølge Greene. En del av hans bevisførsel baserer seg på eksperimenter hvor han konfronterer college-studenter med moralske dilemma samtidig som hjerneaktiviteten deres dokumenteres ved hjelp av en fMRI-maskin. En naturlig innvending vil være å påpeke at hva som foregår i hjernen til en gjennomsnittlig amerikansk college-student når vedkommende blir konfrontert med et moralsk problem er én ting. Det er derimot en helt annen ting når en av historiens mest innflytelsesrike tenkere skal bedrive samme øvelse. Dette er ikke en grunnløs innvending, og det er interessant. For hva ville skjedd om Kant selv, på en eller annen måte, kunne delta i Greenes originale eksperiment?

Per dags dato er dette umulig, og selv om det var mulig virker Greene sikker i sin sak:

> «*Of course, deontologists may regard themselves and their minds as exceptions to the statistically significant and multiply convergent psychological patterns identified in these studies, but in my opinion the burden is on them to demonstrate that they are psychologically exceptionally in a way that preserves their self-conceptions.*»[68]

På den ene siden er det rimelig av Greene å legge bevisbyrden på de deontologene som hevder at de er unntaket fra regelen. Akkurat som at det er umulig å bevise at alle svaner er hvite er det likeså umulig å bevise at alle karakteristiske deontologiske resultater er et produkt av system 1s underliggende prosesser. Det som dog er mulig, i det minste i teorien, er å bevise er unntaket fra regelen. Det vil si eksistensen av en sort svane, eller i dette tilfelle; en karakteristisk deontologisk dom som produseres uten noen særskilt befatning med system 1.

På den annen side viser det seg at også denne strategien er problematisk. Greene utelukker på ingen måte at system 2 er involvert i deontologisk etikk, men da kun i form av post hoc-rasjonalisering. Det vil si at deontologen kan møte sin del av bevisbyrden og si at «her er en deontologisk dom som kan vise til system 2-aktivitet på lik linje med en helt ordinær konsekvensialistisk- eller ikke-moralsk dom», uten at dette vil imponere Greene. I et slikt tilfelle vil system 2-aktiviten skyldes et forsøk på post-hoc å rettferdiggjør en dom som i utgangspunktet er produsert av ubevisste prosesser i system 1. Slik han ser det er tross deontologi et naturlig uttrykk for våre dypeste moralske emosjoner.[69] Greene mener den påfallende korrelasjonen mellom deontologisk teori og våre moralske emosjoner er bevis for nettopp dette. Det finnes for eksempel en komplisert og abstrakt teori om rettigheter som tilsier at det er ok å ofre de fem i *switch*, men ikke i *footbridge*, og det har seg *tilfeldigvis* slik at vi har sterke negative impulser forbundet til sistnevnte, men ikke førstnevnte. Videre finnes det en teori om plikter som tilsier at vi har en plikt til å redde barnet som drukner i kulpen i Peter Singers berømte eksempel, men ingen tilsvarende plikt til å redde sultende barn i den tredje verden, og *tilfeldigvis* har det seg slik at vi har sterke emosjonelle responser til førstnevnte, men ikke til sistnevnte. På samme måte forbyr det kategoriske imperativ ifølge

[68] Greene 2008: 59
[69] Greene 2008: 63

Kant masturbering fordi det involverer å bruke seg selv som et middel, og *tilfeldigvis* fant Kant masturbering å være frastøtende.[70]

Greenes argument inneholder med andre ord ikke bare en korrelasjon mellom hjerneaktivitet forenelig med affektive prosesser og deontologiske dommer, men også en filosofihistorisk korrelasjon mellom deontologisk teori og moralske dommer som er intuitivt tilfredsstillende. I den sammenhengen er det verdt å påpeke at det finnes filosofihistoriske moteksempel. Deontologiske teorier kan også produsere kontraintuitive resultater. Blant de mest berømte er morderen på dørstokken, som Kant hevdet at det kategoriske imperativ forbyr oss å lyve til, til tross for at det vil innebære at morderen finner og dreper vennen din som har søkt tilflukt hos deg.[71]

Det kan være fristende å se til slike moteksempel, slå seg på brystet og proklamere at det tross alt er mulig å ta utgangspunkt i deontologisk teori og utlede moralske dommer som ikke bare er uavhengige av system 1s moralske emosjoner, men faktisk er stikk i strid med deres budskap. Likevel finnes det ressurser for et motargument[72]: Selv om deontologisk etikk er et naturlig uttrykk for våre moralske emosjoner så må den være selektiv med tanke på hvilke emosjoner og intuisjoner den ønsker å akkomodere. Moralske emosjoner eksisterer i forbindelse- og i relasjon til scenarier som påkrever en moralsk vurdering, og som den filosofiske litteraturen er et prov på er en uendelig rekke slike scenarier tilgjengelig for en kreativ sjel. Følgelig er det svært urealistisk å ta utgangspunkt i enkelte moralske emosjoner, knyttet til konkrete saker, rasjonalisere fram et moralfilosofisk rammeverk som bygger opp under disse, for så å forvente at det samme rammeverket skal støtte opp under system 1s moralske emosjoner i alle tenkelige situasjoner.

Selv om Kant utviser intellektuell integritet i sin lojalitet til det kategoriske imperativs direktiver så er det ting som tyder på at post hoc-rasjonalisering er et fenomen heller ikke fagfilosofer er immune mot. I et eksperiment utført av Eric Schwitzgebel og Fiery Cushman

[70] Greene 2008: 68
[71] Greene 2008: 66
[72] Greene velger ikke denne strategien, men fokuserer heller på at det nettopp er slike kontraintuitive resultater moderne kantianere har forsøkt å bortforklare

ble moralfilosofer og lekfolk konfrontert med moralske dilemma slik som *footbridge* og *switch.* Målet var å undersøke om begge grupper i like stor grad ble påvirket av såkalte *order effects*. Teorien er, slik Greene beskriver det, at når du konfronteres med *footbridge* først, så opplever du en sterk negativ impuls som sier «nei». Du sier derfor nei til offeret i *footbridge*. Når du så får *switch* opplever du ingen slike impulser. Du prosesserer det heller på en nøytral måte, og selv om det kanskje tilsvarer å si «ja» til offeret, fordi fem liv er mer enn ett liv, så innser du at det ikke er noen moralsk relevant forskjell mellom de to scenariene, og du velger derfor å være konsistent og sier følgelig «nei» til offeret også i *switch*. Dersom du derimot blir presentert med *switch* først opplever du ingen sterke negative impulser, og dømmer derfor offeret som tillatelig. Når du så blir presentert med *footbridge* opplever du igjen sterke negative impulser, og selv du er klar over at du for å være konsistent bør si «ja» til offeret også her, blir denne betraktningen trumfet av de sterke negative impulsene du opplever, og du sier derfor «nei» til offeret.[73]

Denne tendensen ble også bekreftet. Når testsubjektene fikk høre *footbridge* først hadde de større sannsynlighet for å dømme offeret som utillatelig i både *footbridge* og *switch*. Når de derimot fikk høre *switch* først hadde de større sannsynlighet for å dømme offeret som tillatelig i *switch*, men ikke i *footbridge* – helt i tråd med den generelle tendensen for denne type moralske dilemma. Dette viste seg også gjeldene for ekspertene så vel som lekfolk.[74] Begge gruppene ble så spurt om de anerkjente doktrinen om dobbelteffekten, som støtter det generelle mønsteret hvor en sier «ja» til *switch* og «nei» til *footbridge*. For ikke-filosofene hadde det ingen effekt hvorvidt de fikk det ene eller det andre scenariet først, men filosofene hadde 50% høyere sannsynlighet for å anerkjenne doktrinen om dobbelteffekten dersom de ble konfrontert med *switch* først og derfor hadde avgitt de dommene som foreskrives av doktrinen om dobbelteffekten. Dette får Greene til å konkludere med at:

> «*Their experiment shows that philosophers are different from lay moralists and that they do indeed think harder (…)* the *folk are happy to let "popular prejudice" be "popular prejudice," but philosophers are motivated to translate that popular prejudice into principle.*»[75]

[73] Greene 2014: 720
[74] Schwitzgebel og Cushman 2012: 141
[75] Greene 2014: 720

Det faktum at deontologisk system 2-aktivitet kan bortforklares som post hoc-rasjonalisering på denne måten gjør at man sitter igjen med en bevisbyrde ingen av partene i debatten virker å være i stand til å bære i sin helhet. Greene kan vise til sine eksperimentelle resultater, mens deontologen kan hevde at hans testsubjekter er dårlige representanter for deontologien og at det virkelig må finnes unntak i de store deontologiske tenkerne. Men til og med om deontologen kunne dokumentere dette gjennom sin egen fMRI-studie kan Greene igjen trekke på skuldrene og si at annet ikke er å forvente, men at det kun er snakk om post hoc-rasjonalisering. Jeg kommer tilbake til akkurat dette, men først skal jeg dvele mer med studien til Schwitzgebel og Cushman.

3.2. Refleksjon

Deres studie viser nemlig ikke bare at fagfilosofer bedriver post hoc-rasjonalisering på lik linje med lekfolk. Den viser også at heller ikke fagfilosofers moralske vurderinger er immune mot faktorer som ikke er moralsk relevante. Selv om studiepoeng eller fagfilosofisk kompetanse i seg selv ikke er tilstrekkelig for å herde våre moralske vurderinger mot moralsk irrelevante faktorer så er det med det ikke gitt at det ikke finnes andre faktorer som kan bidra til å gjøre nettopp denne jobben. For å forstå nøyaktig hva dette kan være vil det være gunstig i å ta utgangspunkt i en større studie av Amos Tversky og Daniel Kahnemans, og især deres *asian disease*-eksempel. Dette er en del av en serie eksempler som illustrer hvordan vi som beslutningstakere er sårbare for *framing effects*. I studien ble testdeltakere fortalt at USA forbereder seg på et sykdomsutbrudd som er forventet å ta livet av 600 mennesker. Det finne to forskjellige måter å bekjempe utbruddet på: Dersom program A velges vil 200 mennesker bli reddet, men dersom program B velges er det 1/3 sjanse for at 600 mennesker vil reddes, mens det er 2/3 sjanse for at ingen mennesker vil bli reddet. Under denne formuleringen vektlegges det hvor mange som vil bli reddet av hvert program. Da valgte 72% programmet med minst risiko, program A, mens 28% valgte det med mest risiko, program B. En annen gruppe testdeltakere fikk derimot presentert alternativene formulert i form av dødstallene. Program A* ville da bety at 400 mennesker mister livet, mens program B* innebærer en 1/3 sjanse for at ingen vil dø, og en 2/3 sjanse for at 600 mennesker vil dø. Når antall døde mennesker – det potensielle tapet – ble vektlagt viste det seg at kun 22% valgte det sikre alternativet, program A*, mens 78% nå valgte alternativet med høyest risiko, program B*.[76]

[76] Tversky og Kahneman 1981: 260

Denne formen for *framing effects* har mye til felles med den formen for *order effects* jeg allerede har tatt opp. Likeledes har den også mye til felles med det originale *trolley problem: B* og B* er identiske. Akkurat som A og A*. Det eneste som skiller de er måten de er formulert på; om de formuleres i form av det man har å tape eller det man har å tjene. Om man godtar at forskjellen i formuleringen ikke er moralsk relevant er det heller ikke vanskelig å se parallellene til *the trolley problem*: Akkurat som *asian disease* og *asian disesase** er ikke *switch* og *footbridge* forskjellige i et moralsk henseende, men til tross for dette så vurderer vi de annerledes. Enkelte har vel og merke argumentert for at sistnevnte er forskjellige i et moralsk henseende. Doktrinen om dobbelteffekten har da vært den vanligste forklaringen, men den har igjen mistet mye av sin kredibilitet som en konsekvens av Thomsons berømmelige *loop*-variant av *the trolley problem*[77]. I mangel på gode alternativer virker det derfor å være rimelig å anta, slik også Greene gjør, at *switch* og *footbridge* er like i et moralsk henseende.

Poenget med å sammenligne disse scenariene fra Tversky og Kahnemans studie med *switch* og *footbridge* er at det er gjort mye forskning på hvordan effekten av *framing effects* i Tversky og Kahnemans studie kan begrenses. For eksempel har *Kazuhisa* Takemura utført en rekke eksperimenter med utgangspunkt i Tversky og Kahnemans *asian disease*-eksempel, for å se på hvordan denne formen for *framing effects* kan motvirkes. Hans hypotese var at effekten av *framing effects* ville være begrenset om testsubjektene la ekstra tid og omtanke i problemene de ble konfrontert med. I et eksperiment ble for eksempel halvparten av testsubjektene instruert til å tenke over hvordan de ville begrunne avgjørelsen deres. De ble også fortalt at de etter hver avgjørelse ville måtte skrive ned denne begrunnelsen. Den andre halvparten fikk ingen slike instruksjoner og valgte kun mellom to alternativer. Mens den sistnevnte gruppen bekreftet resultatene fra Tversky og Kahnemans eksperiment var ikke dette tilfelle for gruppen som ble bedt om å begrunne sin avgjørelse. Her spilte det ingen rolle hvordan alternativene var formulert, og testsubjektene fordelte seg jevnt på de forskjellige alternativene.[78] I et lignende eksperiment ble en del av deltakerne bedt om å tenke over problemet i 3 minutt, før de ga sin dom, mens den andre delen ble bedt om å tenke over

[77] Thomson 1985: 1402
[78] Takemura 1994: 36

problemet i 10 sekunder. Også her viste det seg at de som tenkte lenger, i motsetning til de som tenkte kortere, ikke i like stor grad ble påvirket av hvordan alternativene var formulert.[79]

På et lignende vis har det også blitt argumentert for at mennesker med en mer analytisk eller systematisk tenkestil er mindre sårbare for *framing effects*. Todd McElroy og John J. Seta tok i sin studie utgangpunkt i Zenhauserns preferansetest for å dele testsubjektene i to grupper: de som hadde en analytisk tenkestil, og de som hadde en holistisk- eller intuitiv tenkestil. I tråd med deres hypotese viste det seg at gruppen som foretrakk en analytisk tenkestil i mye mindre grad lot seg påvirke av måten alternativene var formulert på enn det som va tilfelle for den andre gruppen.[80] Lignende funn er også blitt gjort av Stephen M. Smith og Irwin P. Levin, som ikke fokuserte på subjekter med analytisk tenkestil, men heller hadde en «*natural tendency to engange in and enjoy thought*»[81] som de kalte «*need for cognition (NC)*»[82]. Testdeltakerne ble delt opp i to grupper ut ifra hvordan de graderte en rekke påstander, slik som: «*Thinking is not my idea of fun*»[83] og «*I really enjoy a task that involves coming up with new solutions to problems*»[84]. Som forventet viste resultatene at subjekter som ble definert som å ha høy NC tilla det mindre vekt om alternativene ble gjengitt i form av potensielt tap eller potensiell fortjeneste.[85]

Summen av disse eksperimentene må i det minste kunne sies å indikere at det finnes faktorer som kan begrense i hvilken utstrekning vår normative dømmekraft kan manipuleres. Selv om det ikke er snakk om en enhetlig faktor kan det være nyttig å bemerke at alle disse faktorene – å måtte begrunne sin dom, ta seg bedre tid til å avgi dom, foretrekke en analytisk tenkestil, høy *need for cognition* – har det til felles at de kan sies å stimulere til refleksjon. En god samlebetegnelse kan derfor være nettopp refleksjon. Dette er gunstig fordi refleksjon langt ifra er et fremmedlegeme i den moralfilosofiske litteraturen. Særlig David Ross, og senere Robert Audi, men også konsekvensialistiske tenkere som Henry Sidgwick, har vektlagt rollen refleksjon spiller innen etikken. Eksempelvis fremhever både Ross og Audi at vi kan begripe

[79] Takemura 1994: 38
[80] McElroy og Seta 2003: 614-615
[81] Smith og Levin 1996: 284
[82] Smith og Levin 1996: 284
[83] Smith og Levin 1996: 285
[84] Smith og Levin 1996: 285
[85] Smith og Levin 1996: 287

selvinnlysende moralske proposisjoner kun i den grad vi fullt ut forstår den aktuelle proposisjonen. Vi kan forstå proposisjonen gjennom refleksjon, og jo mer kompleks proposisjonen er, jo mer krevende er det å begripe den på en tilstrekkelig måte.[86] Referansen til Ross og Audi er viktig siden deres forståelse av refleksjon fremhever at refleksjon også er en del av den opprinnelige dannelsen av moralske vurderinger, og ikke kun en ekstern prosess som går moralske vurderinger i sømmene på jakt etter eventuelle feilslutninger. Det vil si at refleksjon ikke bare må være en post hoc-affære som retter på ellers feilaktige moralske slutninger, men også bidrar til at de feilaktige slutningene ikke oppstår i utgangspunktet.

