Krzysztof Saja

Eksperymenty myślowe a utylitaryzm : odpowiedź na artykuł Włodka Rabinowicza

Analiza i Egzystencja 12, 39-60

2010

Artykuł został opracowany do udostępnienia w internecie przez Muzeum Historii Polski w ramach prac podejmowanych na rzecz zapewnienia otwartego, powszechnego i trwałego dostępu do polskiego dorobku naukowego i kulturalnego. Artykuł jest umieszczony w kolekcji cyfrowej bazhum.muzhp.pl, gromadzącej zawartość polskich czasopism humanistycznych i społecznych.

Tekst jest udostępniony do wykorzystania w ramach dozwolonego użytku.

MUZEUM HISTORII POLSKI

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

ARTYKUŁY

KRZYSZTOF SAJA*

EKSPERYMENTY MYŚLOWE A UTYLITARYZM

ODPOWIEDŹ NA ARTYKUŁ WLODKA RABINOWICZA *UTYLITARYZM PREFERENCJI POPRZEZ ZMIANĘ PREFERENCJI?*

Słowa kluczowe: preskryptywizm, utylitaryzm, Hare, Rabinowicz, metaetyka
Keywords: prescriptivism, utilitarianism, Hare, Rabinowicz, metaethics

Wlodek Rabinowicz w artykule *Utylitaryzm preferencji poprzez zmianę preferencji* rozważa zarzut, który teorii myślenia moralnego Richarda Hare'a postawiony został wcześniej przez Georga F. Schuelera (1984) i Ingmara Perssona (1989). Nosi on miano problemu braku konfliktu. W myśl owej krytyki, nawet jeśli założylibyśmy prawdziwość przesłanek, od których wychodzi Hare przy fundowaniu utylitaryzmu jako teorii normatywnego namysłu moralnego, to i tak istnieje luka między jego przesłankami a oczekiwanym wnioskiem. Choćbyśmy założyli, iż sądy moralne są uniwersalnymi, nadrzędnymi preskrypcjami i zgodzili się na pewne kontrowersyjne wnioski, które Hare z nich wyciąga, nadal nie uzyskamy odpowiedzi na

* Dr Krzysztof Saja, ur. 1979, studiował filozofię na Uniwersytecie Szczecińskim, stopień doktora uzyskał w 2006 r. na UMK w Toruniu. Obecnie pracuje jako adiunkt w Instytucie Filozofii US, specjalizując się w metaetyce oraz utylitarystycznych teoriach etycznych. Autor monografii *Język moralności a utylitaryzm. Filozofia moralna Richarda M. Hare'a*.

pytanie, dlaczego powinniśmy być utylitarystami. Celem artykułu jest odparcie omawianego zarzutu.

Aby uzasadnić utylitaryzm preferencji, należy uzasadnić wszystkie jego cechy: są to konsekwencjalizm (przy ocenie czynu istotne są jedynie jego konsekwencje); welferyzm (jedynie stan dobrobytu jednostek jest wartościowy sam przez się, a dobro ogółu jest sumą dóbr jednostek); uniwersalizm (istotne jest dobro ogółu – wszystkich jednostek, których czyn dotyczy); agregacjonizm (dobra należy ze sobą sumować, a dobro ogółu jest pochodną sumy dóbr jednostek); egalitaryzm (dobrobyt każdej jednostki jest tak samo istotny); maksymalizację (czyn jest słuszny, jeśli maksymalizuje daną wartość); preferencjalizm (jedyną nieinstrumentalną wartością jest spełnienie preferencji). Hare, aby ufundować teorię normatywną o powyższych cechach, przeprowadził następujący tok rozumowania. Według niego:

(1) Racjonalny wybór egoistyczny, np. decyzja, czy powinienem wydać 100 zł na kino i kolację czy na książki, ma charakter zgodny z konsekwencjalizmem, preferencjalizmem, agregacjonizmem i zasadą maksymalizacji. W praktyce wyraża się to tym, iż powinienem wybrać ten czyn, którego skutki dostarczą najwyższego spełnienia sumy moich preferencji.

