Włodek Rabinowicz

Utylitaryzm preferencji poprzez zmianę preferencji?

Analiza i Egzystencja 12, 7-37

2010

„Analiza i Egzystencja" 12 (2010)
ISSN 1734-9923

ARTYKUŁY

WLODEK RABINOWICZ*

UTYLITARYZM PREFERENCJI POPRZEZ ZMIANĘ PREFERENCJI?**

Słowa kluczowe: Hare, utylitaryzm preferencji, zasada refleksji, Vendler, Persson, Elstein
Keywords: Hare, preference utilitarianism, principle of reflection, Vendler, Persson, Elstein

* Wlodek Rabinowicz urodził się w Warszawie, tam też studiował filozofię w latach 60. Relegowany z Uniwersytetu Warszawskiego po wypadkach marcowych, wyemigrował do Szwecji w 1969 r. Kontynuował studia w Uppsali, gdzie obronił doktorat na temat zasady uniwersalizacji i pracował potem jako docent do 1995 r., gdy został mianowany profesorem na uniwersytecie w Lundzie. W latach 1999–2002 był przewodniczącym Europejskiego Towarzystwa Filozofii Analitycznej, a w latach 2005–2007 przewodniczącym Szwedzkiego Towarzystwa Filozoficznego. Jest redaktorem czasopisma „Theoria" i byłym redaktorem „Economics and Philosophy". Członek Institut International de Philosophie, Szwedzkiej Akademii Nauk i Academia Europaea. Ma liczne publikacje z dziedziny etyki, teorii wartości, teorii decyzji i logiki filozoficznej w takich czasopismach, jak „Journal of Philosophy, Ethics, Theory and Decision", „Synthese", „Erkenntnis", „Philosophy of Science", „Proceedings of Aristotelian Society", „Journal of Philosophical Logic", „Economics and Philosophy", „Analysis", „Utilitas", „Philosophy and Phenomenological Research", „Philosophical Quarterly", „Theoria" i „Polish Journal of Philosophy".

** Angielski oryginał tego artykułu – *Preference Utilitarianism by Way of Preference Change?* – ukazał się w: *Preference Change: Approaches from Philosophy, Economics and Psychology*, Seria: Theory and Decision Library A, vol. 42, red. T. Grüne-Yanoff, S.O. Hansson, Springer 2009, s. 185–206. Tekst jest publikowany za życzliwą zgodą Autora i wydawnictwa Springer. Translation is published by permission of the author and publisher.

W niniejszym artykule dokonam ponownego przeglądu klasycznego i często dyskutowanego argumentu Richarda Hare'a za utylitaryzmem preferencji (Hare 2001). Argument ten opiera się na koncepcji namysłu moralnego jako eksperymentu myślowego, któremu towarzyszy zmiana preferencji. Argument jest problematyczny pod różnymi względami. Tutaj skupię się głównie na jednej z trudności: na pozornej luce w rozumowaniu Hare'a, która może być nazwana „problemem braku konfliktu". W artykule, który napisałem kilka lat temu z Bertilem Strömbergiem, staraliśmy się tę lukę wypełnić (Rabinowicz, Strömberg 1996). Nasza propozycja koncentrowała się na idei, iż namysł moralny wymaga rewizji preferencji: dany stan preferencji musi ulec zmianie, tak aby spełnił pewien wymóg jednolitości. Z tego względu rewizji preferencji powinno się nadać teoretyczne ramy. Zasugerowaliśmy, że można założyć, iż taka rewizja kieruje się zasadą minimalnej zmiany: stan wyjściowy musi spełniać nałożone ograniczenia (w naszym przypadku ograniczenie jednolitości) oraz winien się on jak najmniej różnić od stanu wejściowego. Jeżeli miara odległości między stanami preferencji zostanie wybrana we właściwy sposób, można wykazać, iż stan wyjściowy jest zgodny z utylitaryzmem. Jak się jednak okazuje, propozycja ta prowadzi do wielu trudności: już sam wybór odpowiedniej miary odległości pomiędzy stanami preferencji stanowi poważny problem. Nie jest również oczywiste, czy zasada minimalnych zmian, którą zakłada się często jako główną zasadę kierującą rewizją przekonań, może być zastosowana do zmiany preferencji. Te i inne problemy podają naszą pierwotną propozycję w wątpliwość. Z tego względu rozważę również alternatywne rozwiązanie, które ostatnio zasugerował mi Daniel Elstein. Propozycja Elsteina jest bliższa sposobowi rozumowania Hare'a i może okazać się najlepszym sposobem na wypełnienie luki w jego argumencie.

W artykule napisanym ze Strömbergiem badamy także, czy owa luka rzeczywiście istnieje: może rozważany problem powinien zostać usunięty, a nie rozwiązany? Propozycja ta sięga pomysłu Zeno Vendlera (Vendler 1988). Okazuje się niestety, że sugestia Vendlera nie chroni Hare'a przed krytyką. Oddala problem braku konfliktu, powodując jednak inne, potencjalnie poważniejsze, trudności[1].

[1] Jestem wdzięczny Danielowi Elsteinowi, Christianowi Listowi, Toniemu Rönnow-Rasmussenowi, Markowi Schroederowi i Bertilowi Strömbergowi za bardzo przydatne komentarze. Wcześniejsze wersje tego artykułu zostały przedstawione podczas warsztatów na temat zmiany preferencji, zorganizowanych z okazji kongresu Gesellschaft für Analytische Philosophie w Berlinie w roku 2006, a następnie na spotkaniu British Society

1. Argument Hare'a i związany z nim problem

Argument Hare'a opiera się na jego interpretacji sądu moralnego jako uniwersalnej nadrzędnej preskrypcji[2]. Zasada uniwersalizowalności powoduje, iż zalecenie moralne dotyczące konkretnej sytuacji stosuje się również do jej hipotetycznych wariantów, w których osoby zamieniły się rolami. Aby wydać moralny sąd dotyczący rozważanej sytuacji, muszę najpierw rozpatrzeć wszystkie owe warianty. Tak więc załóżmy, iż rozważam działanie, które oprócz mnie dotyczy również innych osób, powiedzmy Jana i Marii. Sąd, iż powinienem to działanie wykonać, zaleca je zarazem w sytuacji, w której jestem osobą działającą *oraz* w hipotetycznych sytuacjach, w których będę zajmował miejsce osób, których owo działanie dotknie. W konsekwencji, aby wydać osąd moralny, muszę zapytać siebie: Co zrobiłbym, jeśli byłbym w skórze Jana lub Marii? Jak by to było, gdybym był wystawiony na to działanie? Sądy moralne, ze względu na ich uniwersalne zastosowanie, muszą być oparte na eksperymentach myślowych tego rodzaju lub być przetestowane przez takie eksperymenty.

Sformułowanie „być w czyjejś skórze" jest w tym kontekście nieco mylące. Uniwersalizowalność zobowiązuje mnie jedynie do rozszerzenia mojej preskrypcji z danej sytuacji na jej dokładnie podobne warianty. Jak ujął to Hare:

> jeśli teraz mówię, że powinienem coś komuś zrobić, to zobowiązuje mnie to do poglądu, iż to samo powinno się zrobić mnie, gdybym znalazł się w dokładnie takiej sytuacji jak ten ktoś i miał te same cechy osobowe, a w szczególności takie same stany motywujące. Jednakże czyjeś stany motywujące mogą przecież całkowicie kłócić się z moimi (Hare 2001: 139).

Dlatego też, kiedy wyobrażam sobie, jak to jest być Janem, muszę założyć, że w tej hipotetycznej sytuacji nie tylko jestem w zewnętrznych okolicznościach, w których znajduje się Jan, ale także posiadam jego ciało, psychikę,

for Ethical Theory w Edynburgu w 2008 r. Chciałbym podziękować uczestnikom tych spotkań oraz organizatorom: Tillowi Grüne-Yanoffowi i Svenowi Ove Hanssonowi oraz Ellinorze Mason i Michaelowi Ridge'owi. Na koniec pragnę podziękować recenzentom tego tomu, Peterowi Dietschowi oraz Martinowi Petersonowi, którzy obaj uprzejmie zgodzili się ujawnić swoją tożsamość. Ich uwagi były bardzo pomocne.

[2] Zob. Hare (2001), s. 74.

charakter, przekonania, uczucia i pragnienia – wchodzę nie tylko w jego skórę, ale przejmuję całe jego wnętrze. Staram się wyobrazić sobie, jak to by było być wystawionym na dane działanie, jeśli byłbym w sytuacji, w której on jest aktualnie.

Aby uczynić te rozważania mniej abstrakcyjnymi, dodajmy do naszego przykładu kilka szczegółów. Załóżmy, że zgodziłem się na spotkanie z Janem i Marią, dwojgiem moich studentów, dzisiaj w instytucie. Nie ustaliliśmy konkretnego czasu naszego spotkania, ale sekretarka zadzwoniła do mnie do domu i poinformowała, że studenci właśnie przybyli i na mnie czekają. Ponieważ pogoda jest ładna, zdecydowanie wolałbym pojechać do pracy rowerem niż samochodem. Studenci jednak są niechętni oczekiwaniu: woleliby, abym przyjechał możliwie jak najszybciej. Rozwiązanie oparte na utylitaryzmie preferencji zalecałoby działanie, które najlepiej zrównoważy nasze preferencje: siła mojej preferencji, aby jechać rowerem jest ważona względem siły przeciwnych preferencji studentów. Załóżmy, że każda z osobna preferencja studentów jest słabsza niż moja, jednak razem ważą od niej więcej. W tych okolicznościach powinienem powstrzymać się od jazdy rowerem i pojechać samochodem.

Równoważenie presuponuje, że siła ludzkich preferencji może być porównywana i mierzona na wspólnej skali. Oczywiście założenie to jest bardzo kontrowersyjne, ale, ponieważ Hare traktuje je jako mniej lub bardziej oczywiste (patrz rozdz. 7 [w:] Hare 2001, s. 157), zamierzam uczynić to samo, przynajmniej ze względu na rozważany argument. Załóżmy więc, że siła moich preferencji, aby pojechać rowerem jest równa +4, natomiast intensywność przeciwstawnych preferencji Jana i Marii to odpowiednio –3 i –2. Znaki plus lub minus określają kierunek preferencji – czy jest ona za lub przeciw branemu pod uwagę działaniu. Rachunek utylitaryzmu preferencji implikuje, że powinienem powstrzymać się od możliwości jazdy rowerem: $+4 -3 -2 < 0$.

Hare pragnie dowieść, że osiągnę taki sam rezultat, jeśli do oceny moralnej dojdę poprzez eksperymenty myślowe, w których przyjmę pozycję różnych osób zaangażowanych w rozważaną sytuację. Jeśli postąpię w ten sposób, oraz jeśli jestem racjonalny, dobrze poinformowany i posiadam wystarczającą wyobraźnię, nie mogę nie uzyskać moralnego zalecenia, które jest wynikiem utylitaryzmu preferencji.