3.3. Eksperiment X

Det er feil å gi forskjellige dommer i *asian disease* og *asian disease** fordi de i et moralsk henseende er like. Greene opererer med en antagelse om at det samme er tilfelle i *switch* og *footbridge*. Grunnen til vi feller ulike dommer i *footbridge og switch* er at vi gjennom vår evolusjonshistorie er betinget til å reagere på moralsk ikke-relevante faktorer som er til stede i *footbridge* men ikke i *switch. I asian disease-eksempelet* viste det seg at effekten måten handlingsalternativene ble formulert på hadde på dommene som ble avsagt ble eliminert eller kraftig redusert når testsubjektene ble stimulert til mer refleksjon forut for domsavsigelsen. Derfor er det betimelig å spørre om hva som ville blitt resultatet om testsubjektene i Greenes eksperiment fikk de samme forutsetningene. En gjennomgang av de aktuelle resultatene er avslørende:

A) Absolutt ingenting endrer seg fra Greenes originale eksperiment. Dette vil naturligvis styrke Greenes posisjon.
B) System 2-aktiviteten øker også når *footbridge* vurderes. De fleste dømmer likevel offeret som utillatelig. Greene kan likevel hevde at den opprinnelige dommen er produsert av system 1. Den økte system 2-aktiviteten skyldes post hoc-rasjonalisering. Greenes posisjon står like støtt som tidligere.
C) System 2-aktiviteten øker også når *footbridge* vurderes. De fleste dømmer nå offeret som tillatelig. Greene kan anse dette som den endelige seieren for hans teori. Når vi tvinges over i system 2-tenkning innser vi vanviddet i deontologien og konverterer til konsekvensialismen.

[86] Audi 2004: 35

Dette viser hvordan Greenes teori tilsynelatende er umulig å falsifisere. Uavhengig av hvilke eksperimentelle resultater man ser for seg så virker de å bekrefte Greenes teori. Men dette er kun én måte å tolke de fiktive eksperimentelle resultatene på. Jeg lanserer derfor denne alternative tolkningen:

> B*) System 2-aktiviteten øker også når *footbridge* vurderes. De fleste dømmer likevel offeret som utillatelig. Refleksjon har fått subjektene til å se bort fra faktorene som ikke er moralsk relevante og som ellers ville aktivert system 1-impulsene. Etter en tids refleksjon har de likevel kommet fram til at offeret ikke er tillatelig – enten i kombinasjon med at offeret i *switch* heller ikke er tillatelig, eller fordi det finnes andre faktorer, faktorer som er moralsk relevante, som skiller *switch* og *footbridge*.
>
> C*) System 2-aktiviteten øker også når *footbridge* vurderes. De fleste dømmer nå offeret som tillatelig, men mange bygger denne dommen på deontologiske vurderinger.
>
> D*) System 2-aktiviteten øker også når *footbridge* vurderes. Refleksjonen over *footbridge* kaster også nytt lys over *switch*, og de fleste subjektene vurderer det nå som utillatelig å ofre den ene for å redde de fem i begge scenariene. Det er tross alt denne konklusjonen Judith Thomson – som har reflektert mer over *the trolley problem* enn de aller fleste – kom fram til etter flere tiår.[87]

Det store spørsmålet blir så: Hvilke grunner har vi til å foretrekke Greenes tolkning over denne – eller snarere; finnes det en grunn til å foretrekke den ene over den andre?

Dette fiktive eksperimentet er et godt eksempel på hvordan de metodiske utfordringene hefter denne debatten. For alt positivt som kan sies om fremveksten av moralpsykologi og den kunnskapen dette har brakt med seg så må det fortsatt påpekes at det er en ung disiplin som fortsatt har sine utfordringer. I konteksten av denne teksten er det særlig to utfordringer som er særlig presserende. For det første er det som jeg har vært inne på et problem at vi mangler en klinisk metode for å skille genuin system 2-aktivitet fra post hoc-rasjonalisering. For det andre mangler vi en metode for å knytte testsubjekter og deres moralske dommer direkte til normative teorier. En nærmere analyse av disse problemene i forbindelse med mitt fiktive

[87] Thomson 2008

eksperiment vil avsløre at de ikke bare er gjeldende her, men også i Greenes originale eksperiment og påfølgende argument.

Alternativ A er uproblematisk og ikke særlig interessant. Et slikt resultat vil indikere at refleksjon ikke har noen effekt. Dette strider dog med de empiriske resultatene jeg allerede har presentert, og jeg vil derfor ikke fokusere mer på dette alternativet. Alternativ D* er heller ikke særlig kontroversielt. Om det forutsettes at offeret i *switch* faktisk er den mest nyttemaksimerende handlingen kan det ikke tolkes som en konsekvensialistisk dom å dømme offeret i *switch* som utillatelig. Dette kan sees på som et stort fremskritt i diskusjonen, siden refleksjon i dette tilfeller gjør dommene i begge scenariene deontologiske. Problemet med dette alternativet er at det står i sterk kontrast til Greenes originale eksperiment, hvor testsubjektene tross alt dømte offeret som tillatelig, samtidig som hjerneaktiviteten deres allerede viste system 2-aktivitet tilsvarende det som er vanlig når ikke-moralske problem behandles. Derfor virker heller ikke dette alternativet veldig sannsynlig.

Da virker derimot alternativ B/B* mer sannsynlig. Selv om vi per dags dato mangler eksperimentelle metoder for å påvise at en gitt system 2-aktivitet faktisk er genuin refleksjon så finnes det mer pragmatiske metoder for å fastslå når genuin refleksjon i det minste ikke er mulig. Som en minimumsbetingelse kan man si at dersom genuin refleksjon rundt et konkret moralsk problem skal være aktuelt forutsetter det en viss forståelse av det aktuelle problemet. Dette kan for eksempel innebære en viss kjennskap til, eller i det minste evnen til å gjenkjenne, moralske prinsipper eller distinksjoner som er relevant for hvordan dom som avgis i det aktuelle problemet. Dersom en slik grunnleggende forståelse ikke foreligger kan man også betvile at genuin refleksjon rundt det moralske problemet er mulig. Man kan med andre ord ikke peke på en konkret psykologisk prosess i forbindelse med et moralsk problem, og si at det nødvendigvis dreier seg om genuin refleksjon, men man kan i det minste peke på moralske problem hvor genuin refleksjon virker usannsynlig.

Det er nettopp dette Fiery Cushman, Liane Young og Marc Hauser undersøkte i sitt eksperiment. De ønsket å undersøke hvorvidt forskjellige moralske prinsipp i det hele og det store er tilgjengelig for bevisst refleksjon. De fokuserte på tre mer eller mindre velkjente

prinsipper, som de kalte: (1) Handlingsprinsippet, som tilsier at skade forårsaket av en aktiv handling er verre enn skade forårsaket av handlingsunnlatelse; (2) intensjonsprinsippet[88], som tilsier at skade ment som et nødvendig middel for å oppnå et høyere mål er verre enn skade som kun er en forutsett bieffekt av å nå et høyere mål; (3) kontaktprinsippet, som tilsier at skade forårsaket av fysisk kontakt med offeret er verre enn skade som ikke innebærer fysisk kontakt med offeret.[89]

Subjektene ble så konfrontert med en rekke moralske dilemma som var ordnet slik at et inneholdt en av disse faktorene, mens et annet ikke gjorde det. I et dilemma ble for eksempel skaden voldt ved at en person dyttet en annen av en gangbro, slik som i *footbridge*, mens i et annet dilemma kunne den samme skaden påføres ved å trekke i en spake som utløste en fallem i gangbroen. Lignende dilemma-par ble konstruert for de andre prinsippene. I de tilfellene hvor subjektenes vurderinger samsvarte med prinsippene (de dømte handling som verre enn ikke-handling, fysisk kontakt verre enn ikke-kontakt og skade som middel til et mål verre enn skade som en bieffekt) ble subjektene presentert med ordlyden i de to dilemmaene side om side, minnet om hvordan de hadde vurdert begge dilemmaene, og så bedt om å forklare hvorfor de behandlet de to dilemmaene forskjellig.[90]

Tanken bak dette designet er at om subjektene gjør bruk av et prinsipp uten å være i stand til i det minste å gjenkjenne det så er det vanskelig å se for seg at det aktuelle prinsippet skal være tilgjengelig for bevisst refleksjon. Resultatene viste at de aller fleste var i stand til å rettferdiggjøre de vurderingene som omhandlet handlingsprinsippet. Når det kom til intensjonsprinsippet var det under en tredel av subjektene som var i stand til å rettferdiggjøre deres vurderinger. 60% av subjektene var i stand til å rettferdiggjøre sine vurderinger vedrørende kontaktprinsippet, selv om 13% av disse mente at prinsippet likevel ikke var moralsk relevant.[91] Her er det likevel viktig å påpeke at selv om mange er i stand til å gjenkjenne et konkret moralsk prinsipp så betyr ikke det at de nødvendigvis er i stand til å

[88] Også kjent som doktrinen om dobbelteffekten
[89] Cushman, Young og Hauser 2006: 1083
[90] Cushman, Young og Hauser 2006: 1083
[91] Cushman, Young og Hauser 2006: 1086-1087

bedrive genuin refleksjon rundt det aktuelle prinsippet. Muligheten for at rettferdiggjørelsen av prinsippet er en post hoc-rasjonalisering er fortsatt til stede.

Det dette eksperimentet derimot kan si noe om er hvilke prinsipper som tilsynelatende ikke er tilgjengelig for genuin refleksjon. Intensjonsprinsippet, sammen med graden av personlighet i den aktuelle skaden, som kontaktprinsippet i det minste er en del av, er faktorene som ifølge Greene utløser system 1-impulsene i *the trolley problem*. Derfor kan det sees på som problematisk at så få er i stand til å rettferdiggjøre intensjonsprinsippet, for om dette prinsippet ikke er åpent for bevisst refleksjon, hvordan kan det forventes at refleksjon skal motvirke feilslutninger i *the trolley problem*, hvor intensjonsprinsippet tross alt later til å spille en sentral rolle? Samtidig må man spørre seg om hvor mye man skal forvente. At én av tre gjenkjenner intensjonsprinsippet og én av to gjenkjenner kontaktprinsippet er tross alt ikke så galt. Ingen har sagt at moralfilosofi skal være lett, og som Ross har påpekt kan selvinnlysende moralske proposisjoner være komplekse og vanskelige å gripe, og noen krever sågar refleksjon tilsvarende et helt livsløp.[92]

Dette kan i beste fall sies å være en halvgod løsning på kontroversen rundt post hoc-rasjonalisering og genuin refleksjon, og det er selvfølgelig problematisk at det ikke finnes en bedre løsning all den tid det er muligheten for genuin refleksjon som aktualiserer «eksperiment x» og den alternative tolkningen av de eksperimentelle resultatene. Det positive er likevel at denne problematikken er mindre når det gjelder det mest sannsynlige resultatet av «eksperiment x». Alternativ C/C* tilsier nemlig at offeret dømmes som tillatelig i både *footbridge* og *switch*. I Greenes originale eksperiment ble til sammenligning offeret i *footbridge* som regel vurdert som utillatelig, og siden det samme offeret dømmes som tillatelig i alternativ C/C* i «eksperiment x» kan det umulig være snakk om en post hoc-rasjonalisering i dette tilfellet. Det kan fortsatt diskuteres om det er relevant nøyaktig hvor i prosessen hvor den moralske dommen dannes den aktuelle refleksjonen finner sted, slik at refleksjonen fortsatt er post hoc, selv om det ikke kan sies å være en rasjonalisering. Om refleksjon er noe som kommer etter at dommen er felt for å søke etter feilslutninger i rettferdiggjørelsen av dommen, eller om refleksjon inngår på et tidligere stadium hvor de relevante faktorene identifiseres og de forskjellige hensynene veies opp mot hverandre. Jeg

[92] Ross 2002: 29

tror begge disse funksjonene er mulige, og sannsynligvis er det vanskelig å skille klart mellom disse. Det skyldes i stor grad at det er vanskelig å avgjøre nøyaktig når en moralsk dom dannes. Jo mer kompleks problemet er jo mer kompleks vil naturligvis dommen være, men i mange tilfeller virker det klart at en moralsk dom ikke kan stadfestes til et konkret tidspunkt, men heller må sees på som en dialektisk prosess hvor forskjellig hensyn veies mot hverandre helt til en konklusjon kan trekkes. I så måte virker den rimeligste konklusjonen å være at dannelsen av en moralsk dom må ha en start og en slutt, og at refleksjon slik jeg har skissert det må finne sted et sted mellom disse to punktene. I denne sammenhengen virker uansett det aller mest sentrale å være at den aktuelle dommen ikke er den samme før og etter refleksjon innføres som en faktor, og følgelig at vi ikke har med en post hoc-rasjonalisering å gjøre.

3.4. Karakteristisk deontologiske dommer

Det er nå på tide å forklare hvorfor alternativ C/C* er det mest sannsynlige resultatet av «eksperiment x». Jeg har argumentert for, eller rettere sagt vist til forskning som tyder på, at det finnes faktorer som reduserer sannsynligheten for at våre normative vurderinger lar seg påvirke av faktorer som ikke er moralsk relevante. Men det finnes også eksperimentelle resultater som tyder på at de samme faktorene også øker sannsynligheten for det som av Greene karakteriseres som konsekvensialistiske dommer. I et eksperiment utført av Renata S. Suter og Ralph Hertwig viste det seg for eksempel at personer som ble konfrontert med moralske dilemma ga mer deontologiske svar når responstiden ble begrenset, mens de ga mer konsekvensialistiske svar når responstiden ikke ble begrenset.[93] Siden testsubjekter som får bedre tid på å avgi sin dom har større sannsynlighet for å dømme det som tillatelig å ofre én for å redde fem i personlige moralske dilemma, slik som *footbridge*, så er det sannsynlig at alternativ C/C* ville vært resultatet av «eksperiment x». Selv om Suter og Hertwig konkluderer med at lengre betenkningstid fører til mer konsekvensialistiske vurderinger så betyr det likevel ikke at C bør foretrekkes over C*.

Problemet er at denne studien heller ikke har en sikker måte å knytte moralske dommer til normative teorier på, annet enn å vise til karakteristisk deontologiske dommer og karakteristisk konsekvensialistiske dommer. Faktisk bygger studien tungt på Greenes originale forskning, og benytter ikke bare mange av de samme dilemmaene, men definerer

[93] Suter og Hertwig 2011: 456

også deontologiske- og konsekvensialistiske dommer i henhold til deres funksjon i personlige- og ikke-personlige moralske dilemma. At også denne studien baserer seg på Greenes karakteristiske dommer på denne måten understreker mangelen på gode alternativ. Det er også verdt å merke seg at når jeg i forrige kapittel viste til eksterne bekreftelser på Greenes teori, slik som at enkelte psykopater og andre med skader på deler av hjernen som forbindes med system 1 i større grad enn andre gjør konsekvensialistiske vurderinger, så siktes det også her til karakteristisk konsekvensialistiske dommer. Derfor er det rimelig å si at rollen de karakteristiske dommene spiller i Greenes prosjekt knapt kan overdrives.

Forrige avsnitt avslører at C/C* er det mest trolige resultatet av mitt fiktive eksperiment. Samtidig avslører det også en av de største utfordringene for Greenes prosjekt og hvorfor det er så vanskelig å avgjøre hva som er den beste tolkningen av C og C*. Til tross for at Greenes uttalte studieobjekt er normative teorier, i dette tilfellet deontologi og konsekvensialisme, opererer han altså med det han kaller karakteristiske deontologiske- og karakteristisk konsekvensialistiske dommer. Grunnen til dette er at normative teorier er immaterielle og konstruerte størrelser som i seg selv ikke kan knyttes direkte til forskjellige prosesser i- eller deler av hjernen. Derfor definerer han de ut ifra deres funksjon som han igjen utleder fra den grunnleggende praktiske uenigheten som *the trolley problem* er tuftet på. Konsekvensialister *er* vanligvis mer villige til å ofre individuelle rettigheter om det er til det felles beste, og deontologer er vanligvis mindre villige til å gjøre det samme offeret. Selv om det ikke nødvendigvis må være slik, er det det som følger mest naturlig av de respektive teoriene uten at de må legge til mye *fancy filosofering* for å nå sin posisjon, ifølge Greene.[94] Det nærmeste han kommer å kunne legge deontologisk etikk på disseksjonsbordet i sitt laboratorium er med andre ord å studere hjerneaktiviteten bak karakteristisk deontologiske dommer.