(2) Sądy moralne są nadrzędnymi uniwersalnymi preskrypcjami.

(3) Wybór moralny podlega racjonalności indywidualnej ograniczonej wymogiem uniwersalizowalności i nadrzędności. Choć myślenie moralne nie sprowadza się jedynie do zasad racjonalnego wyboru, są one jego bardzo ważnym elementem. Utylitarystyczna teoria moralna kieruje się namysłem nad dobrem ogółu, który rządzony jest kryteriami racjonalności indywidualnej.

(4) Aby uzyskać sąd moralny – tzn. uniwersalną i nadrzędną preskrypcję – należy przeprowadzić specyficzne eksperymenty myślowe polegające na uniwersalizacji naszych wstępnych preskrypcji. Polegają one na wyobrażeniowym wejściu w skórę innych osób i zadaniu sobie pytania, czy również z ich perspektywy chcielibyśmy, aby uniwersalnie obowiązywało dane zalecenie. W eksperymentach musimy zinternalizować preferencje wszystkich zaangażowanych osób (uniwersalizm), traktując je jako jednakowo ważne (egalitaryzm).

(5) W rezultacie takich eksperymentów międzyosobowe konflikty preferencji sprowadzone zostają do konfliktów wewnętrznych.

W konsekwencji preferencje mogę poddać zsumowaniu (agregacjonizm), a konflikty między nimi rozstrzygnąć tak, jak opisane jest to w (1).

(6) Owo rozstrzygnięcie, jeśli byłoby przeprowadzone przez idealny podmiot – Archanioła – wyznaczyłoby czyn, który jest moralnie słuszny i zgadza się z utylitaryzmem preferencji.

W omawianej krytyce profesor Rabinowicz akceptuje na potrzeby artykułu punkty (1)–(4). Zarzut braku konfliktu, który wysuwa, uderza jedynie w (5). Z tego względu w niniejszym artykule skupię się jedynie na przezwyciężeniu wysuwanych przeszkód, które stoją na drodze akceptacji kwestionowanej tezy.

Konflikt międzyosobowy w konflikt wewnętrzny

Olbrzymią rolę w uzasadnieniu punktu (5), zamieniającego międzyosobowy konflikt preferencji w konflikt wewnętrzny, pełni, oprócz założeń (1)–(4), zasada warunkowego namysłu. W *Myśleniu moralnym* Richard Hare wyraził ją tak:

> 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 (Hare 2001, s. 139, 129).

Powyższe sformułowanie naraża nas jednak na dwie interpretacje, przed myleniem których przestrzegał Hare:

(1) Zgodnie z moją obecną preferencją o natężeniu *N*, gdybym znalazł się w tej sytuacji, wolałbym, aby *x* raczej się zdarzyło, niż nie zdarzyło;

(2) Gdybym znalazł się w tej sytuacji, to zgodnie z moją preferencją o natężeniu *N* wolałbym, aby *x* raczej się zdarzyło, niż nie zdarzyło (Hare 2001, s. 124).

Według Hare'a zdania te nie są tożsame, ale nie mógłbym wiedzieć, że (2), gdyby (1) nie było prawdą (2001, s. 124). Innymi słowy, moje hipotetyczne

preferencje – jeśli wiem, że w sytuacji hipotetycznej bym je posiadał – znajdują odzwierciedlenie w moich rzeczywistych preferencjach dotyczących danego przypadku.

Biorąc powyższe pod uwagę, mój stan preferencji może składać się z trzech rodzajów preferencji:

> *preferencji bazowych* (PB) – preferencji, które aktualnie posiadam i które odnoszą się do mojej obecnej sytuacji;
>
> *preferencji hipotetycznych* (PH) – preferencji, które bym posiadał, jeśli znalazłbym się w hipotetycznej sytuacji;
>
> *preferencji warunkowych* (PW) – preferencji, które posiadam aktualnie za działaniem w hipotetycznej sytuacji.