W jaki sposób Hare opisuje proces, który prowadzi do moralnego sądu? Pozwolę sobie zacząć od cytatu z *Myślenia moralnego*. Opisuje

on tam „dwustronny" przykład, w którym ja – podmiot – rozważam, czy przesunąć czyjś rower w celu stworzenia miejsca na parkingu dla mojego samochodu. Żadne dodatkowe osoby nie są w to wmieszane. Utylitaryzm preferencji zakłada, że powinienem przesunąć rower wtedy i tylko wtedy, gdy moja preferencja, by wykonać to działanie, jest silniejsza niż preferencja rowerzysty przeciwko przemieszczaniu roweru. Hare komentuje:

> Nie widzę powodu, żeby w tej sytuacji nie przyjąć rozwiązania takiego jak wtedy, gdy w grę wchodzi konflikt naszych własnych różnych preferencji. Zmodyfikujmy na przykład sytuację i przyjmijmy, że chodzi o mój motocykl i że przemieszczenie go jest umiarkowanie kłopotliwe, ale wysoce kłopotliwa jest niemożność zaparkowania mojego samochodu; motocykl oczywiście przestawię, przyjmując, że właśnie to – mówiąc rozważnie – powinienem uczynić czy że to w sumie chcę zrobić. Powróćmy teraz do relacji dwustronnej (Hare 2001: rozdz. 5.5, 122–125): ustaliliśmy, że jeśli mam pełną wiedzę na temat preferencji drugiej osoby, to sam zyskuję preferencje – równe jej preferencjom – względem tego, co powinno być zrobione mnie, gdybym znalazł się w jej sytuacji; i są to preferencje, które teraz kłócą się z moim pierwotnym nakazem. Mamy więc w efekcie nie interpersonalny konflikt preferencji czy nakazów, lecz konflikt intrapersonalny; obie ścierające się preferencje należą do mnie. Potraktuję go zatem tak, jakby dotyczył dwóch moich preferencji.
>
> Tak oto przypadki relacji wielostronnych [które dotyczą wielu osób] przedstawiają w tej chwili mniejszą trudność, niż wydawało się na początku. Również tutaj, bez względu na złożoność konfliktu i liczbę osób, wszystko sprowadza się, przy pełnej wiedzy na temat preferencji innych, do konfliktów intrapersonalnych (Hare 2001: rozdz. 6.2, 140).

Spróbujmy teraz przeanalizować ten fragment, używając przykładu, którym rozpoczęliśmy nasze rozważania. Zastanawiam się nad pojechaniem rowerem do pracy w sytuacji, którą nazwiemy s_1. Posiadam preferencję *dla* tego działania o sile 4. Jednakże, ponieważ sądy moralne są uniwersalne, zalecają one dokładnie podobne działania dla dokładnie podobnych sytuacji. W konsekwencji sąd moralny dotyczący tego, co powinienem uczynić w s_1 ma również zastosowanie do hipotetycznej sytuacji, w której role zostałyby odwrócone. Dlatego muszę sobie wyobrazić, że jestem odpowiednio w skórze Jana i Marii, tj. unaocznić sobie dwie hipotetyczne sytuacje s_2 i s_3, w których znajduję się jako osoba doświadczająca działania. Zdaję sobie

sprawę, że jeśli byłbym na miejscu Jana z jego pragnieniami itp., miałbym taką samą preferencję, jaką w sytuacji aktualnej ma Jan: *przeciwko* rozważanemu działaniu o sile 3. Analogicznie, jeśli byłbym w skórze Marii, posiadałbym preferencję *przeciwko* działaniu o sile 2.

Następny krok w procesie namysłu zakłada to, co Allan Gibbard nazwał „zasadą warunkowej refleksji" (Gibbard 1988). Zasadę tę wprowadza sam Hare, nie nadając jej jednak żadnej nazwy.

> *Warunkowa refleksja:* O ile w pełni wiem, co preferowałbym w sytuacji hipotetycznej, muszę posiadać odpowiednią preferencję (ten sam znak, taka sama siła) w odniesieniu do tej sytuacji hipotetycznej[3].

Innymi słowy, moje hipotetyczne preferencje – jeśli wiem, że w sytuacji hipotetycznej bym je posiadał – znajdują odzwierciedlenie w moich rzeczywistych preferencjach dotyczących danej sytuacji. Jeśli teraz dostrzegam, że gdybym był na miejscu Jana, to miałbym preferencję przeciwko jeździe rowerem, zyskuję preferencję przeciwko temu działaniu w odniesieniu do tej hipotetycznej sytuacji.

Hare traktuje zasadę warunkowej refleksji jako prawdę pojęciową. Obowiązuje ona ze względu na rzekome występowanie elementu preskryptywnego w samym pojęciu „ja": „Sugeruję, że 'ja' nie jest słowem całkowicie opisowym, ale w części nakazowym". (Hare 2001: 126)

> [...] nazywając jakąś osobę „ja", wyrażam zdecydowanie większą troskę o jej preferencje niż o preferencje tych ludzi, których tak nie nazywam. A zatem gdyby w normalnym, jednoznacznym wypadku, gdy ktoś jest maltretowany i to mu się nie podoba, zapytano mnie: „A gdyby tak ciebie postawić w jego sytuacji wraz z jego preferencjami?", odpowiedziałbym, że jeśli to byłbym *ja*, to teraz odczuwam tę samą niechęć, aby być wystawionym na to działanie, jaką on czuje obecnie (Hare 2001: 128). [Tekst tłumaczenia jest częściowo zmieniony – K.S.]

[3] Por. Hare (2001), s. 129: „[...] nie mogę poznać skali i jakości cudzych cierpień oraz, ogólnie rzecz biorąc, motywacji i preferencji, jeżeli nie kieruje mną taka sama motywacja w odniesieniu do tego, co mogłoby mnie spotkać, gdybym znalazł się na miejscu kogoś innego wraz z jego motywacjami i preferencjami". Ta sama zasada została również nazwana „zasadą hipotetycznego samouznania" (Persson 1989) oraz „zasadą warunkowego samouznania" (Rabinowicz 1989 i Rabinowicz, Strömberg 1996).

Warunkowa refleksja ugruntowana jest na mojej elementarnej trosce własnej, którą Hare interpretuje jako troskę o spełnienie moich preferencji, zarówno faktycznych, jak i hipotetycznych. Myśląc o hipotetycznych preferencjach jako *własnych*, tym samym je popieram. Czy jest to twierdzenie przekonujące? Można mieć wątpliwości: takie poparcie preferencji wydaje się zawieszone, kiedy rozważam hipotetyczne sytuacje, w których moje preferencje w świetle mego *obecnego* rozeznania wydają się wadliwe lub w jakiś sposób zniekształcone. Gdybym miał sadystyczne usposobienie, chciałbym zadawać ból. Ale wiedza ta wcale nie czyni, że pragnę, abym zadawał ból, gdybym był sadystą: nie popieram mojej hipotetycznej preferencji, jeśli obecnie oceniam, iż jest ona wadliwa lub nieracjonalna.

Sugeruje to, że jeśli warunkowa refleksja miałaby zostać przyjęta, powinna zawierać co najmniej pewne ograniczenia. Co więcej, możliwe, że należałoby ją interpretować nie jako prawdę pojęciową, lecz jako wymóg racjonalności: podczas gdy troska własna, która leży u podstaw warunkowej refleksji, wydaje się ważnym elementem racjonalności, można z powodzeniem kwestionować, czy postawa taka gra rolę w określaniu znaczenia terminów takich, jak „ja" czy „moje". Jeśli warunkową refleksję postrzegamy jako wymóg racjonalności, może być ona interpretowana jako warunek, który winny spełniać preferencje idealnie zintegrowanego i pewnego siebie podmiotu[4]. Niemniej jednak, przynajmniej teraz możemy pozostawić tę zasadę w formie, w jakiej została zaprezentowana. W każdym razie jest oczywiste, że zarzut z „wadliwości" w naszym przykładzie nie

[4] Co łatwo zauważyć, ograniczająca preferencje warunkowa refleksja jest ściśle związana z dobrze znaną zasadą refleksji w odniesieniu do przekonań. Zgodnie z tą ostatnią, wiedza o tym, jakie ktoś miałby przekonania w hipotetycznej sytuacji, zobowiązuje go do przyjęcia analogicznego i równie silnego przekonania warunkowego dotyczącego danej sytuacji. Bas van Fraassen wykazał, że osoba, która narusza tę zasadę, jest narażona na zarzut „Dutch Book", jeśli tylko przypisuje ona pozytywne prawdopodobieństwo zdarzeniu się danej sytuacji hipotetycznej (por. van Fraassen 1984). W Rabinowicz (1989) sugeruję, że podobny argument odwołujący się do „Dutch Book" może być użyty, aby uzasadnić analogiczną zasadę refleksji dla preferencji. Jednakże, co z łatwością przyznałby Hare, prawdopodobieństwo, iż zajmę dokładnie taką samą pozycję jak ktoś inny zajmuje w aktualnej sytuacji, jest równe zeru. Z tego względu nie jest możliwe użycie tego argumentu odwołującego się do zarzutu „Dutch Book" przeciw komuś, kto w odniesieniu do takiej hipotetycznej sytuacji narusza zasadę warunkowej refleksji. Jeśli chce się bronić warunkowej refleksji, choćby dla tych eksperymentów myślowych, to obrona nie może się opierać na takich czysto pragmatycznych podstawach.

ma zastosowania: preferencje studentów, abym przyjechał wcześniej, są w pełni rozsądne.

Warunkowa refleksja zakłada, że po rozważeniu, jak to byłoby być w skórze moich studentów, dochodzę do kilku preferencji w odniesieniu do rozważanych działań – tak wielu, jak wiele sytuacji muszę wziąć pod uwagę. Nadal posiadam moją pierwotną preferencję dla jazdy rowerem o sile +4, ale teraz – po rozważeniu hipotetycznych sytuacji s_2 i s_3 – również nabywam dwie preferencje przeciw temu działaniu, odpowiednio o mocy –3 i –2.