Det som er interessant er at det Greene kaller *fancy filosofering* også kan være det jeg har kalt refleksjon. Jeg har argumentert for at refleksjon har en viktig rolle å spille innen deontologisk etikk, og det faktum at Greene eksplisitt ekskluderer en slik faktor fra den deontologien han kritiserer sår tvil om hvorvidt Greene faktisk kritiserer den sterkest mulige utgaven av deontologisk etikk. Greenes karakteristiske dommer er altså en løsning på et metodisk problem, men spørsmålet er om det er en god løsning. For hva er egentlig forbindelsen

[94] Greene 2008: 39

mellom Greenes karakteristiske dommer og deres respektive opphavs-teorier? En karakteristisk konsekvensialistisk dom er å dømme det som tillatelig å ofre den ene personen i *switch*. Det kan argumenteres for at en slik dom er i overenstemmelse med konsekvensialisme, men det må likevel tas visse forbehold. At en slik dom er i overenstemmelse med konsekvensialisme betyr ikke at den ikke samtidig kan være i overenstemmelse med mange deontologiske teorier. Det betyr heller ikke at en konsekvensialistisk begrunnelse nødvendigvis må være den beste begrunnelsen for hvorfor denne handlingen er tillatelig. C er med andre ord den mest naturlige tolkningen fordi det er enklest å forsvare denne dommen i konsekvensialistiske termer. Det betyr likevel ikke at det er det eneste eller en gang det beste forsvaret av denne dommen. For det første er det konsekvensialistiske vokabularet tilgjengelig for deontologien så lenge den holder fast ved at det finnes minimum ett tilfelle hvor det trumfes av andre hensyn, og for det andre er det også mulig å forsvare denne dommen ved hjelp av et eksklusivt deontologisk vokabular. Derfor bør det nå være klart at det i utgangspunktet er et åpent spørsmål hvorvidt man bør foretrekke C eller C*.

Det er her min kritikk av Greene møter mye av den øvrige kritikken hans prosjekt har blitt utsatt for, for hans karakteristisk deontologiske dommer er utvilsomt en av de mest kontroversielle delene av prosjektet hans. I artikkelen *Methodological Issus in the Neuroscience of Moral Judgement* poengterer for eksempel Guy Kahane og Nicholas Shackel at på en lignende måte at det skal mer til enn at et subjekt vurderer *footbridge* og *switch* i overenstemmelse med konsekvensialistisk teori for at vedkommende skal kunne regnes som konsekvensialist, nettopp fordi en slik vurdering også er i overenstemmelse med mange deontologiske teorier.[95] I *Neuroscience and Moral Reasoning: A Note on Rescent Research* skriver også Frances Kamm nøyaktig det samme: «*the judgment that it is permissible to turn the trolley may as likely be a deontological response as a consequentialist response.*»[96]

Kamm og Kahane og Shackel mener disse problemene bør svekke vår tiltro til Greenes prosjekt. Riktig nok har også jeg argumentert for at C* er en like plausibel tolkning som C, men det er likevel et spørsmål om det ikke er mulig å bygge en bro over dette problemet. For

[95] Kahane og Shackel 2010: 574-575
[96] Kamm 2009: 337

her ønsker jeg å spekulere i om det ikke nok en gang er mulig for Greene å skyve bevisbyrden foran seg. Problemet er at det på dette tidspunktet ikke lenger er like tydelig at bevisbyrden hviler på deontologen. Om en part i en diskusjon skal være rettferdiggjort i å skyve bevisbyrden foran seg på denne måten ved å si at «da er det opp til deg å argumentere for hvorfor det ikke har seg slik» så fordrer det i det aller minste at vedkommende selv har sannsynliggjort at det er nettopp slik. Men med tanke på de metodiske problemene som så langt er blitt belyst er det ikke nødvendigvis like klart at Greene har gjort dette i tilstrekkelig grad.

Man kan se for seg at Greene i utgangspunktet vil argumentere på den følgende måten. Fordi deontologisk etikk er et produkt av ubevisste prosesser i system 1 kan ikke økt system 2-aktivitet i forbindelse med aksepten av å ofre den ene i *footbridge* tolkes som C*. Dette er fordi C* forstår dommen som en deontologisk dom, men en deontologisk dom må nødvendigvis må være et produkt av system 1. Derfor er C en bedre tolkning av «eksperiment x». Men når muligheten for at C* er et fullgodt alternativ til C reises nettopp for å illustrere muligheten for at deontologisk etikk ikke nødvendigvis må være et produkt av ubevisste prosesser i system 1 så vil denne argumentasjonsrekken være utilstrekkelig siden den unngår det sentrale spørsmålet om hvorfor C* ikke er et fullgodt alternativ til C kun ved å anta at den endelige konklusjonen er sann.

Men når jeg argumenterer for at et slikt argument blir sirkulært kan det igjen innvendes fra Greene at jeg overser hans eksperimentelle resultater som selvstendig bevis for konklusjonen som igjen støtter opp om C som den beste tolkningen. For selv om problemet med de karakteristisk deontologiske dommene betyr at han ikke kan påstå å ha påvist en absolutt og nødvendig sammenheng mellom system 1 og deontologisk etikk så kan han påstå å ha påvist en sammenheng mellom system 1 og moralske dommer som i det minste minner veldig om deontologiske dommer og som de aller fleste vil klassifisere som deontologiske dersom de måtte velge. For selv om dette i seg selv kan betraktes som en liten seier for deontologien så er det likevel deontologens byrde å presentere en forklaring på dette, dersom C* skal være et fullgodt alternativ til C. Hvorfor er det slik at moralske vurderinger som vanligvis er mer appellerende for deontologer kan kobles til høyere system 1-aktivitet enn moralske vurderinger som vanligvis er mer appellerende for deontologer?

Kahane og hans kolleger har for eksempel foreslått at det er en tilfeldighet og at grunnen til at karakteristisk konsekvensialistiske dommer så langt har blitt koblet til system 2 er at man så langt har fokusert på situasjoner hvor karakteristisk konsekvensialistiske dommer faktisk er støttet av system 2. Deres hypotese var at dersom man så på andre typer dilemma så ville man også finne tilfeller hvor karakteristisk deontologiske dommer kunne kobles til høyere system 2-aktivitet, på lik linje med karakteristisk konsekvensialistiske dommer.[97] Et eksempel på et slikt dilemma er en situasjon hvor du kan være ærlig mot vennen din ved å fortelle vedkommende sannheten om hva den slemme onkelen hans egentlig mener om han, selv om det vil gjøre han knust og følgelig ikke produsere de beste konsekvensene. Men som Greene siden har poengtert så var resultatene fra deres studie tvetydige, og de karakteristisk deontologiske dommene i dilemmaene til Kahane og hans kolleger ble faktisk gjort på enda kortere tid enn de karakteristisk deontologiske dommene i *footbridge*, i strid med deres teori.[98]

Dette viser at det er lettere sagt enn gjort å konstruere en alternativ forklaringsmodell for Greenes eksperimentelle resultater som ikke forutsetter en kausal relasjon mellom deontologien og system 1, og følgelig favoriserer C over C* som tolkning av «eksperiment x». I det følgende skal jeg likevel forsøke på nettopp dette. Jeg skal ikke vise til dilemmaene som er brukt, men heller strukturelle forskjeller mellom deontologi og konsekvensialisme, representert med distinksjonen mellom det jeg kaller rigide- og fleksible normative teorier.

3.5. Rigide- og fleksible normative teorier

I hans siste artikkel *The-rat-a-gorical-imperative* påpeker Greene at konsekvensialisme er en nokså rigid teori med en rigid struktur og at det nettopp er dette som har gjort den så sårbar for den type moteksemepel jeg refererte til i innledningen av denne oppgaven.[99] Han har selvfølgelig rett. Konsekvensialisme i sin enkleste form er en utrolig minimalistisk teori. Den trenger i teorien kun å ta utgangspunkt i disse to spørsmålene: 1) Hva er det som er relevant i et moralsk henseende? 2) Hva er det som defineres som godt? Under forutsetningen om at man har tilgang til all den ikke-moralske informasjonen, altså hva som faktisk vil maksimere

[97] Kahane et al. 2012: 395
[98] Greene 2014: 708
[99] Greene 2017: 10

det som defineres som godt, kan man med utgangspunkt i disse antakelsene felle en konkret dom i alle tenkelige moralske problem. Det hører med til denne historien at dette levner lite rom for tolkning, skjønn eller intuisjon. Dette spiller kanskje en viss rolle under utarbeidelsen av de opprinnelige antagelsene som må gjøres, men derifra er det en rent mekanisk prosess. Derfor er også de konsekvensialistiske dommene så ufravikelige og udiskutable. Du forer den konsekvensialistiske algoritmen med den ikke-moralske informasjonen på den ene siden og ut kommer dommen på den andre siden. Derfor er konsekvensialismen populær. Men derfor er konsekvensialismen også sårbar for moteksempel. Fordi det er lett å konstruere scenario hvor dommen virker kontra-intuitiv og teorien i seg selv er så rigid at det er umulig å bortforklare resultatene. Men på den annen side; derfor ser vi også de fMRI-resultatene vi ser i Greenes eksperiment.

For hva om man snur på det? Om konsekvensialisme er en mekanisk prosess vi også kan programmere en robot til å utføre er det ikke rart at system 2 er så dominerende når vi foretar karakteristisk konsekvensialistiske dommer. Deontologiske teorier, derimot, er som regel ikke like rigide. David Ross opererer for eksempel med åtte *prima facie* plikter. Blant disse finnes det for eksempel en plikt til å forbedre andres situasjon[100], ikke så ulik konsekvensialismens grunnleggende doktrine, samt en plikt til ikke å skade andre.[101] Man kan se for seg *the trolley problem*, eller et hvilket som helst annet moralsk problem, som en test for å skille mellom rigide og fleksible normative teorier. En fleksibel teori er en teori som ikke er bundet med nødvendighet til en konkret dom, mens en rigid teori er nettopp det. Det kan selvfølgelig tenkes at for eksempel Ross, eller en annen deontologisk teoretiker, vil være veldig klar på nøyaktig hva deres teori vil kreve i møte med et konkret moralsk problem slik som *the trolley problem*, men poenget er at det for en som vurderer teorien utenfra ikke er åpenbart, men snarere et noe som kan diskuteres. Ross sier riktig nok at plikten til ikke å skade andre vanligvis vil være sterkere enn plikten til å forbedre andres situasjon, slik at vi som regel ikke bør se på det som tillatelig å drepe en person for å redde en annen. Samtidig påpeker han at også denne interne rangeringen kun er *prima facie*, og at det ikke finnes en hierarkisk ordning av de forskjellige pliktene.[102] For en som setter seg fore å tolke Ross er med andre ord, i alle fall i teorien, alle muligheter åpne med tanke på *the trolley problem*.

[100] Det Ross kaller *duty of beneficence*
[101] Ross 2002: 21
[102] Ross 2002: 22

Dette gjelder ikke bare Ross. For eksempel sier Kants humanitetsformulering at man skal handle slik at man bruker menneskeligheten, i seg selv så vel som i andre, ikke bare som et middel men alltid også som et mål i seg selv.[103] Dette bør binde Kant til å forkaste offeret i *footbridge*, og i så måte kvalifisere Kants etikk som rigid. Likevel har det blant kantianere vært uenighet om hva humanitetsformuleringen, og især også universaliseringsprinsippet, faktisk innebærer. Det må heller ikke glemmes at humanitetsformuleringen også er grunnen til at Kant mente å ha et teoretisk grunnlag for å fordømme masturbering. Om man sammenligner humanitetsformuleringen med konsekvensialismens prinsipp om nyttemaksimering er det langt fra like klart hvilke handlinger som faktisk bruker menneskeligheten i en selv ikke bare som et middel men også som et mål i seg selv. Kanskje viser dette at *the trolley problem* alene ikke er den beste indikatoren på hvilke teorier som kvalifiserer som rigide og ikke. For selv om Kants humanitetsformulering er rigid når det kommer til *footbridge* er det definitivt rom for en viss fleksibilitet i formuleringens anvendelse.

Slik jeg ser det er den vesentlige forskjellen mellom rigide- og fleksible normative teorier i denne sammenhengen at fleksible teorier har mye «ledig» eller «uokkupert» rom som rigide teorier ikke har. Med ledig rom mener jeg plass som kan fylles av intuitiv- og affektiv tenkning; system 1, kort fortalt. Man kan se på det på den følgende måten: Deontologisk etikk behøver nødvendigvis ikke være et resultat av de bakenforliggende prosessene i system 1. Kanskje er det mer presist å si at deontologisk etikk ofte akkompagneres av system 1, fordi deontologiske teorier er fleksible og har plass til å romme denne innflytelsen. Se for deg at du anvender din foretrukne deontologiske teori på et moralsk dilemma, men fordi teorien er vag – det vil si fleksibel – er det ikke klokkeklart hva som vil være teoriens endelige anbefaling. Det kan for eksempel avhenge av hvordan man forstår «menneskeligheten i en selv», eller en annen deontologisk formulering. Siden de teoretiske forpliktelsene på denne måten er uklare, men du likevel har dine moralske intuisjoner, så er det altså mulig for deg å tolke de teoretiske forpliktelsene på en slik måte at de kler de moralske intuisjonene du allerede har. Det er fullt mulig å se for seg at det er på dette stadiet den mest innflytelsesrike system 1-tenkningen, med sine intuisjoner og emosjoner, i hovedsak kommer inn i bildet.

[103] Kant 1998: 4:429

For alle ønsker i utgangspunktet at ens bevisste og deliberative holdninger skal være i overenstemmelse med ens følelser. Om jeg er sint på deg fordi du baksnakket meg med til felles bekjent så ønsker jeg (i det minst å tro) at det finnes et rasjonelt grunnlag for meg å ha denne følelsen. Om det føles forferdelig å skulle dytte en person ned fra gangbroen i *footbridge* ønsker jeg (i det minst å tro) at det finnes et rasjonelt grunnlag for meg å ha denne følelsen. På samme måte er det motsatte sant. Når jeg skal bedrive normativ etikk så ønsker jeg at mine normative konklusjoner er i overenstemmelse med mitt moralske følelsesliv – uavhengig av om jeg er konsekvensialist eller deontolog. Forskjellen er at konsekvensialisme er en rigid teori som ikke er i stand til å tilpasse seg intuisjoner og følelser slik som mer fleksible deontologiske teorier kan. Dette trenger ikke bety at deontologisk etikk i større grad enn konsekvensialistisk etikk er et direkte produkt av intuisjoner og emosjoner, kun at de har større rom til å tilpasse seg emosjoner og intuisjoner som de i utgangspunktet utsettes for på lik linje som konsekvensialistisk etikk.

Oppfølgingsspørsmålet blir så om deontologiske teorier også kan være rigide. I prinsippet er det ingen grunn til at de ikke kan være det. Likevel er det, som jeg har argumentert for, ikke normen. Det kan være flere grunner til det, men for å se at det i det minste er en prinsipiell mulighet kan man starte med å sammenligne med lovverket i et land. Loven er riktignok ikke et moralsk system i seg selv, og mye kan sies om relasjonen mellom lov og moral, men all den tid loven regulerer ikke bare hvordan vi bør opptre men også hvordan de som ikke følger loven bør behandles er den er like fullt et normativt system. Loven er selvfølgelig ikke fremmed for konsekvensialistiske betraktninger, men det er heller ikke de fleste deontologiske teorier, og i motsetning til hva som er tilfelle for konsekvensialistiske teorier er ikke nytte-maksimeringen lovens eneste- eller overordnede mål. Loven har mye til felles med deontologisk teori i den forstand at den opererer med klare forbud og påbud – uavhengig av hva som fører til de beste konsekvensene i den gitte situasjonen. Ifølge loven er det for eksempel forbudt å stjele og drepe, og dersom det finnes unntak fra det opprinnelige forbudet er dette tydelig formulert og regulert.

Den mest åpenbare grunnen til at de deontologiske teoriene vi opererer med i dag ikke har den samme presisjonen og detaljrikdommen moderne lovverk har er sannsynligvis at moralske

teorier som regel er enkeltmannsforetak mens eksempelvis norsk lov har et helt statsapparat bak seg. Men om vi satte oss denne bragden fore, og med tiden konstruerte «Norges offisielle nasjonal-deontologiske folkemoral (NONDF)», støpt i formen til-, og dermed like stringent som loven selv, så ville NONDF i teorien kunne avsagt en konkret og udiskutabel dom for alle tenkelige moralske problem. Det må dog sies at rettspraksisen også opererer med et visst rom for skjønnsmessige vurderinger, og det må til stadighet settes nye rettpresedenser. Likevel må loven sies å operere med et helt annet nivå av rigiditet enn fleksible deontologiske teorier.

En rigid deontologisk teori trenger heller ikke være så omfattende og ressurskrevende. Man kan også se for seg Ross' åtte *prima facie* plikter hvor de er nøye presisert og gitt en hierarkis ordning. Det øverste prinsippet skal alltid følges, og gjelder over alle andre. Det nest-øverste prinsippet skal alltid følges så fremst det ikke er i konflikt med prinsippet øverst. Slik følges lista helt til vi kommer til det laveste prinsippet, som kun skal følges dersom det ikke er i konflikt med noen av de øvrige prinsippene. Et slikt moralsk system vil kunne gi moralsk rådgivning i de fleste situasjoner som vil oppstå. Et samfunn tuftet på et slikt moralsk system vil sannsynligvis også være relativt velfungerende. Et slikt moralsk system vil være et deontologisk moralsk system, og det vil i enkelte scenarier produserer kontraintuitive resultater, men enda viktigere vil det være et rigid moralsk system og en representant for dette systemet vil neppe være i tvil om hvordan dets prinsipper skal anvendes i møte med et gitt moralsk problem.