Do powyższego zbioru dodajmy jeszcze jeden rodzaj preferencji. Jak trafnie w innym miejscu zauważył Rabinowicz, zasada warunkowego namysłu implikuje bowiem, że wiedza o preferencji pociąga posiadanie tej preferencji (WPPP): „Jeśli obecnie (w pełni) wiem, że aktualnie posiadam lub będę posiadał preferencję na korzyść A, to teraz posiadam równie silną preferencję na korzyść A" (Rabinowicz 1989, s. 148–149). Jak dalej zaznacza: „w WPPP moja obecna preferencja nie jest jedynie warunkowa, ponieważ wiem, że sytuacja, w której posiadam preferencję za A, zaistniała lub zaistnieje" (tamże, s. 149). Innymi słowy, moje hipotetyczne preferencje – jeśli wiem, że w sytuacji hipotetycznej bym je posiadał oraz że *sytuacja hipotetyczna zaszła lub zajdzie* – znajdują odzwierciedlenie w moich rzeczywistych preferencjach dotyczących danego przypadku. Choć założenie to należałoby ograniczyć do preferencji racjonalnych (Rabinowicz 1989, s. 149–150), z czym godził się Hare (1989, s. 218–219), oddaje ono jego przekonanie, iż pełna reprezentacja przyszłej preferencji pociąga za sobą jej teraźniejszą kopię (Hare 2001, s. 124, przypis 16). Taką preferencję nazwiemy zaktualizowaną preferencją warunkową (APW):

> *zaktualizowana preferencja warunkowa* (APW) – preferencja, którą posiadam aktualnie za działaniem w hipotetycznej sytuacji, o której wiem, że zaszła lub zajdzie.

Dokonawszy koniecznych rozróżnień, przejdę teraz do wykazania, iż problem braku konfliktu nie dotyczy eksperymentów myślowych Hare'a. Poniżej rozważę ich trzy interpretacje: pierwszą nazwę interpretacją wielo-

światową, drugą wielosytuacyjną interpretacją jednoświatową, trzecią jednoświatową interpretacją Vendlera.

Problem braku konfliktu w wieloświatowej interpretacji eksperymentów myślowych Hare'a

Problem braku konfliktu związany jest z pewnym sposobem rozumienia „wejścia w skórę innych", którą zaproponował Ingmar Persson. Na czym polega więc sedno wieloświatowego odczytania eksperymentów myślowych Hare'a? Profesor Rabinowicz przedstawia następujący przykład, który stanowi podstawę jego rozważań:

> 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. [...] 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 (Rabinowicz 2010).

W myśl pierwszej interpretacji eksperymentów myślowych Hare'a, „stawiając się na miejscu innych osób" w powyższym przykładzie, powinienem wziąć pod uwagę trzy odrębne sytuacje, interpretowane jako odrębne światy możliwe: moją sytuację rzeczywistą (s_1) oraz dwie sytuacje hipotetyczne, w których byłbym w skórze Jana (s_2) i Marii (s_3). Następnie powinienem zastosować do nich zasadę warunkowej refleksji. W wyniku tego procesu mój stan preferencji będzie składał się z:

- preferencji bazowej $a = +4$ odnośnie do rzeczywistej sytuacji s_1,
- preferencji warunkowej b = –3 odnośnie do hipotetycznej sytuacji s_2,
- preferencji warunkowej c = –2 odnośnie do hipotetycznej sytuacji s_3.

W tym miejscu jednak, według krytyków Hare'a, powstaje problem braku konfliktu (Rabinowicz, Strömberg 1996, s. 100–103). Przejawia się on, co sugeruje Persson (1989, s. 163), niemożnością, co chciałby uzyskać Hare, zsumowania preferencji *b* i *c* przeciwko *a* oraz, generalnie, niemożliwością zsumowania (zbalansowania) jakichkolwiek preferencji (nawet dwóch) z różnych sytuacji hipotetycznych (Schueler 1984, s. 80).