W cytowanym powyżej fragmencie Hare wydaje się sugerować, że ostatni krok w procesie dochodzenia do moralnego osądu polega na roztropnościowym *równoważeniu.* Oto jestem z preferencjami, które wiodą mnie w przeciwstawnych kierunkach – w stronę czynu oraz z dala od niego. Moja racjonalna preferencja „końcowa", jak ujął to Hare, jest funkcją owych preferencji wejściowych. W naszym przykładzie oznacza to, że dochodzę do końcowej preferencji, aby nie jechać rowerem: moje preferencje przeciw temu działaniu są wspólnie silniejsze niż moja preferencja, aby pojechać rowerem. Dzięki eksperymentom myślowym, które doprowadziły mnie do uzyskania nowych preferencji, a następnie dzięki zbalansowaniu tych preferencji z moją pierwotną preferencją, wydaję się osiągać takie samo rozwiązanie jak to, które jest wynikiem utylitaryzmu preferencji. Hare pragnie podkreślić, że jego rekonstrukcja procesu namysłu moralnego przekształca pierwotny *między*osobowy konflikt preferencji w konflikt *wewnętrzny*. Ten ostatni zaś jest rozwiązywalny w standardowy sposób – poprzez proste równoważenie preferencji.

Jednakże, jak zauważyli Schueler (1984) oraz Persson (1989), argument ten – w formie, w jakiej został przedstawiony powyżej – zawiera istotną lukę. Hare'a porównanie namysłu moralnego do standardowych problemów decyzji, w których podmiot przeżywa konflikt preferencji, jest zwodnicze. W standardowym przypadku moje sprzeczne pragnienia dotyczą jednej i tej samej sytuacji – tej, w której dokonuję wyboru. W argumencie Hare'a natomiast różne preferencje, które nabywam poprzez eksperymenty myślowe, nie są powiązane w ten sposób. Posiadam preferencję, aby jechać rowerem w odniesieniu do rzeczywistej sytuacji s_1, preferencję przeciw temu działaniu w odniesieniu do hipotetycznej sytuacji s_2, w której jestem w skórze Jana, oraz jeszcze jedną preferencję przeciw temu działaniu w odniesieniu do s_3, w której zajmuję miejsce Marii. Moje pragnienia dotyczą różnych sytuacji i dlatego nie stoją ze sobą w konflikcie. Inaczej niż w przypadku, w którym

w grę wchodzi roztropność, nie istnieje tutaj konflikt preferencji, który muszę rozwiązać poprzez równoważenie. Załóżmy zatem, że zdecydowałbym się udać do pracy na rowerze. Działanie to spełniłoby moją preferencję w odniesieniu do rzeczywistej sytuacji s_1, ale w żaden sposób nie stałoby w konflikcie z moimi preferencjami dotyczącymi czysto hipotetycznych sytuacji s_2 i s_3. To właśnie, krótko mówiąc, jest „problem braku konfliktu", który zagraża argumentowi Hare'a.

Ale czy o czymś nie zapomnieliśmy? Co z zasadą uniwersalizowalności? Jeśli preskrypcja ma być zaleceniem moralnym, uniwersalizacja wymaga, aby w odniesieniu do różnych rozważanych sytuacji s_1, s_2 i s_3 preskrypcja ta była *jednolita*. Zatem tak długo, jak długo moje preferencje dotyczące s_1 różnią się od tych dotyczących s_2 i s_3, nie doszedłem jeszcze do sądu moralnego. Chociaż w wyżej cytowanym fragmencie nie jest to widoczne, Hare w innym miejscu wydaje się sugerować, że jednolita preskrypcja może zostać osiągnięta przez proces *próbnej ekstrapolacji*: staram się przenosić moje preferencje dotyczące jednej sytuacji na inne. Pytanie dotyczy wtedy tego, czy przenoszona preferencja jest wystarczająco silna, aby przetrwać wszelkie konflikty preferencji, które przez ten manewr mogą być stworzone[5]. Jeśli nie jest, zamiast niej staram się przenieść jedną z moich innych preferencji – jedną z tych, które nabywam w odniesieniu do sytuacji, w której role zostają odwrócone. Czy powyższe rozwiązanie może nam tu pomóc?

[5] „[...] jeśli teraz mówię, że powinienem coś komuś zrobić, to zobowiązuje mnie to do poglądu, iż to samo powinno się zrobić mnie, gdybym znalazł się w dokładnie takiej samej sytuacji jak ten ktoś. Jednakże [...] on na przykład może bardzo nie chcieć, żebym zrobił mu to, co – jak twierdzę – powinienem mu zrobić. [...] Jeśli w pełni przedstawiam sobie jego sytuację, w tym też jego motywacje, sam zyskuję podobną motywację, co znajduje wyraz w nakazie, że nie zostanie mi to zrobione, gdybym natychmiast znalazł się w tej właśnie sytuacji. Ale ten nakaz jest sprzeczny z moim pierwotnym stwierdzeniem 'powinno się'.[...] Mogę uniknąć tej sprzeczności w 'naszej własnej woli' (por. Kant 1785: 58) jedynie porzucając moje pierwotne stwierdzenie 'powinno się', zważywszy na moją obecną wiedzę na temat sytuacji, w której mógłbym być ofiarą" (Hare 2001: 139, pkt 6.2.). Należy przyznać, że powyższy fragment nie jest krystalicznie jasny. Kiedy moja pierwotna preferencja jest najpierw zuniwersalizowana, tj. przeniesiona na sytuację, w której jestem przedmiotem działania, i kiedy ta uniwersalna preferencja jest następnie porzucona w celu uniknięcia „sprzeczności w woli", Hare nie stwierdza wyraźnie, że ta próba uniwersalizacji preferencji zostaje przerwana ze względu na to, iż przeciwna preferencja jest silniejsza.

Pomysł z ekstrapolacją faktycznie jest pomocny, ale tylko w przypadkach *dwustronnych*. Jeśli tylko jeden student, powiedzmy Jan, czeka na mnie w instytucie, istnieją tylko dwie możliwości, o które muszę się martwić: rzeczywista sytuacja s_1 oraz hipotetyczna sytuacja s_2, w której jestem w skórze Jana. W takim przypadku moją preferencję, aby jechać rowerem, mogę z powodzeniem przenieść z s_1 do s_2, ponieważ jest ona silniejsza niż moja przeciwna preferencja dotycząca s_2, którą nabyłem zgodnie z warunkową refleksją. Gdyby natomiast ta ostatnia preferencja była silniejsza, wtedy byłbym w stanie z powodzeniem przenieść ją zamiast pierwszej. W związku z tym mogę utrzymać jednolite zalecenie w odniesieniu do obydwu sytuacji i będzie miało ono w pełni utylitarny charakter.

Jednakże zaproponowane rozwiązanie w przypadkach, w których zaangażowanych jest wiele osób, prowadzi nas na manowce (por. Persson 1989). Rozważmy zatem ponownie przykład dwojga studentów, którzy czekają na mnie w instytucie. Nietrudno spostrzec, że moja preferencja w odniesieniu do s_1, aby jechać rowerem, może być z powodzeniem przeniesiona zarówno do s_2, jak i do s_3. Jest ona silniejsza niż moje wzięte z osobna przeciwstawne preferencje dotyczące tych dwóch sytuacji, choć jest ona słabsza niż one obie wzięte razem. Przeniesiona preferencja wygrywa dlatego, że za każdym razem spotyka tylko jedną przeciwstawną preferencję. Przeciwstawne preferencje, jeśli można tak powiedzieć, nigdy nie mają możliwości połączenia sił. Jednoznaczne zalecenie, aby pojechać rowerem, pozostaje więc niepokonane, pomimo jego nieutylitarystycznego charakteru.

Persson (1989) sugeruje, że luka w argumencie Hare może zostać wypełniona przez wprowadzenie „zasłony niewiedzy" – narzędzia, które zostało rozsławione przez Johna Rawlsa i Johna Harsanyiego. Zasłona niewiedzy Perssona jest identyczna z modelem równych prawdopodobieństw Harsanyiego (patrz Harsanyi 1953 i 1977): po nabyciu preferencji dotyczących trzech sytuacji s_1, s_2 i s_3, powinienem udać, iż nie jestem pewny, która z tych trzech sytuacji jest sytuacją rzeczywistą. Jak sugeruje Harsanyi, niepewność powinna być reprezentowana jako przypisanie równych prawdopodobieństw (a nie jako pełna niewiedza, czyli brak znajomości prawdopodobieństw, jak ma to miejsce u Rawlsa). Powinienem zatem traktować owe trzy sytuacje tak, jak gdyby były one równie prawdopodobne. Podobnie jak Harsanyi, w celu zidentyfikowania działań, które mają być wykonane, Persson w następnym kroku stosuje standardową zasadą maksymalizacji oczekiwanej użyteczno-

ści[6]. W naszym przykładzie oznacza to, że powinienem powstrzymać się od jazdy do pracy rowerem. Działanie to zaspokoi moje preferencje, jeśli s_1 jest sytuacją rzeczywistą, ale zapobiegnie ich spełnieniu, jeśli zamiast tego jedna z dwóch pozostałych sytuacji jest rzeczywista, co, biorąc pod uwagę moją pozorowaną ignorancję, jest dwa razy bardziej prawdopodobne. Tak jak model Harsanyiego, koncepcja ta zapewnia standardowe utylitarystyczne rozstrzygnięcie.

Propozycja Perssona nie została przyjęta przez Hare'a, mimo że, podobnie jak Rawls i Harsanyi, Hare również pragnie ugruntować moralność na koncepcji racjonalnego wyboru. Jednakże, w odróżnieniu od nich, Hare w swojej racjonalnej rekonstrukcji rozumowania moralnego pragnie uniknąć jakichkolwiek elementów pozoru czy udawania. Jak zaznacza sam Persson, użycie zasłony niewiedzy, biorąc pod uwagę Hare'a, projekt aby ufundować etykę na całkowicie racjonalnych podstawach, wprowadziłoby obcy element do jego myśli:

> [...] dodanie PEP [= zasady równego prawdopodobieństwa] do założeń Hare'a wydaje się wysoce problematyczne, ponieważ gdy racjonalność [...] domaga się, aby preferencje były formowane na podstawie wszystkich istotnych dostępnych dla podmiotu informacji, PEP wymaga od podmiotu abstrahowania od niektórych informacji (dotyczących tożsamości numerycznej zaangażowanych osób) (Persson 1989, s. 170).

Wyrazić można to również w taki oto sposób: od udawania się zaczyna, na udawaniu się kończy. Jeśli udajemy, że akceptujemy przesłanki, będziemy również tylko udawali, że akceptujemy wniosek[7]. Dlatego też nie dziwi, iż Hare w komentarzu do artykułu Perssona próbuje w inny sposób wypeł-

---

[6] Harsanyi uważa, że takie udawanie ignorancji pasuje do idealnego obserwatora. Trudniej jest zrozumieć, w jaki sposób takie udawanie może być w ogóle możliwe w kontekście praktycznego namysłu moralnego: kiedy podejmuję decyzję, czy wykonać działanie, nie mogę w tym samym czasie udawać, że – o tyle, o ile wiem – możliwe jest, iż będę przedmiotem rozważanego przeze mnie czynu!