En annen plausibel kandidat til å være en rigid deontologisk teori er, kanskje noe overraskende, regel-konsekvensialisme. En mulig formulering av regel-konsekvensialisme tilhører Brad Hooker: «*An act is wrong if it is forbidden by the code of rules whose internalization by the overwhelming majority of everyone in each new generation has maximum expected value in terms of well-being*»[104]. Merk for det første at dette ikke kan være en ren konsekvensialistisk teori, siden det finnes tilfeller hvor denne teorien vil hevde at den handlingen som maksimerer nytten ikke er den riktige handlingen. Grunnen til dette er at hvor ren konsekvensialisme er opptatt av de faktiske konsekvensene er regel-konsekvensialisme opptatt av de hypotetiske konsekvensene dersom en regel som tillater den konkrete

[104] Hooker 2000: 32

handlingen skal internaliseres, slik Hooker beskriver. Å ofre den ene i *footbridge* vil kanskje produsere de beste konsekvensene i den konkrete situasjonen, men om det ikke ville gjort det i en hypotetisk verden hvor de aller fleste internaliserte en regel som tillot dette så spiller det ingen rolle for regel-konsekvensialismen. Likevel har regel-konsekvensialisme mye til felles med ren handlingskonsekvensialisme, og det er nettopp derfor den kan klassifiseres som en rigid teori. Akkurat slik som med konsekvensialismen kan man se for seg en regel-konsekvensialistisk algoritme. Så lenge man har tilgang på all den relevante ikke-moralske informasjonen kan regel-konsekvensialismen foreskrive en ikke-forhandlingsbar moralsk dom i alle tenkelige scenarier. Den regel-konsekvensialistiske algoritmen tar vel og merke hensyn til flere faktorer, og må derfor være mer komplisert enn en standard handlingskonsekvensialistisk algoritme, men prinsippet om rigiditet er likevel det samme.

Distinksjonen mellom rigide- og fleksible normative teorier kan med andre forklare hvorfor ikke-konsekvensialistiske moralske dommer ofte akkompagneres av høy system 1-aktivitet, uten at det innebærer en kausal forbindelse mellom system 1 og deontologisk etikk. Siden Greenes eksperimentelle resultater da kan forklares på en måte som ikke forutsetter en nødvendig forbindelse mellom system og deontologisk etikk betyr det også at man kan trekke de karakteristiske rollene i tvil, og påstå at C* er et fullgodt alternativ til C, og si Greene nødvendigvis ikke kritiserer deontologien i sin beste form.

Enkelte vil kanskje reagere på den kunstige konstruksjonen av mine eksempler på potensielle rigide deontologiske teorier. De virker å gjøre mange normative antakelser som ikke nødvendigvis følger naturlig av hverandre, slik at selv om anvendelsen i seg selv kanskje er mekanisk og system 1-fri, så må likevel den opprinnelige erkjennelsen av prinsippene være basert på en eller annen form for intuitiv tenkning. Men her er det et poeng at dette også er et problem for konsekvensialistisk teori. For konsekvensialismen opererer også med sine grunnleggende normative antagelser, og også disse må erkjennes på en eller annen måte. Det er dette som vil være tema for neste kapittel. For dersom en slik innvendingen som dette avsnittet representerer skal være mulig å rette mot mine rigide deontologiske teorier så må konsekvensialisten på en eller annen måte argumentere for at det samme ikke er et problem for konsekvensialismens grunnleggende intuisjoner.

Kapittel 4: Finnes det relevante forskjeller mellom konsekvensialismens grunnleggende intuisjoner og deontologiens tilsvarende intuisjoner?

Forrige kapittel kulminerte med distinksjonen mellom rigide- og fleksible normative teorier: Selv om dual processing-theory er korrekt, og system 1 og system 2 representerer de bakenforliggende prosessene innen moralfilosofien, så trenger ikke deres funksjon å være representert ved deontologiske- og konsekvensialistiske teorier, men snarer ved rigide- og fleksible teorier. Derfor argumenterte jeg for at deontologiske teorier ikke nødvendigvis må være fleksible teorier. Et problem for disse deontologiske- og rigide teoriene jeg introduserte er at de virker å være kunstige konstruksjoner som er avhengig av å gjøre mange forskjellige normative antagelser. Selv om anvendelsen av teoriene kanskje er rigid og følgelig i stand til å begrense uønsket system 1-inflytelse, så kan ikke det samme sies om erkjennelsen av teorienes grunnleggende prinsipp og verdier. Et naturlig tilsvar for deontologien er dog å hevde at dette problemet også angår konsekvensialistisk teori. Dersom denne kritikken av deontologiens rigide teorier skal være tilgjengelig for konsekvensialismen må med andre ord konsekvensialismen forklare hvorfor dens grunnleggende intuisjoner stiller i en annen klasse enn deontologiens tilsvarende intuisjoner. Det er dette som er målet for inneværende kapittel; å undersøke om det finnes relevante forskjeller mellom de to respektive teorienes grunnleggende moralske intuisjoner. Jeg starter med å slå fast at dette faktisk er et felles problem, før jeg undersøker to forskjellige forklaringer, fra henholdsvis Henry Sidgwick og Joshua Greene, på hvorfor konsekvensialismens intuisjoner ikke er like problematiske som deontologiens intuisjoner. Jeg vil argumentere for at ingen av disse forklaringene i seg selv er tilstrekkelig for å gjøre denne jobben, men at en syntese mellom de to er i stand til det, og at dette faktisk gir deontologien et forklaringsproblem.

4.1. Et felles problem

En av kandidatene jeg i forrige kapitel lanserte som en potensiell deontologisk rigid normativ teori var en presis, nøye spesifisert, hierarkisk ordnet og eksplisitt ikke-*prima facie* modell av David Ross′ åtte *prima facie* plikter. Derfor er det passende å ta utgangspunkt i hans egent forsvar av sin teori mot lignende anklager. I *The Right And The Good* skriver han blant annet at der hvor hans egen teori fremstår usystematisk og uten et overordnet moralsk prinsipp så står det uansett ikke bedre til med rivaliserende teorier: «*The list of good put forward (…) is*

reached by exactly the same method—the only sound one in the circumstances—viz, that of direct reflection on what we really think.»[105] Det samme poenget kan også finne støtte i argumenter som forutsetter at *alle* moralske proporsjoner enten må ha sitt opphav i en eller annen form for grunnleggende moralsk intuisjon eller inngå i en evig regress.[106] Det sentrale ordet i den foregående setningen er «alle», siden dette er et problem som i utgangspunktet er like aktuelt for konsekvensialistisk tenkning som for deontologisk tenkningen. Dette er også grunnen til at jeg innledningsvis i denne oppgaven beskrev Greenes prosjekt som å ha potensial til å representere framgang innen normativ etikk – men samtidig være grobunn for skeptisisme.

Til tross for at konsekvensialisme er en minimalistisk teori som må gjøre svært få normative antakelser så er det et viktig poeng at den fortsatt må gjøre *noen* normative antagelser. For om det er et problem for en rigid deontologisk teori å forklare hvordan prinsippene som anvendes med slik en rigiditet samtidig kan erkjennes på en intuisjon- og system 1-fri måte så må det samme være tilfelle for konsekvensialistisk teori. Om man for eksempel skal svare på hva som er godt eller hva som er relevant i et moralsk henseende, som jo er sentrale spørsmål for enhver konsekvensialist, så er det vanskelig å se for seg hvordan dette kan gjøres uten en referanse til en eller annen form for intuisjon. At også konsekvensialismen på denne måten er avhengig av enkelte moralske intuisjoner er heller ikke et synspunkt som kun eksisterer blant deontologer og moralske skeptikere. For eksempel skriver Sidgwick i *The Method of Ethics*: «*the utilitarian method (…) could not (…) be made coherent and harmonious without (…) a fundamental intuition.*»[107] I en lignende tone skriver også Peter Singer: «*Even a radical ethical theory like utilitarianism (…) rest on a fundamental intuition about what is good.»*[108]

På bakgrunn av dette konkluderer jeg med at innvendingen som kan reises mot rigide, deontologiske teorier – at de minimerer system 1-innflytelsen under den praktiske anvendelsen av teorien kun ved å flytte system 1-innflytelsen til nivået for moralske førsteprinsipp – er et felles problem som i utgangspunktet rammer konsekvensialismen i like stor grad. Dersom kritikken jeg introduserte innledningsvis i dette kapitlet skal være

[105] Ross 2002: 23
[106] Sinnott-Armstrong 2008: 49
[107] Sidgwick 1907: xvi-xvii
[108] Singer 2005: 349

tilgjengelig for konsekvensialisten er det med andre ord konsekvensialistens byrde å forklare hvorfor den samme kritikken ikke også rammer konsekvensialistisk etikk.

Denne taktikken kan kanskje best betegnes som geriljafilosofi. Den gjør ingen selvstendige grep for å fremheve sitt eget alternativ og frikjenne det fra kritikk, men søker ganske enkelt å dra konkurrerende alternativ ned i den samme sumpen. Det kan innvendes at dette er en vel negativ taktikk. Selv om taktikken lykkes så gjør ikke det deontologi til en bedre normativ teori, og den bør fortsatt forkastet. Men mitt mål er ikke å konvertere noen til deontologien. Mitt mål er å undersøke om Greenes argument kan regnes som et framskritt innen normativ etikk, og skeptisisme er definitivt ikke dette.

4.2. Sidgwick

Dersom Greenes prosjekt skal utgjøre et framskritt innen normativ etikk må han forklare hvorfor konsekvensialismens grunnleggende intuisjoner stiller i en annen klasse enn deontologiens tilsvarende intuisjoner; hvorfor førstnevnte er til å stole på mens sistnevnte ikke er det. En populær forklaring på hvorfor nettopp dette er tilfelle, som Greene også viser til[109], er Sidgwicks distinksjon mellom filosofiske- og dogmatiske intuisjoner. Av disse er det de filosofiske intuisjonene som er Sidgwicks foretrukne intuisjoner, og han mener konsekvensialismens grunnleggende intuisjoner er av denne typen.

Nøyaktig hvor mange filosofiske intuisjoner Sidgwick opererer med er omstridt.[110] Likevel er det noen som er ganske klare. En tilsier for eksempel at det er galt for A å behandle B på en måte som det ville vært galt for B å behandle A på, kun på grunnlag av at de er to forskjellige individ, uten at det foreligger rimelige grunner for slik forskjellsbehandling.[111] En annen tilsier at alle er forpliktet til å anse det gode for hver enkelt person som like verdifullt som det gode for en selv.[112] En tredje tilsier at en bør sikte på ens eget gode i sin helhet, slik at en for eksempel ikke bør foretrekke et mindre gode nå over et større gode i morgen.[113] Dette utvalget viser at selv om Sidgwick mener hans filosofiske intuisjoner nødvendigvis leder til en

[109] Greene 2017: 10
[110] Skelton 2008: 187-188
[111] Sidgwick 1907: 380
[112] Sidgwick 1907: 382
[113] Sidgwick: 1907 381

konsekvensialistisk normativ teori så er ikke alle intuisjonene i seg selv eksklusive for konsekvensialismen. Den tredje intuisjonen er for eksempel felles for enhver maksimerende teori, enten den er egoistisk eller altruistisk, slik som henholdsvis etisk egoisme og konsekvensialisme. Tilsvarende krever den første av disse intuisjonene upartiskhet, og ekskluderer derfor egoisme, men er samtidig felles for både konsekvensialisme og deontologi. Den andre av disse intuisjonene beskriver Sidgwick som en rasjonell basis for det utilitaristiske systemet.[114]

Dogmatiske intuisjoner er derimot det Sidgwick ser på som folkemoral[115], implisitt i den moralske tenkningen til vanlige mennesker.[116] Som konkrete eksempler på dogmatiske intuisjoner diskuterer Sidgwick blant annet tanken om at vi har en plikt til å overholde løfter[117] og kompensere andre for urett vi påfører dem.[118] Ikke bare er dette arketypiske eksempler på deontologiens idé om spesielle forpliktelser vi har- og får som en følge av vår personlige omgang med andre mennesker, men de inkluderes også av Ross på hans liste over *prima facie* plikter.[119] En av de tydeligste forskjellene mellom filosofiske- og dogmatiske intuisjoner er at mens filosofiske intuisjoner er abstrakte moralske prinsipp og ideer så sier dogmatiske intuisjoner noe om konkrete moralske handlinger og generelle moralske regler i seg selv: «*it is implied that we have the power of seeing clearly that certain kinds of actions are right and reasonable in themselves, apart from their consequences»*[120]. Derfor er det klart at også intuisjoner av typen «det er galt å dytte den fremmede i *footbridge*», som Greene tar for seg i sine eksperiment, også tilhører denne gruppen intuisjoner, selv om Sidgwick, naturlig nok, ikke nevner dette eksplisitt.[121]

[114] SIdgwick 1907: 387
[115] *Common sense-morality*
[116] Sidgwick 1907: 101
[117] Sidgwick 1907: 354
[118] Sidgwick 1907: 293
[119] Ross 2002: 21
[120] Sidgwick 1907: 200
[121] Sidgwick nevner også en tredje type intuisjon, «*perceptual intuitionism», som går ut på at vi direkte kan sanse den moralske valøren til en konkret handling, uten referanse til øvrige prinsipp eller regler. (Sidgwick 1907: 100) Siden Sidgwick ikke sier like mye om denne typen, og det er uansett ikke er bedre stilt med den enn med dogmatiske intuisjoner, som han sier mye mer om, går jeg ikke nærmere inn på det her.*

Sidgwicks problem med dogmatiske intuisjoner er at de, som han selv sier det, mangler karakteristikkene til vitenskapelige aksiom.[122] Når Sidgwick foretrekker de filosofiske intuisjonene må det derfor antas at det er nettopp fordi de har karakteristikkene til vitenskapelige aksiom. Sidgwick beskriver de filosofiske intuisjonene som selvinnlysende prinsipp vedrørende hva som bør være, og at de er tilgjengelige gjennom direkte refleksjon.[123] Med dette mener han dog ikke at en gitt moralsk proposisjon ikke kan være usann eller er ufeilbarlig kun fordi den har status som en intuisjon: «*By calling any affirmation as to the rightness or wrongness of actions 'intuitive' (…) I only mean that its truth is apparently known immediately, and not as the result of reasoning*»[124]. Nettopp derfor, for at de i størst mulig grad skal være til å stole på, stiller Sidgwick krav om at intuisjonene må ha karakteristikkene til vitenskapelige aksiom.[125] I den forbindelse operer Sidgwick med fire kriterier som han hevder hans filosofiske intuisjoner tilfredsstiller, mens dogmatiske intuisjoner derimot ikke tilfredsstiller. For at en intuisjon skal være selvinnlysende sann må den (1) være klar og presis, (2) et resultat av omhyggelig refleksjon, (3) konsistent med andre proposisjoner ansett som selvinnlysende, i tillegg til at (4) uenighet vedrørende dens sannhetsgehalt må være fraværende eller mulig å avvise på et rasjonelt grunnlag.[126] Disse kriteriene er ganske selvforklarende, men siden jeg vurderer de, og især deres relasjon til henholdsvis konsekvensialistiske- og deontologiske intuisjoner, annerledes enn Sidgwick er det nødvendig å gå nærmere inn på hver enkelt.

(3) og (4) hviler på det samme logiske prinsippet. Om A og B er uenige om intuisjon *y* er usann eller sann må en av de ta feil siden *y* ikke samtidig kan være både sann og usann. Om A ikke har noen grunn til å anta at B står tilbake for A på noen måte, så har A like god grunn til å anta at B har rett som at A har rett, og følgelig like god grunn til å tro at *y* ikke er sann som at *y* er sann. Likeledes dersom A har intuisjon y og intuisjon x, og y og x, eller implikasjonene av y og x, kontradikterer hverandre, så er det et bevis på at enten *x* eller *y* må være feil. Det tredje kriteriet vektlegger altså koherens, og det er muligens det som har fått John Rawls til å argumentere for at Sidgwick faktisk argumenterer for en koherentistisk teori som i sin tur leder til Ralwls foretrukne metode; refleksiv likevekt. Som tilsvar til dette er det dog

[122] Sidgwick 1907: 360
[123] *Self-evident*
[124] Sidgwick 1907: 211
[125] Skelton 2010: 494
[126] Sidgwick 1907: 338-341

tilstrekkelig å påpeke at dette kriteriet er utelukkende negativt. Fraværet av koherens mellom moralske intuisjoner gir oss grunn til å betvile dem. Men det faktum at de koherere gir oss i seg selv ingen grunn til å anta at de er sanne.[127] At også deontologiske intuisjoner er i stand til å oppnå intern koherens bør dog være hevet over tvil. Man kan riktig nok si at i den virkelige verden er det mange deontologiske intuisjoner som ikke koherere, og at dette til og med gjelder mange av de deontologiske intuisjonene det vil være vanlig at en og samme person på samme tid har. Likevel finnes det ingen prinsipiell grunn for at man ikke skal kunne ha et koherent sett med deontologiske intuisjoner, og også i den virkelige verden bør dette kunne oppnås ved å eliminere de intuisjonene som ikke koherer med de øvrige intuisjonene i det aktuelle settet. Derfor vil jeg heretter la dette kriteriet ligge.