Zacytujmy fragment, który wskazuje na ów problem:

> 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 sobie w konflikcie. Inaczej niż w przypadku, w którym w grę wchodzi roztropność, nie istnieje tutaj konflikt preferencji, który muszą 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 (Rabinowicz 2010).

Preferencje warunkowe nie wchodzą w konflikt z preferencją bazową, gdyż spełnienie jednej nie udaremnia możliwości spełnienia drugiej. Dotyczą one bowiem różnych możliwych światów. Mogę sobie wyobrazić, że zdecydowałem się pojechać rowerem i w trakcie jazdy stwierdzić, że, po pierwsze, gdybym był w skórze Jana, nie chciałbym, żeby profesor jechał rowerem, gdyż zależy mi na czasie. Znajduję więc w sobie preferencję hipotetyczną przeciwko jeździe rowerem. Co więcej, imaginując sobie, iż jestem Janem, aktualnie, w czasie jazdy rowerem, zauważam u siebie preferencję (warunkową) przeciwko jeździe rowerem, gdybym miał być w sytuacji Jana. Ponieważ jednak nie jestem w sytuacji Jana, jadąc na rowerze mogę spełniać swoją pierwotną preferencję, posiadając również tę, która jest czysto warunkowa. Preferencje te nie wchodzą z sobą w konflikt, gdyż pierwsza nie

udaremnia drugiej: obie mogą być spełnione: jedna w sytuacji rzeczywistej, druga w sytuacji hipotetycznej.

Czy należy wyciągnąć stąd wniosek, że podczas podejmowania racjonalnego namysłu, który według Hare'a musi kierować moralnymi decyzjami, nie liczą się posiadane przeze mnie preferencje warunkowe? Teza taka wydaje się zbyt pochopna. Rozważmy bowiem pewien przykład. Załóżmy, że nie odpowiada mi wielkość mojego mieszkania i odnajduję w sobie preferencję wzięcia kredytu na zakup dużego domu z ogrodem. Moje pragnienie w sytuacji aktualnej s_1 ma siłę +4. Zastanawiam się jednak i stwierdzam, że jeśli byłbym w sytuacji s_2, w której stałbym się kaleką niemogącym pracować, to wolałbym mniejsze mieszkanie niż duży dom i kredyt do spłacenia. Moja preferencja za domem i kredytem w sytuacji s_2 miałaby siłę –3. Przychodzi mi również na myśl sytuacja s_3, w której członkowie mojej rodziny odchodzą ode mnie i zostaję sam, posiadając dom, ogród i kredyt. W tej sytuacji moja preferencja posiadania domu wynosi –2. Zgodnie z zasadą warunkowej refleksji, jeśli w pełni stawiam się w powyższych sytuacjach, aktualnie uzyskuję warunkową preferencje $wp_2 = -3$ odnośnie do sytuacji s_2 oraz $wp_3 = -2$ odnośnie do sytuacji s_3. Czy dokonując racjonalnego namysłu, powinienem wziąć pod uwagę preferencje p_1, wp_2 i wp_3, równoważąc je ze sobą? A może powinienem rozważać jedynie preferencje bazowe? Przy podejmowaniu podobnego wyboru większość z nas zapewne nie zwróciłaby uwagi na preferencje wp_2 i wp_3. Ważne jest jednak wyjaśnienie: racjonalne osoby nie zważałyby na powyższe preferencje warunkowe nie dlatego, że nie wchodzą one w konflikt, lecz z powodu przypuszczenia, że prawdopodobieństwo wystąpienia sytuacji s_1 oraz s_3 jest bardzo małe (powiedzmy 1–2% szans). W takim wypadku siłę aktualnej preferencji p_1 mnoży się przez 0,97 (ponieważ możliwość jej realizacji jest prawie stuprocentowa), siłę preferencji warunkowej natomiast przez 0,01 i 0,02. Biorąc pod uwagę powyższe dane, preferencje warunkowe w omawianym przykładzie w ostatecznej decyzji praktycznie nie grają roli (3,88 > 0,03 + 0,04).