[7] Można postawić zarzut, że nawet podejście samego Hare'a zawiera element udawania. Ostatecznie, czy nie jest tak, że eksperymenty myślowe odgrywają kluczową rolę w jego koncepcji moralnego namysłu? Zarzut taki bazowałby jednak na nieporozumieniu. W koncepcji Hare'a podmiot proszony jest o rozważenie, co by się stało, jeśli role zostałyby odwrócone. Nie jest on zmuszony udawać, że – o tyle, o ile wie – ta hipotetyczna sytuacja być może jest *rzeczywista.* To tutaj rozchodzą się drogi Hare'a z Rawlsem i Harsanyim.

nić ową lukę (por. Hare 1989). Nie będę rozważał jego propozycji z racji ograniczonego miejsca. Pozwolę sobie jedynie powiedzieć, że uważam ją za niezadowalającą[8].

2. Rewizja preferencji

W tym miejscu pozwolę sobie przejść do propozycji, którą w innym artykule przedstawiłem wraz ze Strömbergiem. Wróćmy do punktu, w którym moje eksperymenty myślowe doprowadziły mnie do uzyskania zestawu preferencji dotyczących trzech sytuacji $s_1 - s_3$. Biorąc pod uwagę rozważane działanie, na tym etapie mój profil preferencji może być reprezentowany przez wektor

$$(+4, -3, -2),$$

w którym pierwszy element określa siłę mojej preferencji dotyczącej s_1, drugi element określa siłę mojej preferencji dotyczącej s_2 itd. Znaki plus lub minus określają kierunek preferencji – czy jest ona za lub przeciw rozważanemu działaniu. Na podstawie tego profilu muszę teraz dotrzeć do moralnego sądu, tj. do uniwersalnego zalecenia, aby bądź pojechać rowerem, bądź z tego zrezygnować. Zalecenie to musi być takie samo dla wszystkich trzech sytuacji.

Główna idea, na której oparta jest nasza propozycja, może być sformułowana w następujący sposób: uniwersalne zalecenie, które należy osiągnąć, powinno być możliwie jak najbardziej zgodne z pierwotnym profilem preferencji podmiotu. Pomysł ten może zostać sprecyzowany na kilka sposobów. Dwa z nich naszkicowaliśmy w naszym artykule. To co nazywamy podejściem „rewizji preferencji", odróżniamy od podejścia „ostatecznego werdyktu". W tym miejscu skupię się jedynie na pierwszym z nich.

Dla Hare'a preferowanie i zalecanie są zasadniczo jednym i tym samym. „Wszystkie zalecenia, włączając w to zalecenia moralne, są wyrazami preferencji lub w szerokim sensie pragnień" (Hare 2001, s. 231, zob. także s. 137)[9].

---

[8] Omówienie i krytyczną dyskusję znaleźć można w Rabinowicz, Strömberg (1996), rozdział 3.

[9] Zob. także Hare (1987), s. 73: „Chcieć, aby coś się stało, to być w stanie umysłu, w którym, jeśli miałbym to wyrazić słowami, akceptuje się zalecenie, aby się to stało". Ostatecznie pomysł ten sięga do Hare (1963), sekcja 9.4.

Tak więc, gdy próbuję dotrzeć do jednolitego zalecenia dla naszych przykładowych trzech sytuacji, to szukam jednolitej preferencji w odniesieniu do tych sytuacji[10]. Innymi słowy, staram się zrewidować swoje pierwotne preferencje, które różnią się w odniesieniu do trzech sytuacji, aby dotrzeć do nowego stanu preferencji o jednolitym profilu:

$$(x, x, x)$$

W powyższym wektorze w każdym miejscu występuje ta sama wartość (dodatnia lub ujemna). W poszukiwanym stanie preferencji posiadam dokładnie taką samą preferencję za lub przeciw działaniu w odniesieniu do każdej z trzech sytuacji $s_1 - s_3$.

Jak należy postąpić, aby w ten sposób zmodyfikować moje preferencje? Jaka jest właściwa wartość x?

Zmiana preferencji może być widziana jako proces analogiczny do zmiany *przekonań*. Zwykle uważa się, iż główną regułą rządzącą zmianą przekonań jest zasada minimalnej zmiany. Kiedy muszę zmienić swoje przekonania, aby zrobić miejsce na nową informację lub – bardziej ogólnie – aby uzgodnić je z pewnymi wymaganiami, które muszę spełnić, powinienem być zachowawczy. Moje nowe przekonania powinny w jak najmniejszym stopniu (jak tylko jest to możliwe przy uwzględnieniu wymagań, które muszą być spełnione) odbiegać od moich przekonań początkowych. Mówiąc inaczej, odległość między moimi starymi a nowymi przekonaniami powinna być zminimalizowana, uwzględniając cel, który mamy osiągnąć. Por. Gärdenfors (1988), s. 8:

> przy ocenie zmiany przekonań wymagamy, aby zmiana była tak *minimalna*, jak jest to potrzebne do uwzględnienia epistemicznego wkładu, który zmianę ową generuje.

Jeśli zasada minimalnej zmiany potraktowana zostanie również jako zasada rządząca rewizją preferencji (co jest poważnym założeniem, które będzie omówione poniżej), to w konsekwencji jednolity stan preferencji, który ma być osiągnięty przez podmiot, powinien możliwie jak najmniej odbiegać od stanu pierwotnego. Parafrazując Gärdenforsa, wymagamy, aby

[10] „Zaakceptować uniwersalną preskrypcję" to to samo, co „sformułować uniwersalną preferencję" (Hare 1989, s. 172). Taka identyfikacja uniwersalnego zalecenia z uniwersalną preferencją została również założona w manewrze ekstrapolacji.

zmiana preferencji była tak minimalna, jak jest to potrzebne do spełnienia koniecznego do zmiany warunku jednolitości. Zatem wartość x powinna być wybrana w taki sposób, aby zminimalizować odległość między dwoma stanami preferencji.

W jaki jednak sposób mielibyśmy określić taką odległość? Jeśli, jak zrobiliśmy, stany preferencji będziemy reprezentować jako wektory, to każdy stan może być widziany jako punkt w przestrzeni wektorowej. Punkt w przestrzeni jest opisywalny poprzez współrzędne numeryczne, które określają jego pozycję w różnych wymiarach przestrzennych. W naszym przykładzie operujemy na wektorach trójczłonowych, tj. punktach w przestrzeni trójwymiarowej. Ogólnie rzecz biorąc, liczba wymiarów określana jest przez liczbę sytuacji, które ja – podmiot – muszę rozważyć, tj. ostatecznie przez liczbę osób zaangażowanych w sytuację rzeczywistą. Dla każdej osoby muszę rozważyć sytuację – rzeczywistą lub hipotetyczną – w której byłbym w skórze tej osoby. Gdyby liczba zaangażowanych osób była mniejsza, powiedzmy, tylko ja i Jan, tylko dwie zamiast trzech sytuacji musiałyby zostać wzięte pod uwagę. Wtedy mój stan preferencji byłby reprezentowany jako punkt w przestrzeni dwuwymiarowej.

Jaką miarę odległości między wektorami powinno się przyjąć? Oczywiste jest, że owa miara powinna być „bezstronna". W szczególności nie powinna ona faworyzować preferencji podmiotu w odniesieniu do rzeczywistej sytuacji s_1 kosztem jego preferencji w odniesieniu do hipotetycznych sytuacji s_2 i s_3. Taka stronniczość byłaby wyraźnie sprzeczna z duchem uniwersalizowalności, który przenika przedsięwzięcie Hare'a. Tak więc przyjmujemy, że w argumencie Hare'a uniwersalizacja przejawia się w dwóch miejscach: po pierwsze, jako warunek jednolitości nałożony na ostateczne stany preferencji – jako wymóg, aby ostateczna preferencja w odniesieniu do każdej sytuacji była taka sama niezależnie od pozycji, jaką się zajmuje w tej sytuacji; po drugie, jako warunek bezstronności nałożony na miarę odległości – jako wymóg, aby odległość między punktami w przestrzeni wektorowej nie ulegała zmianom przy permutacjach wymiarów.

Rozważmy *n-wymiarową* przestrzeń stanów preferencji. Jak już wiemy, n jest liczbą sytuacji, które należy rozważyć, ostatecznie zaś liczbą zaangażowanych osób. *Jeżeli* przyjąć, że miara odległości w tej przestrzeni jest standardową miarą euklidesową, taką, którą przywykliśmy używać w innych kontekstach, to odległość między dwoma stanami preferencji $v = (v_1, ..., v_n)$

i $w = (w_1, ..., w_n)$ jest pierwiastkiem kwadratowym sumy kwadratów różnic między odpowiednimi składowymi v i w:

odległość euklidesowa: $[\Sigma_{i = 1, ..., n}(v_i - w_i)^2]^{½}$

Powyższy wzór czyni nasze zadanie rozwiązywalnym: możemy określić, jaką wartość musi przyjąć x, jeżeli euklidesowa odległość między stanem wyjścia i jednolitym stanem dojścia (x, ..., x) ma być zminimalizowana.

Można udowodnić, że odległość euklidesowa jest zminimalizowana, jeśli x stanowi *średnią* wartości z pierwotnego profilu preferencji[11]. To uśredniające rozwiązanie oczywiście bardzo dobrze oddaje ducha utylitaryzmu preferencji: średnia preferencji nabytych przeze mnie zgodnie z warunkową refleksją, w odniesieniu do sytuacji, w której zajmuję miejsce różnych innych osób, równa się średniej preferencji, które osoby te posiadają w sytuacji rzeczywistej, utylitaryzm preferencji implikuje, że czyn powinno się wykonać wtedy i tylko wtedy, gdy ta ostatnia średnia jest dodatnia[12].

Zatem w naszym przykładzie średnia moich preferencji w stanie

$$(+4, -3, -2)$$

wynosi –1/3. W konsekwencji, jeśli właściwa miara odległości między stanami preferencji jest miarą euklidesową, zrewidowany jednolity stan będzie taki:

$$(-1/3, -1/3, -1/3).$$

Oznacza to, że moralne zalecenie zakazuje jazdy rowerem, dokładnie tak, jak chciałby tego utylitaryzm preferencji.

---

[11] Na dowód patrz Rabinowicz, Strömberg (1996).

[12] Powyższe zachodzi, jeśli problem wyboru jest binarny, tj. gdy jedyne alternatywy to wykonanie lub wstrzymanie się od działania. Przy takim wyborze możemy założyć, że pozytywna preferencja względem działania znajduje odzwierciedlenie w równie silnej negatywnej preferencji względem jego alternatywy. Dyskusja nad wyborem pomiędzy kilkoma alternatywnymi czynami znajduje się w następnym rozdziale. Należy również zauważyć, że nie ma znaczenia, czy wyliczamy średnią, czy całkowitą sumę preferencji, tak długo jak tylko omawiamy sytuacje, w których uczestniczy niezmienna liczba osób. W niniejszym artykule rozważamy jedynie przypadki, w których zbiór osób, które należy wziąć pod uwagę, dany jest *ex ante*.