(4) er derimot der hvor dogmatiske intuisjoner får problemer, ifølge Sidgwick. Riktig nok virker mange av de dogmatiske intuisjonene å være preget av universell aksept, men denne universelle enigheten er ikke reell, og kan ifølge Sidgwick kun eksistere så fremst intuisjonene ikke samtidig tilfredsstiller (1):

> «*So long as they are left in the state of somewhat vague generalities, as we meet them in ordinary discourse, we are disposed to yield them unquestioning assent (…) But as soon as we attempt to give them the definiteness which cience requires, we find that we cannot do this without abandoning the universality of acceptance.*»[128]

Dogmatiske intuisjoner har med andre ord et dilemma. De kan tilfredsstille (4) kun ved ikke å tilfredsstille (1) – og omvendt. I forrige kapittel argumenterte jeg for at deontologiske teorier bør være rigide – eller nettopp *klar og presis* – for å unngå unødig system 1-inflytelse. Sidgwick argumenterer altså for noe lignende, men av en helt annen grunn. Her tror jeg min begrunnelse av (1) er bedre. Grunnen til dette er at Sigwicks egne filosofiske intuisjoner virker å få større problemer med (1) og (4), slik han selv begrunner dem, enn det han selv var av oppfatning av.

Sett i lys av den konteksten denne oppgaven utgjør så befinner (4) seg i en besynderlig posisjon. Utgangspunktet for denne oppgaven er den tilsynelatende mangelen på konsensus

[127] Skelton 2010: 504
[128] Sidgwick 1907: 342

som konstituerer Mackies *argument from relativity*. Sånn sett er det vanskelig å se hvordan Sidgwick kan påberope sine filosofiske intuisjoner en slik konsensus, all den tid de er byggeklosser for en klassisk form for utilitarisme – som det jo unektelig hersker stor uenighet rundt. Samtidig poengterer Sidgwick at uenighet kun er graverende dersom den ikke kan bortforklares på et rasjonelt grunnlag, og det er jo nettopp et slikt rasjonelt grunnlag Greenes originale argument utgjør om det er vellykket. Men som jeg har forsøkt å illustrere forutsetter Greenes originale argument at vi på en adekvat måte kan forklare hvorfor konsekvensialistiske intuisjoner ikke rammes av den samme kritikken som deontologiske intuisjoner. Men det er selvfølgelig en uholdbar løsning dersom man må anta at konsekvensialismens intuisjoner tilfredsstiller (4) for å gyldiggjøre Greenes poeng slik at det kan forklare hvorfor den tilsynelatende uenigheten vedrørende de filosofiske intuisjonenes sannhetsgehalt kan bortforklares på et rasjonelt grunnlag, og følgelig tilfredsstille (4).

Anthony Skelton argumenterer også for at enkelte av Sidgwicks filosofiske intuisjoner får problemer med (1) og (4). Når Sidgwick for eksempel skriver at like tilfeller skal behandles likt så fremst det ikke foreligger rimelige grunner til å forskjellsbehandle situasjonene unnlater han for eksempel å forklare hva som vil utgjøre en rimelig grunn i en slik situasjon. Hva som er rimelig kan her tolkes på flere måter og det er langt fra sikkert at rasjonelle aktører vil enes om hvordan det bør forstås. Samtidig bør det påpekes at dette nok er blant de minst kontroversielle av de filosofiske intuisjonene. Denne intuisjonen er tross alt felles for både deontologi og konsekvensialisme, som igjen representerer størstedelen av det moralfilosofiske landskapet. Dette er i det minste en indikasjon på at uenigheten, der den eksisterer, er av en mer teknisk enn fundamental karakter, siden de aller fleste som bedriver normativ etikk virker å ville si seg enig i noe veldig lignende av denne intuisjonen. Det er derfor ikke umulig å se for seg at man kan gi Sidgwicks intuisjon de presiseringene Skelton etterlyser og fortsatt ivareta en viss konsensus.

Andre av Sidgwicks filosofiske intuisjoner er dog av en mer kontroversiell natur. Et eksempel på dette er den følgende intuisjonen: «*Happiness (when explained to mean a sum of pleasures) (…) is the sole ultimate end*»[129]. I motsetning til den foregående intuisjonen er ikke denne en fellesnevner for både deontologi og konsekvensialisme, men er tvert imot en av de

[129] Sidgwick 1907: 402

intuisjonene det er mest konflikt rundt. Det vil si at man ikke, som i forrige avsnitt, kan appellere til en slags grunnleggende enighet om at noe slikt som denne intuisjonen nødvendigvis må være sann. Skelton påpeker riktig nok at også denne intuisjonen har den samme type tekniske utfordringer. For eksempel er det når Sidgwick skriver at lykke (forstått som å bety en sum av nytelse) er det eneste endelige målet uklart om han mener at nytelse er det eneste som har egenverdi eller at lykke er det eneste som har egenverdi, før det deretter argumenteres for at lykke utgjøres av nytelse. Man kan ifølge Skelton se for seg at mange vil være enige i den siste tolkningen, men uenige i den første tolkningen på grunn av måten Sidgwick definerer nytelse på: «*a feeling which, when experienced by intelligent beings, is at least implicitly apprehended as desireable*»[130]. Her kan det til og med tenkes at enkelte rasjonelle aktører vil avvise definisjonen, mens andre vil slutte seg til den, kun på grunn av måten de tolker *intelligent vesen* på.[131] Problemet med denne intuisjonen er likevel at den tekniske delen ikke er det største problemet. Selv om alle rasjonelle aktører skulle blitt enig om tolkningen, samt implikasjonene, av den aktuelle intuisjonen ville enighet vedrørende dens sannhetsgehalt fortsatt vært fraværende nettopp fordi langt fra alle er enige i at lykke er det eneste endelige målet, uavhengig av hvordan man definerer lykke-begrepet og dets relasjon til nytelse.

Siden også konsekvensialismens intuisjoner får problemer med (1) og (4) er det vanskelig å se hvordan disse skal fungere som en effektiv demarkasjonslinje mellom legitime og illegitime moralske intuisjoner. Det kan argumenteres for at deontologien i enda større grad enn konsekvensialismen får problemer med (1), men siden hovedproblemet med (1) ifølge Sidgwick er at det leder til (4), og konsekvensialismen har like store problemer med (4) som deontologien, er ikke dette problematisk. Jeg har også konkludert med at (3) bør være uproblematisk for begge parter, og det som gjenstår er da (2), hvor jeg tror det finnes et i denne sammenhengen uforløst potensial.

I forrige kapittel påpekte jeg at faktorer som kan oppsummeres som refleksjon, ifølge empirisk forskning, bidrar til å redusere i hvilken grad moralske aktører blir påvirket av såkalte *framing effects*. Dette virker også å være Sidgicks poeng; at vi ved å forstå den

[130] Sidgwick 1907: 127
[131] Skelton 2008: 206

aktuelle proposisjonen fullt ut lettere vil være i stand til å oppfatte feil med den gitte proposisjonen. Men å forstå den aktuelle proposisjonen kan også bety noe annet enn bare å tenke seg skikkelig godt om. Det kan også innebære å forstå årsakene til at vi har den aktuelle intuisjonen. Om det for eksempel ligger en bestemt bias bak en konkret moralsk intuisjon er det klart at man har mindre grunn til å tro at den intuisjonen er sann.[132] Det er her Greenes forskning møter Sidgwicks distinksjon. Når Greene argumenterer for at vi bør mistro de deontologiske intuisjonene som ligger bak de negative dommene i *footbridge*, så gjør han det nettopp med referanse til deres kausale opprinnelse. De produseres av en automatisk prosess som utløses av faktorer som nødvendigvis ikke er moralsk relevante. Derfor kan også Greenes bekymring sorteres under (2). Dette medfører i sin tur at selv om Sidgwick var av den oppfatning av at det var (1) og (4) som virkelig skapte problemer for dogmatiske intuisjoner, så kan det vise seg at (2) faktisk er vel så utfordrende.

Et annet prominent argument som også kan sorteres under (2) tilhører Peter Singer. Han viser også til Sidgwicks filosofiske intuisjoner som moralske intuisjoner vi kan stole på, og han argumenterer for at grunne til det er at de, i motsetning til de intuisjonene Greene kritiserer, ikke er et resultat av menneskets evolusjonshistoriske fortid.[133] Greene referer selvfølgelig også til menneskets evolusjonshistorie, men når han gjør det gjør han det for å forklare hvorfor system 1 virker som det gjør og hvorfor det produserer de moralske intuisjonene det produserer, og ikke som en selvstendig grunn til å mistro de samme intuisjonene. I senere publikasjoner har Greene også spilt ned rollen menneskets evolusjon har spilt for utviklingen av våre moralske intuisjoner.[134] Singers argument er også kontroversielt, og det kan for eksempel settes spørsmålstegn ved hvorvidt ikke også konsekvensialismens grunnleggende intuisjoner er et produkt av menneskets evolusjonshistorie, og også om dette nødvendigvis må være så graverende som Singer argumenterer for. Det er selvfølgelig mye mer som kan sies om denne type argument, men av plasshensyn går jeg ikke nærmere inn på dette i denne teksten.

[132] Skelton 2010: 495
[133] Singer 2005: 349-350
[134] Greene 2017

4.3. Modellbaserte- og modellfrie læring- og beslutningsprosesser

Greene refererer ikke bare til Sidgwick for å forsvare de konsekvensialistiske intuisjonenes overlegenhet, men har også utviklet et selvstendig argument. Denne grunngivingen bygger på den siste nyvinningen i Greenes prosjekt, som jeg presenterte mot slutten av kapittel to, nemlig distinksjonen mellom modellbaserte- og modellfrie læring- og beslutningsprosesser. Den enkleste måten å forstå dette argumentet på er ved å ta utgangspunkt i en konkret kritikk rettet mot Greene, utviklet av Peter Railton.

For det er ikke alle som er enige i at det faktum at en moralsk intuisjon har sitt opphav i en konkret nevrologisk prosess i seg selv er tilstrekkelig til å betvile den aktuelle intuisjonen. Enkelte har til og med påstått at dette bør ses på som en styrke, ikke en svakhet. Det er denne strategien Railton velger i artikkelen *The Affective Dog and Its Rational Tale: Intuition and Attunement*, hvor han konfronterer de samme empiriske resultatene som har vært tema for denne oppgaven. Her snur Railton opp-ned på Haidts eksempel med søskenparet Mark og Julie. Mens Haidt fokuserer på at testsubjektene fordømmer søsknene selv om de ikke er i stand til å artikulere gode grunner for dette, så fokuserer Railton på hvorfor testsubjektenes intuisjon faktisk er riktig. For selv om Haidt i eksempelet presiserer at Mark og Julies eksperiment ender godt så kunne Mark og Julie på ingen måte vite dette helt sikkert på forhånd. De spilte faktisk russisk rulett med forholdet sitt, og selv om det endte vel er det en god grunn til å være skeptisk til handlingen.[135] Railtons poeng er at denne formen for intuisjoner reflekterer dyrkjøpt evolusjonær erfaring og derfor er smartere enn det vi gjerne tror.

> «*Emotions are modes of functioning, shaped by natural selection, that coordinate physiological, cognitive, motivational, behavioral, and subjective, responses in patterns that increase the ability to meet the adaptive challanges of situations that have recurred over evolutionary time*»[136]

Railtons poeng er i utgangspunktet et friskt pust slik den fram til nå har framstått. Faktoren som i utgangspunktet blir sett på som en korrumperende faktor er nå det som informerer og gir legitimitet. Men dette er bare midlertidig. For Greenes tilsvar til Railton, om hvorfor dette

[135] Railton 2014: 832
[136] Nesse og Ellsworth 2009: 129

ikke er det beste forsvaret av affektive system 1-intuisjoner, er muligens også den beste grunngivningen av hvorfor konsekvensialismens intuisjoner bør foretrekkes over deontologiens intuisjoner. Slik Greene ser det er hans uenighet med Railton kun en tilsynelatende uenighet. For den historien Railton forteller er ifølge Greene en historie i to deler. Som jeg presiserte allerede i kapittel to vil Greene være den første til å innrømme at det finnes mye visdom i de samme moralske intuisjonene han kritiserer – nettopp fordi de er tuftet på erfaringen av hva som fungerer og hva som ikke fungerer.[137] Derfor er de samme intuisjonene velegnet som moralsk kompass hva angår de dagligdagse, velkjente moralske problemene vi støter på. Dine praktiske intuisjoner om at det er galt å lyve til vennene dine, ha sex med søsknene dine eller å dytte fremmede fra gangbroer ned på jernbaneskinner vil med andre ord være en god veiviser i de fleste tilfeller, nettopp fordi de reflekterer erfaring som tilsier at slik adferd sjelden fører til noe godt, og følgelig bør unngås. Problemet oppstår når omgivelsene endrer seg slik at vi enten møter på situasjoner hvor vi mangler tilstrekkelig erfaring, eller enda verre – det som ifølge erfaringen er veldig dumt faktisk er veldig bra. Det er nettopp dette som er tilfelle i *footbridge*: En handling som ifølge all erfaring vil være negativ blitt satt inn i en særs uvanlig kontekst, slik at det som vanligvis er negativt blir positivt; og derfor går system 1 amok, uten at vi helt skjønner hvorfor.

Mens dogmatiske intuisjoner korresponderer med det jeg tidligere har omtalt som modellfri læring- og beslutningstaking korresponderer filosofiske intuisjoner til det jeg har kalt modellbasert læring- og beslutningstaking. Det innebærer at der hvor dogmatiske intuisjoner knytter verdi direkte opp mot konkrete handlingstyper, og derfor er dogmatiske i den forstand at de ikke er i stand til å håndtere situasjoner hvor omstendighetene og forholdene endrer seg, så rammer ikke dette filosofiske intuisjoner. Når konsekvensialismen i motsetning baserer seg på filosofiske intuisjoner, slik som at for eksempel lykke er godt, og at det *ceteris paribus* er bedre at flere enn færre personer opplever lykke, så tar ikke denne type intuisjon stilling til verdien av konkrete handlingstyper i seg selv, men kun hvorvidt de bidrar til å realisere det endelige målet målet. På grunn av dette er ikke denne type intuisjoner på samme måte sårbare for et beslutningsmiljø som ikke er konstant. Om det som vanligvis vil produsere mest lykke i et konkret tilfelle ikke vil gjøre det i et annet helt konkret tilfelle, så er det for en slik filosofisk intuisjon uproblematisk å tilpasse seg en slik anomali, siden den ikke tillegger

[137] Greene 2017: 3

handlinger i seg selv en gitt verdi, men kun vurderer i hvilken grad det endelig målet – å maksimere det som defineres som godt – realiseres. Derfor vil ikke denne intuisjonen insistere på at handlingen som vanligvis vil produsere mest lykke gjennomføres, siden den kun tar hensyn til det endelige resultatet, og derfor innser at det er irrelevant hva som vanligvis vil produsere mest lykke siden det likevel ikke er tilfelle i den helt konkrete situasjonen som vurderes der og da.[138]

Dette er en konkret grunn til å foretrekke en etisk teori som kun trenger å basere seg på filosofiske intuisjoner. Greene mener igjen dette taler for konsekvensialisme som det overlegne alternative blant normative teorier: «*Consequentialism is what you get when you apply model-based thinking to the general problem of morality at the level of first principles.*»[139] Det er ikke vanskelig å se hvorfor han mener dette, og det er heller ikke vanskelig å se hvordan dette skaper vansker for mitt foreslåtte alternativ med rigide, deontologiske teorier. Jeg lanserte for eksempel en hierarkisk ordnet versjon av Ross åtte *prima facie* plikter som en potensiell rigid og deontologisk teori, men det er åpenbart at disse pliktene fortsatt vil tillegge handlingstyper verdi i seg selv, og i så måte være dogmatiske intuisjoner som rammes av Greenes poeng.

Likevel er det ikke åpenbart at alle intuisjonene i deontologens repertoar må være av en slik karakter. En av Sidgwicks filosofiske intuisjoner er til og med felles for deontologien og konsekvensialismen, og er et godt bevis på dette. Samtidig er denne felles intuisjonen alene ikke nok til å utgjøre en komplett normativ teori, og en deontologisk teori vil derfor være avhengig av flere intuisjoner enn kun denne. Men det samme vil også være sant for konsekvensialismen. Også den trenger flere intuisjoner, og som jeg allerede har poengtert er det disse intuisjonene, de som ikke er felles men snarere er eksklusive for konsekvensialismen, som er de mest kontroversielle. Det er heller ikke sikkert distinksjonen mellom dogmatiske- og filosofiske intuisjoner er så klar som først antatt.