Zmodyfikujmy jednak nasz przypadek, aby, co będę uzasadniał, spełniał wszystkie wymagania myślenia moralnego według Hare'a. Załóżmy więc, że muszę podjąć decyzję o zakupie domu z ogrodem. Po namyśle, tak jak w powyższym przykładzie, rozpoznaję u siebie trzy preferencje: bazową $s_1 = 4$ oraz dwie warunkowe $s_2 = -3$ i $s_3 = -2$. Załóżmy dalej, że dowiaduję się, iż specjalna instytucja naukowców niezbicie dowiodła, iż historia ma fatalistyczny charakter: to, co się zdarzy, zostało już rozstrzygnięte i ma

konieczny charakter. Mój kolega jest członkiem owej zasłużonej grupy ekspertów. Poprosiłem go o naukowy zapis mojej przyszłości. Okazało się, ku mojemu przerażeniu, iż w perspektywie kilku miesięcy najpierw zostanę kaleką, później zaś stracę rodzinę. Gdy otrząsnąłem się z doznanego szoku, stanęła przede mną decyzja: czy powinienem kupić duży dom z ogrodem. Jakiego wyboru powinienem dokonać, jeśli okazałoby się, że prawdopodobieństwo zajścia sytuacji s_2 i s_3 wynosi 100%? Czy suma warunkowych preferencji powinna przeważyć nad preferencją aktualną ($3{,}88 < 2 + 3$)? Odpowiedź oczywiście brzmi – tak.

Dlaczego powyższy przykład jest dla nas istotny? Zdaniem Hare'a uniwersalizacja i nadrzędność implikują ważne elementy dla namysłu moralnego. Chcąc wydać sąd moralny, który jest preskrypcją, daję wyraz swoim preferencjom. Ponieważ namysł moralny jest racjonalny, kierują nim zasady racjonalnego wyboru. Uniwersalizacja wymaga od nas, abyśmy, wydając zalecenie moralne, wydali je tak, aby mogło być przez nas zaakceptowane uniwersalnie. Znaczy to, iż zaakceptować musimy preskrypcję dla wszystkich istotnie podobnych sytuacji, opisanych bez użycia stałych indywiduowych. W praktyce, dla naszych eksperymentów myślowych, oznacza to jednak, iż powinniśmy abstrahować od:

(a) tego, kim jesteśmy w rzeczywistości i jaką rolę tam pełnimy;
(b) tego, która sytuacja jest rzeczywista, a która hipotetyczna – jakie jest prawdopodobieństwo ich zajścia.

Wymóg wyrażony w punkcie (a) jest szeroko rozpoznany i respektowany przez krytyków Hare'a. Inaczej wygląda, moim zdaniem, stosunek do punktu (b). Jeśli w naszym myśleniu moralnym mamy posługiwać się wyłącznie własnościami uniwersalnymi, nie możemy, zdaniem Hare'a, traktować jako istotnych tego, która z sytuacji $s_1 - s_3$ ma większe prawdopodobieństwo realizacji. Nie możemy wyróżniać żadnej z sytuacji ze względu na naszą realną w niej obecność, gdyż musimy abstrahować od nazw własnych i indeksów określających miejsce, czas, hipotetyczność czy aktualność rozważanych przez nas sytuacji. Uniwersalizacja wymaga więc od nas, abyśmy nie rozróżniali sytuacji rzeczywistych od hipotetycznych: te drugie są równie istotne jak pierwsze (Hare 1989, s. 215; 2001, s. 143–147; 1984, s. 84–85).