3. Pytania

Propozycja powyższa zawiera oczywiście wiele kontrowersyjnych elementów. Oto niektóre z pytań, na które należy udzielić odpowiedzi:

(i) Pytania dotyczące *zaleceń*: Czy zalecanie to naprawdę to samo co preferowanie? Wydaje się, iż zakłada to dość uproszczony obraz naszego życia psychicznego. Według niektórych filozofów, jak np. Michaela Bratmana (patrz Bratman 1987), należy starannie rozróżniać takie zjawiska psychiczne, jak pragnienia i zamiary. Wydaje się, że można równie dobrze argumentować, iż preferencje i akceptacje zaleceń są odrębnymi stanami psychicznymi. Rozwiązanie takie stworzyłoby problemy dla powyższej propozycji, która, za Hare'em, zalecenia moralne uznaje za uniwersalne preferencje.

(ii) Pytania dotyczące *minimalnych zmian*: Czy analogia pomiędzy rewizją przekonań a rewizją preferencji jest uzasadniona? W ostatnim przypadku zasada minimalnych zmian nie wydaje się równie wiarygodna jak w pierwszym. W przypadku przekonań, konserwatyzm przy dołączaniu przekonań jest ugruntowany w wymogu poznawczej odpowiedzialności, zaś konserwatyzm w ich porzucaniu swe źródła czerpie w następującej okoliczności: z perspektywy *ex ante*, porzucenie przekonania stanowi poznawczą stratę, gdyż porzucamy sąd, który wówczas uznajemy za *prawdziwy*. W przypadku zmiany preferencji korespondujące uzasadnienie niechęci do wyrzekania się preferencji nie jest dostępne, chyba że zinterpretujemy preferencje w duchu kognitywizmu jako przekonania, że pewne przedmioty (= przedmioty preferencji) są wartościowe. Jednak kognitywna teoria preferencji jest wysoce wątpliwa. Być może alternatywne uzasadnienie konserwatyzmu można byłoby znaleźć w zasadzie psychicznej oszczędności. Zmiany preferencji nie są łatwe, a większe zmiany mogą być trudniejsze do osiągnięcia niż mniejsze.

Istnieje jeszcze inny możliwy powód konserwatyzmu w zmianach preferencji, który szczególnie odpowiada teorii namysłu moralnego Hare'a. Nie pasuje on jednak zbytnio do rozważanej obecnie propozycji. Jak pamiętamy, preferencje w stanie wejściowym są przyjmowane przez podmiot zgodnie z zasadą warunkowej refleksji. Jeśli jednak zasada ta ma być prawdziwa na mocy samych pojęć, jak nalega Hare, niezrozumiałe jest, jak preferencje takie mogą kiedykolwiek ulec zmianie[13]. W rezultacie, przejście do nowego

[13] Za zarzut ten wdzięczny jestem Markowi Schroederowi.

stanu preferencji, w którym preferencje są jednolite, staje się trudne do uzasadnienia. Trudności tej można uniknąć, jeśli przeinterpretujemy warunkową refleksję z wymogu pojęciowego na normatywny – jeśli potraktujemy ją jako wymóg racjonalności, który w pewnych okolicznościach możemy naruszać w celu spełnienia innych ograniczeń, takich jak uniwersalizowalność. Inną alternatywą byłoby traktowanie preferencji w stanie wejściowym jako w jakiś sposób nadal obecnych, nawet po przejściu do stanu wyjściowego. Bardzo niejasne jest jednak, co miałoby to znaczyć.

(iii) Pytania na temat wyboru pomiędzy *kilkoma alternatywnymi działaniami*: Problem dokonania wyboru często nie jest prostym dylematem dotyczącym działania lub powstrzymania się od niego. Zamiast tego zadanie podmiotu polega na dokonaniu wyboru ze zbioru kilku alternatywnych czynów. Preferencje podmiotu dotyczące różnych sytuacji są wtedy reprezentowane przez *macierz*, a nie przez pojedynczy wektor: Jeśli należy rozważyć *n* sytuacji oraz *m* alternatywnych działań, stan preferencji może być przedstawiony jako macierz *n*-kolumnowa (jedna kolumna dla każdej sytuacji) i *m*-wierszowa (po jednym wierszu dla każdego z działań). Wartości numeryczne w różnych komórkach macierzy określają intensywność preferencji. Tak więc wartość wiersza *j* w kolumnie *i* określa siłę preferencji podmiotu w odniesieniu do działania *j*, jeśli idzie o sytuację *i*. Możemy przyjąć, że siła preferencji jest mierzona na skali przedziałowej. Wartości w macierzy będą więc dopuszczały pozytywne transformacje liniowe. (Nie ma tu już potrzeby, aby ustalać, czy podmiot jest za, czy przeciw działaniu. Wystarczy określić, czy preferuje on dane działanie w porównaniu z alternatywami, czy też je dyspreferuje, i z jaką intensywnością. W związku z tym zero w naszej skali może być wybrane dowolnie.)

Aby dzięki uniwersalizowalności podmiot obrał moralny punkt widzenia, musi on przejść od swego wcześniejszego stanu preferencji do nowego, reprezentowanego przez macierz z *jednolitymi wierszami.* Znaczy to, że dla każdego działania *j*, wiersz odnoszący się w nowej macierzy do *j* musi mieć takie same wartości preferencji dla każdej sytuacji: (x_j, x_j, ..., x_j). Ponadto, zgodnie z zasadą minimalnej zmiany, nowa macierz powinna w jak najmniejszym stopniu odbiegać od pierwotnej. Istnieje naturalny sposób, aby zgeneralizować euklidesową miarę do odległości pomiędzy macierzami mającymi *n* kolumn i *m* wierszy: odległość między dwoma macierzami to pierwiastek kwadratowy sumy kwadratów różnic między wartościami w odpowiednich komórkach porównywanych macierzy. Można

wykazać, że odległość ta jest zminimalizowana wtedy i tylko wtedy, gdy wartość preferencji dla każdego działania *j* w wyjściowej macierzy jest średnią wartości preferencji w wierszu *j* macierzy wejściowej[14]. Oznacza to, że działanie, które powinno być wykonane, maksymalizuje średni poziom spełnienia preferencji osób zaangażowanych w daną sytuację, tak jak zalecałby utylitarysta preferencji.

Jednakże, dla uproszczenia, poniżej wrócę do binarnych problemów wyboru, gdzie istnieje tylko jedno działanie, o które podmiot musi się troszczyć. Tak więc zamiast odległości pomiędzy macierzami, będziemy rozważali jedynie odległości między wektorami.

(iv) Pytania o *miarę odległości*: Dlaczego przyjmować, że właściwa miara odległości musi być euklidesowa? Jest to oczywiście tylko jedna z wielu możliwości. Jakie są kryteria adekwatności dla „rozsądnej" miary odległości pomiędzy stanami preferencji? Wymieniliśmy jedno z takich kryteriów: bezstronność. Innym rozsądnym kryterium jest to, iż odległość między dwoma stanami preferencji *v* i *w* powinna być rosnącą funkcją bezwzględnych różnic między odpowiednimi składowymi preferencji w stanach *v* i *w.* Jednakże powyższe wymagania same z siebie nie zaprowadzą nas zbyt daleko.

Najprostszą miarą odległości, której w tym kontekście można użyć, jest tzw. metryka miejska (Manhattan, taksówki). Jest ona wyznaczona przez łączną sumę bezwzględnych różnic pomiędzy wektorami *v* i *w* dla każdego z *n* wymiarów[15]:

odległość miejska: $\Sigma_{i = 1, \ldots, n} \mid v_i - w_i \mid$

Jednakże taka metryka nie zawsze dostarcza unikalnego rozwiązania dla minimalizacji odległości. Faktycznie w przestrzeni dwuwymiarowej rozwiązanie uśredniające jest tylko jednym z nieskończenie wielu, które są możliwe. Jeśli oryginalny wektor ma postać (v_1, v_2), to każde *x* między v_1 i v_2 spełnia nasz warunek. Zatem, na przykład, jeśli uprzedni stan preferencji to (+3, –2), rozwiązanie uśredniające wynosić będzie (+½, +½). Jednak

[14] Na dowód patrz Rabinowicz, Strömberg (1996).

[15] Nazwa pochodzi stąd, iż patrząc z perspektywy euklidesowej, metryka ta podaje odległość między dwoma punktami jako długość najkrótszej trasy z jednego punktu do drugiego, która w każdym swym punkcie przebiega równolegle do jednej z osi. Wygląda to tak, jakbyśmy zmuszeni byli podróżować między punktami wzdłuż ulic miasta, które tworzą regularny wzór kraty.

jednolite wektory, które w metryce miejskiej minimalizują odległość od wektora (+3, –2), stanowią kontinuum rozciągające się od (+3, +3) do (–2, –2). Jeśli liczba wymiarów jest większa niż dwa, wykorzystując metrykę miejską można czasami uzyskać unikalne rozwiązanie dla problemu minimalizacji, *ale* nie ma gwarancji, że rozwiązanie to będzie uśredniające. Oto przykład w trzech wymiarach: załóżmy, że pierwotny wektor to (+6, 0, –3). Jednolity wektor, który minimalizuje odległość miejską od (+6, 0, –3), to (0, 0, 0), podczas gdy rozwiązaniem uśredniającym byłby wynik (+1, +1, +1). (Proszę zauważyć, że (0, 0, 0) nadal byłby minimalizującym rozwiązaniem w metryce miejskiej, nawet jeśli zastąpilibyśmy pierwszą składową w (+6, 0, –3) przez jakąkolwiek wartość wyższą niż 6.) Wniosek jest taki, że jeśli optowalibyśmy za metryką miejską jako naszą miarą odległości, to argument za utylitaryzmem preferencji nie byłby poprawny. Dlaczego więc używać metryki euklidesowej?

Powyższe dwie metryki, miejska i euklidesowa, są członkami dużej rodziny metryk odległości, które mają postać:

odległość Minkowskiego: $[\Sigma_{i=1,\ldots,n} \left| v_i - w_i \right|^k]^{1/k} \quad (k \geq 1)$

Jeśli współczynnik k równy jest 1, otrzymujemy odległość miejską; jeżeli wynosi on 2, uzyskujemy odległość euklidesową itd. Im wyższe jest k, tym większa waga przywiązywana jest do większych bezwzględnych różnic pomiędzy odpowiednimi składowymi wektorów w porównaniu z mniejszymi różnicami. Tylko wtedy, gdy k jest równe 1, jak w metryce miejskiej, wszystkie różnice pomiędzy składowymi są ważone jednakowo, niezależnie od ich wielkości. Ale już dla $k = 2$, jak w metryce euklidesowej, większe różnice bezwzględne mają nieproporcjonalnie większy wpływ, przez potęgowanie, w porównaniu z różnicami mniejszymi[16].