[138] Greene 2017: 10-11
[139] Greene 2017: 11

Om man for eksempel ser for seg en klassisk deontologisk intuisjon, slik som «det er galt å lyve» så er dette både en dogmatisk intuisjon og en modellfri intuisjon, fordi den knytter en konkret verdi til handlingen i seg selv. Men man kan også se for seg andre former for deontologiske intuisjoner. Dette er intuisjoner som ikke er dogmatiske intuisjoner og som ikke tillegger konkrete handlinger moralsk verdi i seg selv. Kanskje er de heller ikke like rene filosofiske intuisjoner som en intuisjon om moralsk upartiskhet, side de er intuisjoner som er eksklusive for deontologisk teori, og derfor også er mer kontroversielle. Spørsmålet er om de på en relevant måte skiller seg fra konsekvensialismens tilsvarende intuisjoner. For eksempel kan man ha intuisjoner som tilsier at «menneskelig verdighet har moralsk verdi» og «det er ikke slik at menneskelig verdighet nødvendigvis bør maksimeres».[140] Samtidig kan man se for seg at man har klassiske konsekvensialistiske intuisjoner slik som «nytelse har moralsk verdi» og «den totale mengden nytelse bør maksimeres». Disse intuisjonene sier ingenting om hvilke konkrete handlinger som er riktige og gale i konkrete situasjoner, men er heller abstrakte moralske prinsipper og ideer, og må derfor i utgangspunktet klassifiseres som filosofiske intuisjoner. Man må likevel ikke glemme at det er en tydelig relasjon mellom dogmatiske- og filosofiske intuisjoner. Selv om en teoretisk intuisjon i seg selv ikke knytter moralsk verdi direkte til en handling, slik som «det er galt å lyve», så konstituerer de kriterier for hvilke handlinger som vil ha hvilken moralsk verdi i en konkret situasjon. Man kan altså se for seg at begge disse settene med intuisjoner kan parafraseres til å si noe som slikt som dette: «Handling x har moralsk verdi y fordi den respekterer menneskelig verdighet i den konkrete situasjonen», eller «handling x har moralsk verdi y fordi den maksimerer den totale mengden nytelse i den konkrete situasjonen».

Til tross for at begge typer filosofiske intuisjoner virker å ha identisk funksjon i det foregående eksempel mener Greene at kun konsekvensialistiske, filosofiske intuisjoner kan være et resultat av modellbasert læring. Bakgrunnen virker å være at det kun er de som er i stand til å navigere et beslutningsmiljø som ikke er konstant, men i kontinuerlig endring. Men også her kan det innvendes at det virker som om enkelte deontologiske intuisjoner kan utføre en lignende funksjon. Dogmatiske intuisjoner som «det er galt å lyve» er dårlige fordi de kommer til kort når det som vanligvis er galt, altså å lyve, på grunn av endrede

[140] Selv om det er mulig at for eksempel Sidgwick ville klassifisert også disse intuisjonene som dogmatiske er det viktig å huske på at Sidgwicks distinksjon får sin kraft fra kriteriene som begrunner distinksjonen. Det essensielle her er med andre ord at intuisjonene gjør en like god jobb med disse kriteriene som sin konsekvensialistiske motpart – ikke hva vi kaller dem.

omstendigheter ikke lenger er det. Filosofiske intuisjoner, slik som «maksimer nytelse» er ikke dårlig på denne måten fordi den ikke tar hensyn til hvordan målet – altså å maksimere nytelse – realiseres. Om det som vanligvis er dårlig er bra fordi det nå under ekstraordinære omstendigheter faktisk maksimerer det som defineres som godt så er denne intuisjonen i stand til å få med seg det. Så er spørsmålet hvordan det stiller seg for intuisjoner som «menneskelig verdighet har moralsk verdi» og «det er ikke slik at menneskelig verdighet nødvendigvis bør maksimeres».

Disse intuisjonen sier i seg selv ikke noe om hvilke konkrete handlingstyper som bør og ikke bør aksepteres, men kun at handlinger som bidrar til at målet – at menneskelig verdighet produseres eller respekteres – er gode. I så måte minner de om den konsekvensialistiske intuisjonen. Den største forskjellen er dog at den allmenne forståelsen av hva hvilke handlinger som faktisk produserer, eller respekterer, menneskelig verdighet – og omvendt – er ganske satt. De fleste vil være enige i at det å spytte på noen, bedra noen eller på andre måter ydmyke en person ikke respekterer vedkommendes verdighet. Fra en slik observasjon er det lett å la seg friste til å utlede at denne intuisjonen knytter moralsk verdi direkte til konkrete handlinger, og derfor må være basert på en modellfri læringsstrategi uegnet til å håndtere er beslutningsmiljø i endring. Lignende vil de fleste være enige i at det å drepe, skade eller torturere en annen person ikke produserer de beste konsekvensene, og derfor kan man kanskje lure på om ikke konsekvensialismens filosofiske intuisjoner også knytter moralsk verdi direkte opp mot konkrete handlinger. Dette er selvfølgelig ikke tilfelle, som de mange tankeeksperimentene som har preget moralfilosofien de siste århundrene er et godt bevis på. Om den eneste måten man kan forhindre en ondskapsfull terrorist i å spre et virus som på den aller mest smertefulle måten vil drepe alt liv på jorda er ved å torturere vedkommende og få han til å røpe motgiften vil de fleste være enige i at handlingen som tidligere ble sett på som utillatelig nå er tillatelig – nettopp fordi den nå produserer de beste konsekvensene. Spørsmålet er om en lignende manøver er tilgjengelig for de deontologiske intuisjonene jeg har diskturt her.

Jeg har allerede nevnt spytting som et eksempel på en handling som vanligvis vil bli ansett som ikke å respektere verdigheten til den du spytter på. En beslektet og enda mer ekstrem handling er urinering. I naturen brukes urinering ofte for å kommunisere dominans og å

markere territorium. Derfor er det ingen overraskelse at det å urinere på et annet menneske oppleves både undertrykkende og uforenelig med respekt for menneskets verdighet. Likevel finnes det ekstraordinære omstendigheter hvor denne relasjonen brytes. Om du på sydenferie går en tur langs stranden og kommer over en fremmed person som nettopp har hatt et ublidt møte med en brennmanet, og vedkommende ber deg urinere på armen som har vært i kontakt med brennmaneten for å begrense skadene, så bør du selvfølgelig, om du har en deontologisk intuisjon av den typen jeg har foreslått, etterkomme dette ønsket.

En potensiell innvending til et slikt eksempel er at det fortsatt kan påstås at det å urinere på en annen person, selv under disse omstendighetene, fortsatt innebærer at vedkommendes verdighet krenkes. Det sitter tross alt ganske langt inne å urinere på andre, selv når de selv ber om det. Konsekvensialisten kan da innvende at grunnen til at det er ok å se bort ifra dette i det foregående eksempelet er nettopp hensynet til konsekvensene, som tross alt maksimeres ved at man urinerer på et annet menneske selv om det ikke respekterer vedkommendes verdighet. Men denne kritikken er ikke verre enn at man kan revidere eksempelet på en måte som utelukker en slik innvending. Man kan for eksempel revidere eksempelet slik at man nå ser for seg at man atter vandrer langs en sydenstrand og kommer over en fremmed person. Nok en gang kommer du over en person brent av en brennmanet, som ber om din hjelp. Men denne gangen er stranda full av mennesker. Siden du befinner deg i et relativt konservativt (men ikke så konservativt at du risikerer straffeforfølgelse) land, slik at du rimeligvis kan anta at folkemengden på stranda vil la seg indignere i en slik grad at deres sammenlagte ubehag vil veie tyngre enn den smerten du kan lindre hos den ulykksalige personen foran deg. Nok en gang bør det være åpenbart at du, dersom du er en deontolog som tror at disse intuisjonene er korrekte, bør etterkomme ønsket til den fremmede, og denne gangen kan heller ikke en slik vurdering tilskrives konsekvensialistiske hensyn.

På et overordnet nivå kan foregående avsnitt virke paradoksalt, da det vitterlig later til å være basert på en dogmatisk intuisjon av den typen som Greene vil til livs i utgangspunktet. Her er det viktig å poengtere at dette eksempelet på ingen måte er ment å bevise at den aktuelle moralske dommen, at du bør urinere på den fremmede, er den riktige vurderingen. Slik eksempelet er formulert vil en konsekvensialist nødvendigvis være uenig i at du bør urinere på den fremmede, og i denne sammenhengen anser jeg det som et åpent spørsmål hva som vil

være korrekt i dette tilfelle. Derimot er mitt poeng ganske enkelt at det er slik en deontolog med utgangspunkt i de deontologiske intuisjonene jeg har lansert vil argumentere, noe som i sin tur er ment å illustrere hvorfor deontologiske intuisjoner ikke nødvendigvis må tilskrive konkrete handlinger moralsk verdi i seg selv.

Mitt poeng så langt kan oppsummeres på den følgende måten. Både de deontologiske intuisjonene jeg har operert med her så vel som de mer klassiske konsekvensialistiske intuisjonene later til å kunne parafraseres som kriterium for hvilke handlinger som vil være moralsk tillatelig og ikke. Den konsekvensialistiske intuisjonen tilsvarer at en tillatelig handling er den handlingen som bidrar til den største mulig lykke eller nytelse, mens de deontologiske intuisjonene tilsvarer at en tillatelig handling er en handling som produserer- eller ikke krenker menneskelig verdighet. Her kan det også tenkes at den konsekvensialistiske intuisjonen kan suppleres av den deontologiske intuisjonen i en potensiell deontologisk teori, slik at sistnevnte utgjør en begrensning for førstnevnte. Da vil man stå igjen med en komplett normativ teori som kan si noe om hva en bør gjøre i enhver situasjon.

Enkelte vil kanskje insistere på at den deontologiske intuisjonen fortsatt er basert på en modellfri læring- og beslutningsstrategi. Min feil, vil de si, er at jeg undervurderer den modellfrie strategien, og det faktum at en intuisjon kan vise til en viss fleksibilitet er ikke ensbetydende med at den er et resultat av en modellbasert læring- og beslutningsstrategi. For en modellfri strategi må ikke vurdere handlinger i seg selv, slik at en konkret handling nødvendigvis har en konstant verdi, men kobler snarere en gitt verdi til en gitt handling i en gitt kontekst. Det å slå en hammer mot kroppen til en annen person kan for eksempel gis en negativ verdi, mens det å slå en hammer mot en spiker gis en positiv verdi. Handlingen, det å svinge en hammer gjennom luften, er identisk i begge tilfellene, men konteksten veldig forskjellig, og følgelig evalueres de to situasjonene vidt forskjellig.[141] Så kan man spørre seg om det ikke er nøyaktig det samme som er tilfelle i mitt eksempel hvor man ender opp med å urinere på en fremmed person. Selv om det definitivt er mulig å nå den konkrete moralske konklusjonen i mitt eksempel ved hjelp av en modellfri strategi er det likevel ikke det som er tilfelle i mitt eksempel. I mitt eksempel gjør den aktuelle intuisjonen ingen referanse til verken konkrete handlingstyper i seg selv eller handlingstyper i konkrete kontekster. Det

[141] Cushman 2013: 283

eneste den gjør er å vise til en konkret moralsk verdi, verdighet, som bortsett fra det faktum at det ikke foreligger en intuisjon som tilsier at den bør maksimeres, har nøyaktig den samme funksjonen som nytelse eller lykke har innen konsekvensialistiske teorier.

Derfor er det besynderlig at Greene, til tross for at begge intuisjonene på denne måten har en identisk funksjon, mener at den deontologiske intuisjonen i større grad vektlegger handlingstyper i seg selv fordi den er villig til å si noe om hvilke handlinger som er tillatelig uavhengig av hvilke konsekvenser de har.[142] I utgangspunktet er det selvfølgelig riktig at deontologien i motsetning til konsekvensialismen er villig til å si noe om en handlings moralske valør uavhengig av dens konsekvenser, men det avhenger også av hvordan man forstår «konsekvensene» i denne sammenhengen. Jeg kommer tilbake til den alternative forståelsen av «konsekvensene» jeg referere til her og forholder meg enn så lenge til den mest nærliggende forståelsen. For i utgangspunktet representerer fokuset på konsekvensene en grense for hvor langt deontologien kan etterape konsekvensialismens struktur for å sno seg unna Greenes innvendinger. Tidligere presiserte jeg at menneskelig verdighet, i motsetning til for eksempel nytelse, ikke må forstås som et fenomen som nødvendigvis bør maksimeres. Hadde det ikke vært for dette hadde den aktuelle intuisjonen fort kunne blitt en slags konsekvensialistisk intuisjon hvor menneskelig verdighet er svaret på spørsmål om hva som defineres som godt i neste omgang forutsettes som det som skal maksimeres ifølge den aktuelle konsekvensialistiske doktrinen. Derfor er det i dette tilfellet ikke et alternativ å etterligne den konsekvensialistiske formen for å unngå Greene argument.

Problemet med denne innvendingen er at den ikke virker å følge av distinksjonen mellom modellbasert- og modellfri læring, som Greene i utgangspunktet bruker til å skille konsekvensialismens intuisjoner fra deontologiens intuisjoner. Som jeg har vist kan også deontologiske intuisjoner være fleksible slik at de kan fungere selv når beslutningsmiljøet ikke er konstant, og det er uklart hvorfor det skulle være diskvalifiserende at de gjør dette uten nødvendigvis å ta hensyn til konsekvensene. En mulighet er at Greene baserer seg på en form for uformell statistisk representasjon av deontologiens intuisjoner. For deontologien har unektelig en tendens, en tendens som ikke er mulig innen konsekvensialismen, til å tilskrive konkrete handlinger moralsk verdi i seg selv. Men dette leder kun tilbake til

[142] Greene 2017: 10

stråmannproblematikken fra forrige kapittel. Det kan selvfølgelig konstrueres et selvstendig argument for å forklare hvorfor man ikke kan si noe om en konkret handling er tillatelig eller utillatelig uten å ta hensyn til konsekvensene. Men et slikt argument vil i praksis være et argument for konsekvensialismen, og et argumentet som tilsier at konsekvensialistiske intuisjoner er til å stole på fordi de støtter konsekvensialistisk teori, og tilsvarende at deontologiske intuisjoner ikke er til å stole på fordi de ikke støtter konsekvensialistisk teori, vil være sirkulært på en uholdbar måte.

Sidgwick sier noe av det samme som Greene – problemet med dogmatiske intuisjoner er at de sier noe om den moralske valøren til konkrete handlinger uavhengig av deres konsekvenser[143] – og siden Greene refererer eksplisitt til Sidgwicks filosofiske intuisjoner er det mulig at han også stiller seg bak Sidgwicks egen begrunnelse. I et slikt scenario trenger med andre ord ikke dette poenget å følge av Greenes distinksjon, men kan snarere være tenkt å følge av Sidgwicks fire kriterier for akseptable moralske intuisjoner. Dette gir mening siden det kun er Sidgwicks filosofiske intuisjoner som angivelig tilfredsstiller disse kriteriene, og Sidgwicks filosofiske intuisjoner igjen leder til en konsekvensialistisk teori og en konsekvensialistisk teori aldri vil kunne akseptere en moralsk intuisjon som tilskriver handlinger moralsk verdi uavhengig av konsekvensene. Men som jeg allerede har poengtert er det ikke så åpenbart som Sidgwick mente at hans filosofiske intuisjoner tilfredsstiller hans kriterier, noe som igjen betyr at denne grunnen til å akseptere at moralske intuisjoner som verdsette handlinger på andre måter enn ved å vise til deres konsekvenser er uholdbare faller bort.

Det er dog verdt å merke seg at man også kan tolke Greenes utsagn på en måte som gjør at det faktisk følger av den opprinnelige distinksjonen mellom modellbaserte- og modellfrie læring- og beslutningsstrategier. Men om man legger en slik tolkning til grunn er påstanden usann fordi deontologien faktisk tar hensyn til konsekvensene. En vanlig måte å forstå forskjellen mellom modellbaserte- og modellfrie læring- og beslutningsstrategier på er ved å forstå førstnevnte som prospektiv og sistnevnte som retrospektiv.[144] Dette er naturlig siden en modellfri beslutningsstrategi ikke prøver å oppnå det beste resultatet ved å kalkulere konsekvensene av en handling, men heller baserer seg på hva som pleier å fungere best i

[143] Sidgwick 1907: 200
[144] Crockett 2013: 363

lignende situasjoner. Fra dette er det mulig å utlede at kun konsekvensialismens intuisjoner kan være baserte på en modellbasert strategi fordi det kun er de de som analyserer moralske handlinger i form av deres konsekvenser. Riktig nok kan det innvendes at også deontologisk bør ta hensyn til konsekvensene på denne måten. Men samtidig må den, for at den ikke skal være en konsekvensialistisk teori, inneholde intuisjoner som begrenser den konsekvensialistiske intuisjonen, og da må den igjen ty til intuisjoner som her tolkes som dogmatiske.