Nadawanie większej wagi większym różnicom między odpowiednimi składowymi stanów preferencji jest bardzo zbliżone do kierowania się *sprawiedliwością*: tym samym nie sprzyja się późniejszym stanom preferencji, które w wielu ze swoich składowych bardzo nieznacznie odbiegają od odpowiednich preferencji we wcześniejszym stanie, lecz w nielicznych przypadkach odbiegają od nich w sposób istotny. Oznacza to, że nie sprzyja

[16] Przy granicy wszystkie odważniki są umieszczone na największych różnicach. Bardzo prostą miarę odległości, która jest wrażliwa tylko na największe różnice, można zdefiniować następująco: odległość między v i $w = max\{|v_i - w_i|: i = 1, \ldots, n\}$.

się późniejszym stanom preferencji, które wykazują niewielkie odchylenia od preferencji wielu zaangażowanych osób, ale znacznie odbiegają od preferencji kilku z nich. Innymi słowy, nie sprzyja się poświęceniu preferencji kilku osób, aby wiele innych uzyskało korzyść. Kwestia ta jest zagadkowa. Przecież powszechnie wiadomo, że względy sprawiedliwości (*fairness*) obce są perspektywie utylitarystycznej. Dla utylitarysty preferencji istotne jest jedyne to, czy zmaksymalizowany jest ogólny poziom spełnienia preferencji (średnia lub suma całkowita; nie ma znaczenia, która z nich, jeśli tylko populacja nie ulega zmianie). Nie jest istotne, czy cel ten jest zrealizowany poprzez poświęcenie preferencji jakiejś osoby dla korzyści innych: nie liczy się osiągnięcie sprawiedliwego podziału spełnienia preferencji. W jaki sposób można więc wytłumaczyć, że aby uzyskać uśredniające utylitarystyczne rozwiązanie, należy posłużyć się metryką euklidesową, a nie miejską, jeżeli to pierwsza, a nie druga z nich bierze pod uwagę sprawiedliwość? Chciałbym znać odpowiedź na to zagadkowe pytanie![17]

(v) Pytania o *egzegezę tekstów Hare'a*: Na ile powyższa propozycja jest wierna sformułowaniu tego argumentu przez samego Hare'a? Pomijając oczywisty fakt, że Hare nigdy nie rozważał naszego problemu w terminach minimalizacji odległości między stanami preferencji, istnieje poważna różnica między jego podejściem a moim: implementują one wymóg uniwersalizowalności na różne sposoby. Ekstrapolacja preferencji, która w argumencie Hare'a służy jako narzędzie uniwersalizacji, w rozważanej przez nas propozycji nie odgrywa żadnej roli. Zamiast tego jest ona zastąpiona przez rewizję preferencji, w której uniwersalizowalność jest implementowana na dwa sposoby: jako wymóg jednolitości wyniku rewizji i jako wymóg bezstronności przy pomiarze odległości. Oznacza to również, że nie został zachowany pomysł Hare'a, aby do moralnych zaleceń dochodzić poprzez przekształcenie interpersonalnych konfliktów preferencji w konflikty wewnętrzne (intrapersonalne).

4. Jednoczesne przeniesienie preferencji

Czy można zrekonstruować argument w taki sposób, aby był on bliższy oryginałowi? Ponownie wróćmy do punktu, w którym wprowadziłem zbiór

[17] Za zwrócenie na to uwagi wdzięczny jestem Christianowi Listowi.

preferencji dotyczących danego działania, o różnej sile i znakach: jedną w odniesieniu do sytuacji rzeczywistej i pozostałe w odniesieniu do sytuacji hipotetycznych, w których role zostały odwrócone. Jak pamiętamy, sugestia Hare'a polegała na tym, aby w tym punkcie dojść do jednolitego zalecenia poprzez proces próbnej ekstrapolacji: staram się przenieść moje preferencje dotyczące, powiedzmy, sytuacji rzeczywistej na jej warianty hipotetyczne. Jeśli przeniesiona preferencja jest wystarczająco silna, aby przetrwać wszelkie konflikty, które mogą być w ten sposób stworzone, osiągnąłem swój cel. Jeśli nie jest, to zamiast niej staram się przenieść jedną z moich innych preferencji – jedną z tych, które posiadam w odniesieniu do sytuacji, w której role zostały odwrócone. Jak jednak widzieliśmy, propozycję tę można zastosować jedynie w przypadkach dwustronnych: gdy zaangażowanych jest kilka osób, prowadzi ona do niepożądanych rezultatów.

Alternatywą byłoby użycie, jak można ją nazwać, *jednoczesnej ekstrapolacji preferencji.* Propozycja ta pochodzi od Daniela Elsteina[18]. Zilustrujmy, jak wyglądałoby zastosowanie tej procedury do naszego przykładu. Nabyłem preferencje dotyczące rozważanego działania w odniesieniu do każdej z sytuacji $s_1 - s_3$. Formują one wektor o postaci:

$$(+4, -3, -2)$$

Aby spełnić wymóg uniwersalizowalności, przenoszę teraz jednocześnie każdą z preferencji z powyższego profilu do wszystkich trzech sytuacji. Możemy myśleć o tym kroku jako o posunięciu, w którym każda z preferencji, aby stać się zaleceniem moralnym, zostaje zuniwersalizowana. W ten sposób docieram do złożonego stanu preferencji, w którym każda z preferencji w stanie (+4, –3, –2) jest obecnie przyjmowana w odniesieniu do *każdej* sytuacji:

$$(\langle +4, -3, -2\rangle, \langle +4, -3, -2\rangle, \langle +4, -3, -2\rangle)$$

W tym nowym stanie pierwszy element <+4, –3, –2> określa moje preferencje w stosunku do s_1, drugi element, który jest dokładnie taki sam, określa preferencje w stosunku do s_2 itd. Można powiedzieć, że w stanie tym przyjmuję jednocześnie trzy zalecenia, które jednolicie stosują się do wszystkich trzech sytuacji: jedno zalecenie *za* rozważanym działaniem z siłą 4 oraz dwa pozostałe *przeciwko* danemu działaniu o mocy, odpowiednio, 3 i 2.

[18] W prywatnej rozmowie.

Ale jak jest możliwa akceptacja zaleceń, które są wzajemnie sprzeczne? Jak mogę *zarazem* przyjąć, że powinienem pojechać rowerem *i* że nie powinienem tego zrobić? Odpowiedź brzmi, że odpowiednie sądy powinnościowe mają charakter *pro tanto*: Każdy z nich odzwierciedla tylko jeden aspekt sprawy. Innymi słowy, zalecają lub zakazują one czynu *o tyle, o ile* posiada on tę lub inną cechę. Zatem jazda rowerem jest zalecana o tyle, o ile pierwotnie preferuję to działanie w s_1, jest zabroniona o tyle, o ile pierwotnie mam preferencję negatywną wobec tego działania w s_2, kiedy jestem w skórze Jana, a także o tyle, o ile pierwotnie posiadam negatywną preferencję wobec tego działania w s_3, gdzie jestem w skórze Marii.

W odróżnieniu od powinności wszystko-biorących-pod-uwagę, powinności *pro tanto* nie są nadrzędne. Nowość jednoczesnej ekstrapolacji leży właśnie w tym, że wprowadza ona powinności *pro tanto*. Innymi słowy, nowość tej propozycji polega na tym, że wprowadza warunek uniwersalizacji na wcześniejszym etapie niż sam Hare: na etapie, na którym nie zobowiązujemy się jeszcze, nawet na próbę, do nadrzędnego sądu moralnego.

Reszta procesu namysłu wygląda następująco. W stanie:

$$(<+4, -3, -2>, <+4, -3, -2>, <+4, -3, -2>)$$

posiadam wzajemnie sprzeczne preferencje w odniesieniu do każdej sytuacji $s_1 - s_3$. Ów wewnętrzny konflikt preferencji rozwiązywany jest następnie przez proste równoważenie,

$$+4 - 3 - 2 = -1$$

W konsekwencji, kończę z tą samą ostateczną preferencją w odniesieniu do każdej sytuacji:

$$(-1, -1, -1)$$

Moje nadrzędne moralne zalecenie biorące-wszystko-pod-uwagę, które uzyskuję poprzez równoważenie zaleceń moralnych *pro tanto*, głosi, iż nie powinienem jechać do pracy na rowerze. Jest to zgodne z utylitaryzmem preferencji: silniejsza preferencja przegrywa z połączonymi siłami preferencji słabszych.

Powyżej zauważyliśmy, że warunkowa refleksja, jeśli traktujemy ją jako prawdę pojęciową, stawia zmianie preferencji poważne ograniczenia: preferencje w stanie wejściowym są przyjmowane przez podmiot zgodnie z warunkową refleksją. Ale jeśli zasada ta jest prawdziwa na mocy samych

znaczeń, wydaje się niemożliwe, aby kiedykolwiek można było porzucić preferencje wejściowe, dopóki podmiot zachowuje pełną świadomość tego, co by preferował, gdyby role zostały odwrócone. W kroku jednoczesnej ekstrapolacji, który prowadzi podmiot z (+4, –3, –2) do (<+4, –3, –2>, <+4, –3, –2>, <+4, –3, –2>), żadna z preferencji nie zostaje jeszcze porzucona. W tym kroku zostają one rozciągnięte, lecz nie porzucone. Jednakże wydają się one znikać w ostatecznym kroku równoważenia, gdy przechodzimy od (<+4, –3, –2>, <+4, –3, –2>, <+4, –3, –2 >) do końcowego stanu preferencji (–1, –1, –1). Jak można rozwiązać ten problem?

Odpowiedź wymaga, jak sądzę, właściwej interpretacji procesu równoważenia. Błądzimy, jeśli postrzegamy go jako *refleksyjne* formowanie nowej preferencji „biorącej wszystko pod uwagę", do której dochodzimy przez wzgląd na wcześniej nabyte preferencje. Zamiast tego, ów proces należałoby traktować bardziej dosłownie, w ten sam sposób co dodawanie ciężarów: pierwotne preferencje są do siebie dodawane, podobnie jak dodaje się odważniki na szali. Oznacza to, że w końcowym wyniku pierwotne preferencje nie są porzucone, lecz nadal są obecne w preferencji końcowej. Nadal tam są obecne jako jej różne składowe[19].