Problemet med en slutning av denne typen er at den forveksler en nokså alminnelig forståelse av hva «en handlings konsekvenser» innebærer med en utpreget konsekvensialistisk forståelse av hva det samme utsagnet innebærer. En handlings konsekvenser kan forstås på en utpreget konsekvensialistisk måte, som å vise til en kost-/nytteanalyse. Men det kan også forstås på en måte som gjør det mulig for deontologien å ta hensyn til konsekvensene, også foruten om de intuisjonene den har til felles med konsekvensialismen. Når deontologen sier at en handling har dårlige konsekvenser kan det ganske enkelt menes at den har dårlig konsekvenser fra et deontologisk perspektiv. For å gjøre forskjellen fra en konsekvensialistisk forståelsene av «konsekvensene» tydelig kan man heller si at «følgene», heller enn «konsekvensene», er negative. Et eksempel på dette kan være at handlingen ikke respekterer, eller ikke produserer, menneskelig verdighet. Da kan man si at handlingen har negative følger – at en persons menneskelige verdighet ikke respekteres – uten å vise til konsekvensene i konsekvensialistisk forstand. Om man med andre ord ser for seg at verdighet er en deontologisk verdi på lik linje med måten konsekvensialismen verdsetter lykke eller nytelse bør det være ganske åpenbart at også deontologiens intuisjoner kan være prospektive og faktisk kan vurdere handlinger i henhold til deres konsekvenser; endog ikke i konsekvensialistisk forstand. Denne måten å forstå «en handlings konsekvenser på» virker altså å følge av distinksjonen mellom modellbaserte- og modellfrie læring- og beslutningsprosesser, men den gjør til gjengjeld ingenting for å underminere deontologiens grunnleggende intuisjoner.

Det eksisterer åpenbart klare utfordringer for hvordan Greene ønsker å benytte distinksjonen mellom modellbaserte- og modellfri læring- og beslutningsprosesser på. Men når det er sagt tror jeg også han er inne på noe; at det er egenskaper ved konsekvensialismens filosofiske intuisjoner som gjør de spesielt velegnet som moralfilosofiske førsteprinsipp, og at dette er

egenskaper som er eksklusive for konsekvensialismens grunnleggende intuisjoner. Videre tror jeg en syntese av Sidgwick og Greenes i seg selv ufullstendige argument for det samme poenget, sett i relasjon til min distinksjon mellom rigide- og fleksible normative teorier, er det beste argumentet for dette.

4.4. «Greenwick»

I *The Secret Joke of Kants Soul* skriver Greene at selv om emosjoner nødvendigvis spiller en viss rolle innen både deontologi og konsekvensialisme, så fungerer det affektive elementet under sistnevnte mer som en slags valuta, og under førstnevnte mer som en alarm.[145] Selv om Greene siden har blitt kritisert for nettopp denne formuleringen fordi den er fenomenologisk og uvitenskapelig i sin natur,[146] så er den beskrivende. En valuta er kvantifiserbar, den kan byttes i andre valuta og i teorien i alle andre ting av tilsvarende verdi. Det er mulig at det er nettopp denne egenskapen som gjør at konsekvensialismen er i stand til å være rigid samtidig som den kun trenger å basere seg på filosofiske og fleksible moralske intuisjoner. For om man sammenligner med deontologien er det vanskelig å se for seg en tilsvarende intuisjon som kan gjøre samme jobb. En deontologisk intuisjon om menneskelig verdighet kan riktignok være fleksibel og dermed være effektiv selv innenfor et beslutningsmiljø som ikke er konstant, men dersom den er det kan den ikke samtidig være rigid på en slik måte at anvendelsen av den er utvetydig.

Når jeg i forrige kapittel argumenterte for at også deontologiske teorier kunne oppnå rigiditet som begrenser uønsket system-1 innflytelse så virker det som om denne rigiditeten er noe deontologien kan tilegne seg gjennom en rekke dogmatiske intuisjoner. Men som denne kritiske analysen av hvilke typer moralske intuisjoner som er egnet til å spille en aktiv rolle innen moralfilosofien viser så er dogmatiske intuisjoner problematiske i en slik sammenheng. Som det dog har framkommet de siste sidene bør det kunne eksistere deontologiske og filosofiske intuisjoner som er fleksible på den måten Greene etterspør, men problemet med dette er igjen at de da er fleksible nettopp ved ikke å være rigide slik jeg etterlyste i forrige kapittel. Det som etterlyses er altså en moralsk teori som opererer med et førsteprinsipp som er fleksibelt i den forstand at det ikke knytter moralsk verdi direkte til handlinger i seg selv,

[145] Greene 2008: 41
[146] Berker 2009: 308

samtidig som det er rigid i den forstand at anvendelsen av førsteprinsippet er en rent mekanisk prosess som ikke tillater uønsket system 1-innflytelse under den moralske beslutningsprosessen. Der hvor disse to kravene er uproblematiske for konsekvensialismen virker derimot deontologien kun å være i stand til å oppfylle et av kravene om gangen.

Her kan kanskje deontologen innvende at rigiditeten i deontologien ikke trenger å komme direkte fra dogmatiske intuisjoner. Selv om en moralsk proposisjon av typen «handling p er moralsk utillatelig under omstendighet q» kan, og ofte vil, forstås som en dogmatisk intuisjon så må den nødvendigvis ikke det. Dersom en slik proposisjon anses som en moralsk intuisjon antas den å være selvinnlysende slik at man er rettferdiggjort i å tro at proposisjonen er sann uten videre bevisførsel, men den trenger absolutt ikke ansees som sann kun fordi den ansees som selvinnlysende. Et annet alternativ er at man ganske enkelt har andre, eksterne, grunner til å tro at den er sant. For eksempel fordi den kan utledes fra et førsteprinsipp som anses som selvinnlysende sant. I nettopp slik stil argumenterer Robert Audi for at de åtte *prima facie* pliktene til David Ross kan utledes fra noe slikt som det kategoriske imperativ, samtidig som de forskjellige pliktene kan ordnes og systematiseres slik at vi vet hva som kreves av oss også nå de forskjellige pliktene er i konflikt med hverandre.[147] En slik konstruksjon minner om den fra forrige kapittel, hvor Ross' *prima facie* plikter utgjorde utgangspunktet for en rigid deontologisk teori. Forskjellen er selvfølgelig at teorien nå ikke er avhengig av dogmatiske intuisjoner.

Problemet med denne strategien er at den ikke løser noe. Selv om de dogmatiske intuisjonene som i utgangspunktet besørger deontologien med dens rigiditet ikke trenger å være dogmatiske intuisjoner, men heller kan være mellomaksiom som utledes fra filosofiske intuisjoner, så må likevel rigiditeten stamme fra et sted. Siden mellomaksiomene stammer fra en eller flere filosofiske intuisjoner må også rigiditeten komme fra samme sted. Det vil si at alle som vurderer den aktuelle filosofiske intuisjonen må forstå hva den innebærer slik at den kan implementeres på en mekanisk måte som ikke tillater uønsket system 1-innflytelse på fortolkningsprosessen. I forrige kapittel argumenterte jeg for at dette er problematisk for mange av de deontologiske filosofiske intuisjonene, og derfor måtte en rigid deontologisk teori konstrueres via en rekke dogmatiske intuisjoner. Men selv om man antar at det er mulig

[147] Audi 2005: 90-91

å konstruere en filosofisk deontologisk intuisjon som innehar en slik egenskap virker det usannsynlig at den samtidig skal være i stand til å tilfredsstille Greenes krav om fleksibilitet i forbindelse med et beslutningsmiljø som ikke er konstant. Derfor leder en slik strategi kun tilbake til start.

Når Sidgwick skriver at dogmatiske intuisjoner kan tilfredsstille (1) kun ved ikke å tilfredsstille (4), og omvendt; tilfredsstille (4) ved ikke å tilfredsstille (1), så er han ikke så langt unna å treffe spikeren på hodet. (1) innebærer som kjent at en intuisjon må være klar og presis, og Sidgwicks bekymring er at den tilsynelatende enigheten rundt de dogmatiske intuisjonene vil forsvinne straks de blir klare og presise, slik at alle forstår hvordan de skal forstås, hva de innebærer og hvordan de må implementeres i den moralske beslutningsprosessen. Men jeg har også poengtert en annen grunn til å vektlegge (1). Som jeg har vært inne på er nemlig det at en intuisjon er klar og tydelig; herunder også implikasjonene og anvendelsen av intuisjonen, en forutsetning for at intuisjonen skal være rigid. Om man så forstår (1) på denne måten, for så å forstå det i relasjon til den utvidede forståelsen av (2) som Greene representerer, viser det seg at det giftigste dilemmaet for deontologen ikke består i (1) eller (4), men heller (1) eller (2). Om den skal være klar og tydelig må den gi opp evnen til å navigere et miljø i stadig endring fordi den må knytte moralsk verdi direkte til handlinger, og skal den være i stand til å navigere et skiftende miljø kan den ikke lenger knytte moralsk verdi direkte til handlinger men må heller basere seg på vage førsteprinsipp som er åpen for forskjellige fortolkninger.

Jeg tror dette dilemmaet er den best forklaringen på hvorfor det ikke er tilstrekkelig at deontologien kan vise til filosofiske intuisjoner basert på modellbasert læring så lenge de samme intuisjonene fortsatt er villige til å tilskrive handlinger moralsk verdi uavhengig av konsekvensene. Grunnen til at en normativ teori som baserer seg på intuisjoner som kun vektlegger konsekvensene av en moralsk handling er immun mot dette dilemmaet virker å være nettopp de valuta-lignende egenskapene som en konkret handlings konsekvenser representerer for konsekvensialismen. Siden alle handlinger har konsekvenser og konsekvenser, ved å omsettes til for eksempel lykke eller nytelse, fungerer som en felles valuta alle moralske problem kan omsettes til, selv på tvers av tid og rom, er ikke fleksibilitet med tanke på et beslutningsmiljø som ikke er konstant et problem for konsekvensialismens

førsteprinsipp. Likeledes er konsekvensene – som en hvilken som helst annen valuta – kvantifiserbar, noe som igjen gjør en mekanisk og rigid anvendelse av konsekvensialismens førsteprinsipp mulig.

I forrige kapittel lanserte jeg dog regel-konsekvensialisme som en mulig rigid, deontologisk teori, og det er verdt å nevne at regel-konsekvensialismen ikke virker å rammes av dette dilemmaet. Grunnen til dette er at også regel-konsekvensialismen ser på konsekvensene som det som betinger moralsk verdi og dermed høster godene av konsekvensenes valuta-lignende egenskaper. Regel-konsekvensialismen har altså til felles med handlings-konsekvensialismen hva som gir verdi, men skiller seg fra handlings-konsekvensialismen i formelen det som gir verdi inngår i. Det er også det som er grunnen til at regel-konsekvensialisme ikke klassifiseres som konsekvensialisme under den definisjonen jeg har operert med i denne oppgaven. Men selv om regel-konsekvensialisme i teorien er en form for ikke-konsekvensialistisk hybrid-teori er det en mager trøst for deontologiens proponenter. Særlig det faktum at det er dens distinkte likhet med handlingskonsekvensialisme som gjør at den unngår kritikken som er så problematisk for deontologien er lite oppmuntrende, og mange deontologer vil nok se på deontologien som en tapt sak dersom den eneste måten den kan bestå på er som en form for regel-konsekvensialisme.

4.5. Er deontologien usann og konsekvensialismen sann?

Her kan det være fristende å spørre nøyaktig hvor skadelig dette dilemmaet er for deontologen. For her virker det å være et alternativ for deontologen å bite i den berømte kula og si «hva så?» For det er i utgangspunktet ikke helt klart hvorfor det er et fellende problem at deontologien ikke er fleksibel, slik at den ikke like effektivt kan navigere et miljø i endring, for eksempel ved at den knytter moralsk verdi direkte til handlinger i seg selv, og følgelig er basert på modellfri læring- og beslutningsstrategi. Fo hva om det ganske enkelt er slik den moralske virkeligheten ser ut? Som Ross for eksempel skriver: «*it is more important that our theory fit the facts than it be simple*»[148]. Riktig nok mente Ross dette som tilsvar til G. E. Moore som appellerte til utilitarismens evne til å håndtere moralske plikter som kontradikterer

[148] Ross 2002: 19

hverandre, men det kan med enkelthet adapteres til også å angå det anliggende for dette avsnittet.

Dette aktualiserer enkelte spørsmål tilhørende metaetikken. For et slikt tilsvar forutsetter en form for moralsk realisme; at det finnes objektive og uavhengige moralske fakta. Bare da kan man argumentere for at det er viktigere at en normativ teori korresponderer med de moralske fakta enn at den er fleksibel. Da kan man si at det selvfølgelig ville det være praktisk om det var slik Greene etterlyser, at handlinger ikke har moralsk verdi i seg, men kun i den grad de bidrar til å realisere et høyere mål, men hva hjelper det om det ikke er slik den moralske virkeligheten ser ut? Om det virkelig er en moralsk sannhet at det å lyve har en negativ moralsk verdi i seg selv, uavhengig av konsekvensene, så diskvalifiserer det den andre delen av dilemmaet deontologien er konfrontert med i dette kapitelet, kan deontologen hevde.

Selv sier Greene ingenting eksplisitt om hvilken metaetisk bakgrunn hans prosjekt hviler på. Likevel er det enkelte passasjer som tyder på at han ikke nødvendigvis foretrekker konsekvensialismen over deontologien fordi konsekvensialismen er sann og deontologien usann. For som han vedgår i *The Secret Joker of Kant's Soul:* «*I argue that consecuentialist principles, while not true, provide the best available standard for public decision making*»[149]. I *Moral Tribes* følger han også opp dette: *«Morality is not a set of freestanding abstract truths that we can somehow access with out limited human minds.»*[150] Andre steder skriver han igjen som om det det skulle finnes en uavhengig moralsk sannhet:

> «*They will have to say, first, that the correspondence between deontological judgment and emotional engagement is not a coincident and, second, that our moral emotions somehow track the rationally discoverable deontological moral truth.»*[151]

Her virker det dog som at Greene mener dette rent hypotetisk. Med andre ord at det er dette som vil være deontologens byrde om deontologen velger å appellere til objektive moralske sannheter på den måten jeg har foreslått. Det er på mange måter dette Railton gjør når han argumenterer for at de intuisjonene Greene kritiserer er smartere enn vi tror fordi de reflekterer dyrekjøpt erfaring om hva som vanligvis fungerer og ikke fungerer. Men som

[149] Greene 2008: 77
[150] Greene 2013: 329
[151] Greene 2008: 69

distinksjonen mellom modellbaserte- og modellfrie- læring- og beslutningsprosesser illustrerer leder kun tilbake til det originale problemet siden intuisjonene Railton forsvarer er basert på en modellfri strategi. Det betyr nemlig at intuisjonene ikke tilpasser seg et beslutningsmiljø som ikke er konstant, slik at ting som tidligere har vært en god indikator på hvordan en handling fungerer, slik som personlig vold, ikke lenger er det, for eksempel fordi flere typer vold nå er mulig. Om man følger Railtons modell blir derfor byrden nok en gang å forklare hvorfor disse faktorene er moralsk relevante.

Samtidig kan man også snu på det. Hvorfor er det problematisk at deontologien er sensitiv for faktorer som ikke er moralsk relevante om det ikke eksisterer uavhengige moralske fakta? Dette er absolutt et viktig poeng, og det begrenser i det minste Greenes metaetiske valgfrihet. Om en form for ikke-kognitivisme slik som emotivisme for eksempel er den beste forståelsen av naturen til moralske proposisjoner følger det at Greenes kritikk av deontologien er overflødig. På samme tid poengterer Greene at hans kritikk også rammer de som har en mer moderat, antroposentrisk, holdning til moraliteten. Selv om man, som for eksempel Rawls, søker å konstruere moralske prinsipp som heller enn å være sanne er rimelige for oss hjelper ikke det så lenge teorien fortsatt forsøker å forsvare deontologiske intuisjoner. Grunnen til dette er selvfølgelig at disse intuisjonene ikke er fleksible og siden de ikke er fleksible er de kanskje tilpasset en tid der mennesker kun kunne skade hverandre med klubber og spyd, og derfor er disse intuisjonene igjen sensitive for faktorer som ikke er moralsk relevante.[152]

Så hvordan ser Greene for seg at hans foretrukne konsekvensialisme forholder seg til det som nettopp er diskutert? Hans posisjon virker å være at selv om konsekvensialismen nødvendigvis ikke er sannere enn deontologien er den det beste vi har. Med dette later han ikke til å kategorisk utelukke muligheten for at det eksisterer en sann normativ teori som venter på å bli bevist. Men mens vi venter på Godot, som han sier, bør vi gå for det beste alternativet, nemlig konsekvensialisme; eller *deep pragmatism*, som han beskrivende nok også kaller det.[153] Denne agnostisismen tillater Greene langt på vei å unngå metaetiske komplikasjoner. Nettopp derfor har ordlyden i denne oppgaven først og fremst hvert om deontologien er egnet som teori innen normativ etikk eller ei, ikke om den nødvendigvis er

[152] Greene 2008: 75
[153] Greene 2013: 333

sann eller ei. Og for en teori innen normativ etikk er det problematisk at den er sensitiv for faktorer som ikke er moralsk relevante. Jeg har riktig nok argumentert for at denne skaden kan bøtes på ved å forutsette en rigid teori som begrenser uønsket system 1-innflytelse. Men som jeg også har argumentert for er ikke denne formen for rigiditet tilstrekkelig så lenge den ikke samtidig er fleksibel i møte med et beslutningsmiljø som ikke er konstant. Da kan nemlig rigiditeten på et tidspunkt resultere i sensitivitet for faktorer som ikke er moralsk relevante på et annet tidspunkt.