Powyższa rekonstrukcja zachowuje Hare'a koncepcję idealnego namysłu moralnego jako procesu, w którym

> przypadki relacji wielostronnych [...] bez względu na złożoność konfliktu i liczbę osób, sprowadzają się, przy pełnej wiedzy na temat preferencji innych, do konfliktów intrapersonalnych (Hare 2001: 140).

Jednakże koncepcja jednoczesnej ekstrapolacji odbiega od teorii moralnej Hare'a w jednym kluczowym punkcie: w odniesieniu do kwestii nadrzędności. Oczywiście Hare uznaje możliwość istnienia sądów moralnych dających

[19] Za zwrócenie na to uwagi wdzięczny jestem Markowi Schroederowi i Danielowi Elsteinowi. To, iż Hare interpretuje równoważenie w ten zasadniczo „nierefleksyjny" sposób, zasugerowałem w Rabinowicz (1989). Skontrastowałem tam to, co nazywam poglądem na preferencje jako „dane", zgodnie z którym „podmiot traktuje swoje preferencje [...] jako *dane*, jako coś, czemu może nadać pozytywną wagę lub co może zdyskontować, gdy podejmuje decyzję" (tamże, s. 146), z poglądem na preferencje jako „napędowe siły", który traktuje preferencje podmiotu w momencie podejmowania decyzji jako łączne bezpośrednie determinanty wyboru. Zakładam, że bardziej wierna Hare'owi jest interpretacja druga, która nie wymaga ingerencji elementu refleksji.

się uchylać, ale mają one, jego zdaniem, zawsze charakter *prima facie*[20]. Na pierwszy rzut oka obowiązują, jednak w szczególnych przypadkach po dalszym namyśle mogą okazać się nieważne. Tylko w tym sensie mogą być one uchylone: mogą być uznane za niepoprawne. Takie uchylalne sądy ugruntowane są na ogólnych „zasadach *prima facie*", które z reguły obowiązują, choć dopuszczają wyjątki[21]. O ile wiem, Hare nigdy nie rozważał możliwości wprowadzenia sądów moralnych *pro tanto*, które zachowują swoją wagę i ważność nawet w sytuacjach, w których są przeważone przez inne względy moralne[22]. Idea równoczesnej ekstrapolacji wymaga zatem, abyśmy w tym punkcie wyszli poza myśl samego Hare'a.

Podczas gdy zgoda na powinności *pro tanto* może nie być problemem, jest mniej jasne, co w podejściu tym uzasadnia krok od posiadanej przeze mnie preferencji dotyczącej danej sytuacji do jej ekstrapolacji, tj. do odpowiedniego moralnego zalecenia *pro tanto*. Odpowiedź nie może polegać jedynie na tym, iż staram się dojść do sądu moralnego, który musi być uniwersalny. To oczywiście prawda, że szukam ostatecznej uniwersalnej preskrypcji, ale na jakiej podstawie ekstrapoluję każdą preferencję w moim stanie wejściowym, tj. formułuję uniwersalne zalecenia *pro tanto*? W odpowiedzi na tę wątpliwość Elstein sugeruje, że przeniesione preferencje mogą być traktowane jako sądy dotyczące moralnych *racji*, „ponieważ każdy z nich przyczynia się do finalnej oceny moralnej i ma sens powiedzenie, że preferencja ekstrapolowana z mojej preferencji w s_1 nadal stanowi rację na rzecz jazdy rowerem, nawet jeśli finalny sąd nie jest z nią zgodny" (prywatna korespondencja). Jeśli ekstrapolowane preferencje powinny być traktowane w ten sposób, to argument za ekstrapolacją w konsekwencji głosi, że moralne racje, tak jak moralne sądy, są uniwersalizowalne: racja, która ma zastosowanie do jednej sytuacji, musi również stosować się do każdej sytuacji, która jest bardzo podobna, z wyjątkiem faktu, że role osób zostały odwrócone. Zatem jeśli posiadane przeze mnie preferencje w odniesieniu do s_1, s_2 lub s_3 mają funkcjonować jako moralne względy za lub przeciw uniwersalnej zasadzie

[20] Chyba że są one, jak nazywa je Hare, „cudzysłowowymi" sądami moralnymi, które „implikują jedynie, że dane działania są wymagane w celu dostosowania się do norm moralnych obecnych w społeczeństwie" (Hare 2001, s. 77). „Cudzysłowowe" sądy moralne nie są autentycznie moralne, gdyż brak im mocy preskryptywnej.

[21] Por. Hare (2001), s. 78 i n.

[22] Rozróżnienie między racjami *pro tanto* i *prima facie* przedstawione zostało w Kagan (1989), s. 17.

moralnej, która określa, co w tych wszystkich sytuacjach powinno zostać zrobione, każda z nich sama musi być zuniwersalizowana, tj. ekstrapolowana. Jak ujął to Elstein:

> argument wiąże się z założeniem, że racje są uniwersalizowalne, czego (o ile dobrze pamiętam) Hare nie omawia. Jest ono jednak całkiem naturalnym uzupełnieniem poglądu o uniwersalizowalnych sądach moralnych, a nawet mogłoby być płynącym z niego wnioskiem. [...] Myślę więc, w skrócie, że idea jednoczesnej ekstrapolacji może być umotywowana na bliski Hare'owi sposób poprzez namysł nad racjami (Elstein, prywatna korespondencja).

5. Zwrot vendlerowski

Niech mi teraz będzie wolno zwrócić się w stronę komentarzy Zeno Vendlera dotyczących argumentu Hare'a. Wydaje się, że jeśli Vendler ma rację, byliśmy dotąd na fałszywym tropie. Jeśli ma rację, problem braku konfliktu jest jedynie pozorny. Powinien być on nie tyle rozwiązany, co usunięty.

Jak pamiętamy, problem pojawia się dlatego, że eksperymenty myślowe użyte w argumencie Hare'a dotyczą sytuacji czysto *hipotetycznych*. Gdy pytam siebie, jak to byłoby być w czyjejś skórze oraz jaką, dotyczącą owej sytuacji, preferencję teraz posiadam, powinienem rozważyć hipotetyczny stan rzeczy, który różni się od rzeczywistego stanu rolą, jaką w nim zajmuję. Moje preferencje odnośnie do tego, co w takiej hipotetycznej sytuacji należy zrobić, na pierwszy rzut oka nie wydają się sprzeczne z moimi preferencjami dotyczącymi sytuacji rzeczywistej. Na tym polega sedno problemu braku konfliktu.

Obraz ten uległby zmianie, jeśli – jak można argumentować – wyobrażona sytuacja nie różni się w gruncie rzeczy od sytuacji rzeczywistej. Załóżmy, że to, co rozważam w eksperymencie myślowym, to nadal sytuacja *rzeczywista*, ale spostrzegana z *innej perspektywy* – z perspektywy innej osoby. W takim przypadku uformowana przeze mnie preferencja, dotycząca tego, co należy zrobić w takiej sytuacji, jest w konflikcie z preferencją, którą posiadam, gdy spoglądam na tę samą sytuację z mojego własnego punktu widzenia. Wewnętrzny konflikt preferencji, będący tu rezultatem, może zostać następnie rozwiązany w standardowy sposób – poprzez równoważenie.

Zalecone ma zatem być działanie, które w najwyższym stopniu spełnia moje sprzeczne preferencje.

Vendler podkreśla, iż wyobrażanie sobie, że jest się dokładnie w takiej samej sytuacji jak ktoś inny w sytuacji rzeczywistej, nie przenosi nas do innego świata możliwego:

> Jeśli wyobrażam sobie, że jestem tobą, nie wyobrażam sobie „przenoszenia" czegoś do twojego ciała czy „mieszania" dwóch podmiotów. To, co czynię, to przyjmuję, o ile jest to w mocy mojej wyobraźni, spójny zbiór doświadczeń korespondujący z twoją sytuacją (jak można powiedzieć, twoją „humowską jaźń"). Jednak, jak zauważył Hume, w tym zbiorze doświadczeń nie ma szczególnego doświadczenia „ja". Nie istnieje ono również w moim zbiorze. „Ja" jako takie nie ma treści: jest pustą ramą świadomości obojętną na treść. W konsekwencji, budując w swojej wyobraźni obraz tego, czego doświadczałbym w twojej sytuacji, *ipso facto* reprezentuję twoje doświadczenia (Vendler 1988: 176).

> [...] wyobrażając sobie, że jestem tobą bądź Castro, nie dotykam świata: zmieniam jedynie perspektywę patrzenia na świat. To jest powód, mówiąc przy okazji, dla którego utrzymuję przez cały ten artykuł, że wyobrażanie sobie bycia w jakościowo dokładnie takich samych warunkach jak inna osoba jest tym samym, co wyobrażanie sobie, że się jest tą osobą (Vendler 1988: 182).

> [...] *wydaje* się jedynie, że wyobrażamy sobie dwie różne sytuacje [...] Mówi o tym Hare: „Zauważmy, że choć dwie sytuacje są różne, różnią się one jedynie *osobami* zajmującymi obie role, ich własności *uniwersalne* są takie same" (Hare 2001: 141). „Nie", odpowiadam, to jest ta sama sytuacja z tymi samymi osobami; jedyną różnicą jest to, która z nich jest mną: wyobrażając sobie, że jestem nim, wyobrażam sobie tę samą sytuację z innej perspektywy (Vendler 1988: 178).

Stanowisko, które zajmuje Vendler, jest bardzo atrakcyjne. Poprawne wydaje się stwierdzenie, że, obiektywnie rzecz biorąc, eksperymenty myślowe Hare'a nie dotyczą nowych sytuacji. Zamiast tego osiągają one jedynie zmianę subiektywnej perspektywy. To, co „zmienia pozycję" w takim eksperymencie, to nie ja, osoba, którą jestem, lecz tylko „transcendentalne ja", aby użyć kantowskiej terminologii Vendlera – jedynie rama świadomości, która w zasadzie może być wypełniona dowolną treścią. Pamiętajmy, że kiedy wyobrażam sobie bycie takim, jak Jan jest teraz, nie tylko powinienem

przejąć jego zewnętrzne okoliczności, ale także jego cechy psychologiczne: jego przekonania, uczucia, pragnienia itd.

Z subiektywnej perspektywy sytuacja ulega zmianie, gdy jest postrzegana z różnych perspektyw. Lecz obiektywnie jest to nadal ta sama sytuacja. Dlatego kiedy formuję preferencje odzwierciedlające te preferencje, które ja („transcendentalne ja") posiadałbym zajmując różne pozycje, wszystkie owe preferencje dotyczą jednej i tej samej obiektywnej sytuacji. Jako takie, mogą one wchodzić z sobą w konflikt – problem braku konfliktu jest pozorny.

Vendler uważa, że ten „antymetafizyczny" manewr upraszcza argument Hare'a. Wydaje się jednak, że wniosek może być przeciwny. Chociaż problem braku konfliktu znika, zamiast niego mamy nowy i poważniejszy kłopot[23]. Nadal muszę uformować preferencje odzwierciedlające te, które miałbym „ja", gdybym był w różnych pozycjach; muszę posiadać wszystkie te preferencje razem, z jednej i tej samej perspektywy, aby móc je ze sobą zrównoważyć. W tym celu polegać muszę na takiej zasadzie jak warunkowa refleksja. Jednakże, i tu tkwi haczyk, zasada ta nie wydaje się być stosowalna w kontekstach takich jak ten. Dlaczego nie? Cóż, warunkowa refleksja jest wyrazem *troski własnej* – fundamentalnej postawy troski o siebie, którą każda właściwie zintegrowana osoba powinna przejawiać. Troska własna stosuje się nie tylko do sytuacji faktycznej, rozciąga się ona również na sytuacje hipotetyczne, w których dana osoba może się znaleźć. Przejawia się w popieraniu preferencji, które dana osoba miałaby w sytuacji hipotetycznej, dokładnie tak, jak to stypuluje warunkowa refleksja.

Jednak można by zapytać, o kogo się troszczę, gdy troszczę się o *siebie*? O „transcendentalne ja" – ramę świadomości, która może być wypełniona dowolną treścią – czy też raczej o określoną *osobę*, konkretną osobę, którą jestem? Jeśli o tę ostatnią, co wydaje się dość oczywiste (co może mnie obchodzić sama rama świadomości?), wtedy troska własna nie gra żadnej roli w radykalnych eksperymentach myślowych, które proponuje nam rozważać Hare. W eksperymentach tych to, co sobie naprawdę wyobrażam, to bycie kimś innym, inną osobą. Nie chodzi więc o wyobrażenie sobie siebie – osoby, którą się jest – w jakichś hipotetycznych okolicznościach. Lecz w takim razie, jeśli troska własna nie obejmuje transcendentalnych „zmian perspektywy", to takie zmiany pozostają poza domeną stosowania warunkowej refleksji. Oznacza to, że argument Hare'a nie unika luki, wbrew temu, co

[23] Por. rozdział 7 w Rabinowicz, Strömberg (1996).

mógłby myśleć Vendler. Preferencje, które należą do różnych subiektywnych perspektyw, obiektywnie dotyczą tej samej sytuacji, ale jeśli warunkowa refleksja nie jest stosowalna, nie muszą one zostać odzwierciedlone w jednej perspektywie: nie muszą prowadzić do zbioru współistniejących preferencji, które są przyjmowane razem, w jednym stanie preferencji. W konsekwencji, nie muszą prowadzić do konfliktu wewnętrznego, który mógłby być rozwiązany przez równoważenie.

Recenzent tego tomu, Peter Dietsch, w swoich uwagach zakwestionował moje powyższe rozumowanie:

> Nie rozumiem, dlaczego [...] zmodyfikowana zasada warunkowej refleksji [czyli taka, która mogłaby być zastosowana do Hare'a radykalnych eksperymentów myślowych, jeśli słuszność miał Vendler w swym antymetafizycznym manewrze] nie jest ważna. Rozważmy następującą kandydatkę, dostosowaną do terminologii argumentów Vendlera: „O ile w pełni wiem, co w rzeczywistej sytuacji preferowałbym z perspektywy kogoś innego, muszę posiadać odpowiednią preferencję (ten sam znak, taka sama siła)". Nie jest dla mnie jasne, w jakim sensie zasada warunkowej refleksji jest powiązana z pojęciem troski własnej [...], iż uniemożliwiałoby to takie jej zastosowanie.

Powyższa obawa jest zasadna, ale wydaje się, że zmodyfikowana zasada warunkowej refleksji jest zbyt silna, by mogła być zaakceptowana przez kogoś takiego jak Hare – bez względu na to, czy jest ona rozumiana jako prawo pojęciowe (jak warunkową refleksję traktuje Hare), czy jako wymóg racjonalności. Zmodyfikowana zasada w swej istocie mówi, że *empatia pociąga za sobą sympatię*: jeśli w pełni zrozumiem, czego ktoś w rzeczywistej sytuacji pragnie, muszę tym samym stać się posiadaczem odpowiedniego pragnienia. Pogląd ten jest z pewnością obcy Hare'owi. Twierdziłby on, że jak długo nie wydajemy sądów moralnych, tak długo nie jesteśmy zmuszeni, czy to w oparciu o racjonalne, czy też pojęciowe podstawy, do sympatyzowania z naszymi bliźnimi. Sama empatia nie wystarczy. Nie ma sprzeczności w idei racjonalnego amoralisty, który jest empatyczny, gdy mu to odpowiada, lecz nie sympatyzuje ze swoimi bliźnimi.

Podsumowując: nie jestem pewien, czy Vendler w swoim „antymetafizycznym" manewrze ma słuszność. W każdym razie, aby w tej sprawie podjąć decyzję, należałoby odbyć daleką wyprawę na obszar modalnej metafizyki. *Jeśli* ma on rację, tzn. jeśli eksperymenty myślowe Hare'a tak

naprawdę nie przenoszą nas do innych możliwych światów, to argument Hare'a nie będzie mógł być użyty, ponieważ warunkowa refleksja nie stosuje się do transcendentalnych zmian perspektywy. Jeśli, z drugiej strony, nie ma racji, lub jeśli polegające na zmianie ról eksperymenty myślowe można zinterpretować w mniej skrajny sposób, niż to czyni Hare[24], to eksperymenty takie byłyby w stanie przenieść podmiot poza rzeczywistą sytuację do innych hipotetycznych sytuacji, w których znajdzie się on na miejscu przedmiotu działania. Wtedy warunkowa refleksja miałaby zastosowanie, ale stajemy wobec problemu braku konfliktu. Jego rozwiązanie wymaga, aby preferencja danej osoby została zuniwersalizowana bądź poprzez metodę rewizji preferencji, bądź przez metodę ich jednoczesnych ekstrapolacji. W pierwszym przypadku końcowy sąd moralny uzyskiwany jest bezpośrednio, podczas gdy w drugim tylko za pośrednictwem moralnych sądów *pro tanto*. Oba rozwiązania odbiegają od podejścia samego Hare'a do procesu moralnego namysłu, lecz rozwiązanie pośrednie wydaje się lepsze. Po pierwsze, jest ono prostsze i znacznie bliższe pierwotnemu sformułowaniu argumentu Hare'a. Nie pozostawia również tylu wątpliwości. Jak widzieliśmy, metoda rewizji preferencji napotyka na poważne problemy. W szczególności, obrona zasady minimalnej zmiany przy rewizji preferencji oraz wybór odpowiedniej miary odległości między stanami preferencji mogą prowadzić do nieprzezwyciężalnych trudności.

Literatura

Bratman M. (1987), *Intentions, Plans, and Practical Reason*, Cambridge, Mass: Harvard University Press.

Gibbard A. (1988), *Hare's Analysis of 'Ought' and its Implications. Hare and Critics: Essays on Moral Thinking*, red. D. Seanor, N. Fotion, Oxford: Clarendon Press, s. 57–72.

Hare R.M. (1963), *Freedom and Reason*, Oxford: Oxford University Press.

[24] Alternatywą byłoby traktowanie sytuacji hipotetycznych, które podmiot musi rozważyć w celu osiągnięcia sądu moralnego, jako podobnych pod każdym *istotnym* względem do sytuacji rzeczywistej, a nie, jak ma to miejsce u Hare'a, pod *każdym* względem. Jednakże takie odwołanie się do istotnych podobieństw prowadzi do innych znowu problemów: decyzja o tym, jakie aspekty rzeczywistej sytuacji są moralnie istotne, wydaje się zakładać to, co miało być dowiedzione, jeśli podjęta jest na *początku* moralnego namysłu.

Hare R.M. (1981), *Moral Thinking: its Level, Method and Point*, Oxford: Clarendon Press.

Hare R.M. (1987), *Why Moral Language? Metaphysics & Morality*, red. P. Pettit, R. Sylvan, J. Norman, Oxford: Basil Blackwell.

Hare R.M. (1989), *Reply to Ingmar Persson*, „Theoria" 55, s. 171–177.

Hare R.M. (2001), *Myślenie moralne. Jego płaszczyzny, metody i istota*, tłum. J. Margański, Warszawa: Aletheia.

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

Harsanyi J.C. (1977), *Morality and the Theory of Rational Behaviour*, „Social Research" 44, s. 623–656. Przedruk [w:] *Utilitarianism and Beyond*, red. A. Sen, B. Williams, Cambridge: Cambridge University Press, s. 39–62.

Kagan S. (1991), *The Limits of Morality*, Oxford: Clarendon Press.

Persson I. (1989), *Universalizability and the Summing of Desires*, „Theoria" 55, s. 159–170.

Rabinowicz W. (1989), *Hare on Prudence*, „Theoria" 55, s. 145–151.

Rabinowicz W., Strömberg B. (1996), *What if I were in his Shoes? On Hare's Argument for Preference Utilitarianism*, „Theoria" 62, s. 95–123.

Schueler G.F. (1984), *Some Reasoning about Preferences*, „Ethics" 95, s. 78–80.

Van Fraassen B. (1984), *Belief and the Will*, „Journal of Philosophy" 81, s. 235–256.

Vendler Z. (1988), *Changing Places? Hare and Critics: Essays on Moral Thinking*, red. D. Seanor, N. Fotion, Oxford: Clarendon Press, s. 171–184.

PREFERENCE UTILITARIANISM BY WAY OF PREFERENCE CHANGE?

Summary

This paper is a translation of my *Preference Utilitarianism by Way of Preference Change?*, in which I revisit Richard Hare's classical and much discussed argument for preference utilitarianism (*Moral Thinking*, 1981), which relies on the conception of moral deliberation as a process of thought experimentation, with concomitant preference change. The paper focuses on an apparent gap in Hare's reasoning, the

so-called No-Conflict Problem. A solution to this difficulty which was proposed in (Rabinowicz, Strömberg, 1996) is re-examined and shown to lead to a number of difficulties, not least in connection with the choice of an appropriate measure of distance between preference states. The paper therefore also considers an alternative idea, due to Daniel Elstein. This new proposal may well turn out to be the best way of filling the gap in Hare's argument.

The paper also examines whether the gap is there to begin with: The problem should perhaps be dissolved rather than solved. This suggestion goes back to an idea of Zeno Vendler (1988). Unfortunately, it turns out that Vendler's move does not save Hare from criticism: It does dissolve the No-Conflict Problem, but at the same time gives rise to another, potentially more serious difficulty.