Kapittel 5: Konklusjon

5.1. Kritiserer Greene deontologien i sin beste form, og kan den eventuelt tenkes på en måte som gjør den mindre sårbar for Greenes kritikk?

Greene har unektelig visse metodiske utfordringer som igjen fører til usikkerhet rundt hvorvidt hans kritikk virkelig treffer deontologien i sin beste form. Et av Greenes grunnleggende problem er at han ikke kan teste normative teorier direkte. Han kan kun teste subjekter som på en eller annen måte er antatt å representere den aktuelle normative teorien. Greene gjør dette ved det han kaller *karakteristisk deontologiske dommer*, men problemet med denne metoden er at forbindelsen mellom subjekt og teori i Greenes tilfelle er svak. Grunnen til det er at det også er mulig å gi en deontologisk begrunnelse for å tillate offeret i *switch*, så selv om dette offeret lettest lar seg rettferdiggjøre i konsekvensialistiske termer er det i utgangspunktet ingenting som tilsier at det må det. En sterkere relasjon ville vært en hvor man sikkerhet kunne si at subjektet representerte en konkret teori. Dersom det for eksempel var mulig å hypnotisere testsubjekter, og instruere de i henhold til en gitt deontologisk teori, slik at de nødvendigvis vil gjøre deontologiske vurderinger med utgangspunkt i deontologisk teori, ville dette fra et metodisk perspektiv vært gode nyheter for Greene.

En mulighet er derfor at *karakteristisk deontologiske dommer* kun er dårlige deontologiske dommer. Problemet er at uansett om man forutsetter bedre deontologiske dommer, for eksempel ved å introdusere en faktor som refleksjon, som illustrert ved det jeg har kalt *eksperiment x*, kan Greene enten forklare den forbedrede dommen enten som en konsekvensialistisk dom, eller som en rasjonalisering av en allerede eksisterende system 1-impuls. Dette vil være en rimelig påstand om man i utgangspunktet forutsetter at deontologiske dommer nødvendigvis er et produkt av bakenforliggende system 1-prosesser. Da kan man skyve bevisbyrden foran seg og påstå at den tilfaller den som ønsker å forfekte en alternativ tolkning. Greenes eksperimentelle resultateter viser tross alt at når testsubjektene gjør moralske vurderinger som er i overenstemmelse med, og mest naturlig lar seg forklare av, deontologisk teori så er deler av hjernen som forbindes med system 1-aktivitet høyere enn det det som er tilfelle ved moralske vurderinger som mest naturlig lar seg forklare av konsekvensialistisk teori. Om den beste måten å forklare disse eksperimentelle resultatene på er ved å forutsette en kausal relasjon mellom system 1 og deontologisk etikk er med andre ord

Greene rettferdiggjort i å skyve bevisbyrden foran seg og det er opp til deontologen å presentere en alternativ forklaringsmodell.

Jeg forsøkte å presentere en slik forklaringsmodell ved å vise til det jeg kalte distinksjonen mellom rigide- og fleksible normative teorier. Hovedtanken bak dette var at fleksible teorier har lettere for å akkomodere system 1-impulsene som uansett vil være der, og at deontologiske teorier i motsetning til konsekvensialistiske teorier ofte er fleksible, men at de ikke nødvendigvis må være det. Ved første øyekast virker denne strategien å være en suksess, slik at man ville hatt grunnlag til å konkludere med at Greene ikke nødvendigvis kritiserer deontologien i sin beste form og at deontologien kan tenkes på en måte som gjør den mindre sårbar for Greenes kritikk. Samtidig måtte jeg vedgå at rigide deontologiske teorier er avhengig av å gjøre mange normative antagelser og at denne strategien kun er en suksess dersom det forutsettes at konsekvensialismens tilsvarende antagelser er minst like problematiske.

5.2. Finnes det relevante forskjeller mellom konsekvensialismens grunnleggende intuisjoner og deontologiens tilsvarende intuisjoner, som gjør førstnevnte bedre egnet til å bygge normative teorier på?

Sidgwick operer i *The Methods of Ethics* med fire kriterier moralske intuisjoner må tilfredsstille for at de i størst mulig grad skal være til å stole på. De må (1) være klare og presise, (2) være et resultat av omhyggelig refleksjon, (3) ikke være i konflikt med andre intuisjoner som også tilfredsstiller disse kriteriene og ansees som sanne, i tillegg til (4) at uenighet vedrørende deres sannhetsgehalt må være fraværende eller mulig å avvise på et rasjonelt grunnlag. Grunnen til at intuisjonene som er særegne for deontologien får problemer ifølge Sigwick på grunn av (1) og (4): I den grad det hersker enighet rundt deres sannhetsgehalt er det kun fordi de er tuftet på vage og diffuse formulering, og dersom de skulle formuleres med tilstrekkelig klarhet og presisjon vil det straks vise seg at det likevel ikke hersker enighet rundt hvorvidt de er korrekte eller ei. Men også konsekvensialismens grunnleggende intuisjoner får problem med (4). At lykke er det eneste endelige målet i et moralsk henseende er tross alt høyst kontroversielt, og før Greenes prosjekt eventuelt er endelig bevist kan det heller ikke sies å være en uenighet som kan avvises på rasjonelt grunnlag.

Greene supplerer også Sidgwick med et selvstendig argument. Ifølge han forklarer distinksjonen mellom modellbaserte- og modellfrie- beslutning- og læringsprosesser hvorfor deontologiens grunnleggende intuisjoner er dårligere stilt enn konsekvensialismens tilsvarende intuisjoner. Selv om deontologen ofte passer denne karakteristikken, og klassiske deontologiske intuisjoner som «det er galt å lyve» og «det er galt å dytte fremmede av gangbroer» unektelig er et resultat av en modellfri prosess, finnes det også klare unntak. For deontologiens intuisjoner kan også tenkes på en langt mer sofistikert måte, slik at de har en lignende funksjon som konsekvensialismens grunnleggende intuisjoner, og må være et resultat av en modellbasert prosess.

Deontologiske intuisjoner kan være både rigid slik distinksjonen mellom rigide- og fleksible normative teorier krever og fleksible slik distinksjonen mellom modellbaserte- og modellfrie- beslutning- og læringsprosesser krever, men problemet er at de ikke kan være begge deler samtidig. Det betyr at deontologien uansett vil være sårbar for faktorer som ikke er moralsk relevante. Enten fordi den ikke er rigid og derfor tillater uønsket system 1-innflytelse, eller fordi den baserer seg på intuisjoner som kanskje er designet for å fungere under helt andre omstendigheter enn det som er tilfelle for den daværende situasjonen. Ingen av disse kriteriene er problematisk for konsekvensialismen og dens grunnleggende intuisjoner på samme måte. Derfor kan det konkluderes med at det finnes relevante forskjeller mellom deontologien- og konsekvensialismens grunnleggende intuisjoner som gjør sistnevnte bedre egnet til å tufte normative teorier på.

5.3. Framskritt for normativ etikk

Implikasjonen av at det finnes relevante forskjeller mellom deontologien- og konsekvensialismens grunnleggende intuisjoner som gjør sistnevnte bedre egnet til å tufte normative teorier på er selvfølgelig at distinksjonen mellom rigide- og fleksible normative teorier mislykkes i å representere en alternativ forklaring på Greenes eksperimentelle resultateter. Det betyr igjen det ikke er grunnlag for å påstå at Greene ikke kritiserer deontologien i sin sterkeste form. Likevel er det verdt å merke seg at selv om distinksjonen mellom rigide- og fleksible normative teorier til syvende og sist mislykkes i å gjøre den jobben jeg først foreslo at den kunne gjøre så inngår distinksjonen i et originalt argument for

hvorfor konsekvensialismens grunnleggende intuisjoner bør foretrekkes over deontologiens tilsvarende intuisjoner, noe som er interessant i seg selv.

Siden Greenes prosjekt består mine to kritiske innvendinger konkluderer jeg med at det har potensial til å representere sårt tiltrengt framgang for normativ etikk som disiplin. Når ordlyden er såpass moderat og jeg ikke går lenger i å definere Greenes prosjekt som en suksess er grunnen til at det på ingen måte er umulig å falsifisere Greenes prosjekt. En mulighet er eksempelvis at det finnes en annen alternativ forklaring av Greenes eksperimentelle resultater enn den jeg lanserte, selv om dette per dags dato ikke finnes i litteraturen. Om Greenes prosjekt skal falsifiseres tror jeg likevel dette er den mest sannsynlige måten å gjøre det på. Jeg har tross alt ikke gjort noe for å legge skjul på at Greene har visse metodiske utfordringer, og at disse gjør at tolkningen av de eksperimentelle resultatene kan trekkes i tvil. Samtidig må dette følges opp med en alternativ forklaring på hvorfor de eksperimentelle resultatene er som de er, og hvorfor de eventuelt bør tolkes annerledes. Et av mine bidrag har vært å vise at dette er lettere sagt enn gjort.

Det er heller ikke umulig å se for seg at Greenes prosjekt kan møte utfordringer selv om det består mine kritiske prøver. For eksempel ved at det gjennomføres lignende eksperiment som direkte kontradikterer Greenes eksperimentelle resultater. Inntil det skjer bør vi dog være positive til det potensialet Greenes prosjekt representerer. Innledningsvis argumenterte jeg for at normativ etikk som disiplin har et legitimitetsproblem, og at det er dette problemet som gjør Greenes prosjekt så verdifullt, nettopp fordi det har det potensialet det har til å løse dette problemet, og det har det kun i den grad det består mine to kritiske prøver. Det er denne potensielle verdien som gjør Greenes prosjekt vel verdt en ekstra dose tålmodighet hva angår de utfordringene det måtte stå overfor, slik som eksempelvis de metodiske problemene jeg har adressert i denne oppgaven. Det betyr ikke at vi innbitt skal holde på alle argument som søker å falsifisere deontologisk- eller konsekvensialistisk etikk, kun fordi de i teorien også har et slikt potensial, uansett hvor dårlige de er. Men når et prosjekt er kommet så langt som det Greenes er kommet, og kan vise til de konkrete empiriske resultatene det kan vise til, og i tillegg har det potensialet som det har, så er vi best tjent med å la tvilen komme det til gode og la det få muligheten til å utvikle seg videre.

Litteratur:

- Audi, R. (2005). The Good in the Right. Princeton, Princeton University Press.
- Bentham, J. (2000). An Introduction to the Principles of Morals and Legislation. Kitchener, Batoche Books.
- Berker, S. (2009). "The Normative Insignificance of Neuroscieience." Philosophy & Public Affairs **37**(4).
- Crockett, M. (2013). "Models of morality." Trends in Cognitive Sciences **17**(8).
- Cushman, F., et al. (2006). "The Role of Conscious Reasoning and Intuition in Moral Judgment " Psychological Science **17**(12).
- Cushman, F. (2013). "Action, Outcome, and Value: A Dual-System Framework for Morality." Personality and Social Psychology Review **17**(3).
- Dostojevskij, F. (1993). Brødrene Karamasov. Oslo, Solum Forlag.
- Foot, P. (1967). "The Problem of Abortion and the Doctrine of the Double Effect." Oxford Review(5).
- Joshua Greene, B. S., Leigh Nystrom, John Darley, Jonathan Cohen (2001). "An fMRI Investigation of Emotional Engagement in Moral Judgment." Science **293**.
- Joshua Greene, B. S., Leigh Nystrom, John Darley, Jonathan Cohen (2001). "An fMRI Investigation of Emotional Engagement in Moral Judgment." Science **293**. Supplerende materiell: http://science.sciencemag.org/content/suppl/2001/09/13/293.5537.2105.DC1
- Greene, J. (2008). The Secret Joke of Kants Soul. Moral Psychology. W. Sinnott-Armstrong. Camebridge, Massachusetts, The MIT Press. **3**.
- Greene, J. (2008b). Reply to Mikhail and Timmons. Moral Psychology. W. Sinnott-Armstrong. Camebridge, Massachusetts, The MIT Press. **3**.
- Greene, J., et al. (2009). "Pushing moral buttons: The interaction between personal force and intention in moral judgement." Cognition **111**(3).
- Greene, J. (2013). Moral Tribes. New York, The Penguin Press.
- Greene, J. (2014). "Beyond Point-and-Shoot Morality: Why Cognitive (Neuro)Science Matters for Ethics." Ethics **124**(4).
- Greene, J. D. The rat-a-gorical imperative: Moral intuition and the limits of affective learning. Cognition (2017), http://dx. doi.org/10.1016/j.cognition.2017.03.004
- Haidt, J. (2001). "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." Psychological Review **108**(4).
- Hooker, B. (2000). Ideal Code, Real World. Oxford, Oxford University Press.
- Kahneman, D. (2011). Thinking, Fast and Slow. New York, Farrar, Straus and Giroux.
- Kahane, G. and N. Shackel (2010). "Methodological Issues in the Neuroscience of Moral Judgement." Mind & Language **25**(5).

- Kahane, G., et al. (2012). "The neural basis of intuitive and counterintuitive moral judgment." SCAN **7**.
- Kamm, F. (2009). "Neuroscience and Moral Reasoning: A Note on Recent Research " Philosophy & Public Affairs 37 **37**(4).
- Kamm, F. (2016). The Trolley Problem Mysteries. New York, Oxford University Press.
- Kant, I. (2008). Groundworks of the Metaphysics of Morals. Practical Philosophy. M. Gregor. New York, Camebridge University Press
- Mackie, J. (1977). Ethics: Inventing right and wrong. London, England, Penguin Books.
- McElroy, T. and J. Seta (2003). "Framing Effects: An analytic-holistic perspective." Journal of Experimental Social Psychology **39**.
- Nesse, R. M. and P. C. Ellsworth (2009). "Evolution, Emotion, and Emotional Disorders." American Psychologist **64**(2).
- Nozick, R. (1974). Anarchy, State, and Utopia. Oxford, Blackwell.
- Parfit, D. (1987). Reasons and Persons. Oxford, Oxford University Press.
- Railton, P. (2014). "The Affective Dog ant Its Rational Tale: Intuition and Attunement." Ethics **124**(4).
- Rawls, J. (1999). A Theory of Justice. Cambridge, Massachusetts, Harvard University Press.
- Ross, D. (2002). The Right And The Good. Oxford, Oxford University Press.
- Sandel, M. (2009). Justice. New York, Farrar, Straus and Giroux.
- Schwitzgebel, E. and F. Cushman (2012). "Expertise in Moral Reasoning? Order Effects on Moral Judgment in Professional Philosophers and Non-Philosophers." Mind & Language **27**(2).
- Sidgwick, H. (1907). The Methods of Ethics. London, Macmillian.
- Singer, P. (2005). "Ethics and Intuitions" The Journal of Ethics(9).
- Sinnott-Armstrong, W. (2008). Framing Moral Intuitions. Moral Psychology, volum 2: The Cognitive Science of Morality: Intuition and Diversity. W. Sinnott-Armstrong. Camebridge, Massachusetts, MIT Press. **2**.
- Skelton, A. (2008). "Sidgwick's Philosophical Intuitions." Etica & Politica **X**.
- Skelton, A. (2010). "Henry Sidgwick's Moral Epistemology." Journal of the History of Philosophy **48**(4).
- Smith, S. and I. Levin (1996). "Need for Cognition and Choice Framing Effects." Journal of Behavioral Decision Making **9**.
- Suter, R. S. and R. Hertwig (2011). "Time and moral judgment." Cognition(119)
- Takemura, K. (1993). "Influence of Elaboration on the Framing of Decision." The Journal of Psychology **128**(1).
- Thomson, J. J. (1976). "Killing, Letting Die, and The Trolley Problem." The Monist **59**(2).
- Thomson, J. J. (1985). "The Trolley Problem." The Yale Law Journal **94**(6).

- Thomson, J. J. (2008). "Turning the Trolley." Philosophy & Public Affairs **59**(4).
- Tversky, A. and D. Kahneman (1986). "Rational Choice and the Framing of Decisions." The Journal of Buisness **59**(4).