W celu zinterpretowania eksperymentów myślowych Hare'a tak, aby sytuacje rzeczywiste i hipotetyczne traktowane były jako równie istotne, postąpić możemy na kilka sposobów. Istnieje możliwość posłużenia się metodą zasłony niewiedzy w stylu Rawlsa czy Harsanyiego, którą zaproponował

Hare'owi Ingmar Persson. Pożądany rezultat moglibyśmy uzyskać wyobrażając sobie, iż, po dokonaniu wyboru za zasłoną niewiedzy, znajdziemy się w którejś z sytuacji $s_1 - s_3$ z 33% prawdopodobieństwem. Ponieważ jednak Hare uważał, iż poprawne sformułowanie racjonalnej decyzji musi być czynione w duchu teorii racjonalności zakładającej pełną wiedzę (*full informational theory of rationality*), nie zgadzał się na interpretowanie myślenia moralnego jako procedury symulowania niewiedzy. Nie powinniśmy traktować prawdopodobieństwa wystąpienia danej sytuacji jako mniejszego niż 1. Z tego też względu Hare, odpowiadając na artykuły Schuelera i Perssona, w których odnaleźć można omawianą krytykę, stwierdził, iż w naszych eksperymentach myślowych sytuacje hipotetyczne powinniśmy traktować tak, jakby *miały stać się rzeczywistymi* (Hare 1984, s. 84–85; 1989; 2001, rozdz. 6.4.). Wymóg uniwersalizowalności oraz pełnowiedzowej teorii racjonalności mógłby więc spełnić eksperyment, w którym wyobrażamy sobie, że jesteśmy w skórze wszystkich osób zaangażowanych w naszą decyzję oraz że wystąpienie każdej z sytuacji $s_1 - s_3$ jest w 100% prawdopodobne. W ten sposób, abstrahując od aktualności czy hipotetyczności sytuacji s_2 i s_3, w konsekwencji traktujemy je jako równie istotne jak sytuację rzeczywistą s_1, gdyż zakładamy pewność, że zostaną urzeczywistnione. Jeśli wyobrażę sobie, że każda z sytuacji $s_1 - s_3$ dotyczy *mnie*, i do każdej z tych sytuacji zastosuję zasadę warunkowej refleksji, uzyskam trzy preferencje p_1, awp_2 i awp_3. Racjonalna osoba powinna traktować je jako równie istotne, a w celu dokonania właściwego wyboru, odpowiednio je zbalansować. W wyniku tego uzyska się zalecenie, które, z jednej strony, będzie uniwersalne (obowiązuje w sytuacji $s_1 - s_3$), z drugiej strony – zgodne z utylitaryzmem preferencji.

Błąd zwolenników problemu braku konfliktu polega więc na tym, iż zakładają oni, że w wyniku eksperymentów myślowych Hare'a nasz stan preferencji będzie się składać z preferencji bazowych i warunkowych, które dotyczą różnych światów możliwych i nie mogą być wspólnie zrealizowane. Hare'a eksperymenty myślowe powinniśmy formułować jednak tak, aby w ich wyniku nasz stan preferencji składał się z preferencji bazowych i *zaktualizowanych* preferencji warunkowych. Jeśli bowiem jest prawdą, iż preferencje warunkowe nie muszą wchodzić z sobą w konflikt, gdy założymy, że korespondujące z nimi hipotetyczne sytuacje nie zajdą, każdy racjonalny podmiot podczas moralnego namysłu powinien uwzględnić zaktualizowane preferencje warunkowe: preferencje, o których wie, że są

lub staną się aktualne. Nie powinniśmy więc traktować sytuacji hipotetycznych jako różnych światów możliwych, to bowiem sugeruje, iż nie mogą się one zarazem zaktualizować: nie mogę wszak zarazem jechać na rowerze i być oczekującym studentem na profesora cyklistę – preferencja bazowa profesora nie zostaje udaremniona przez preferencję warunkową związaną z byciem studentem.

Wielosytuacyjna interpretacja jednoświatowa

Inaczej wygląda jednak nasz eksperyment, gdy sytuacje hipotetyczne i aktualne zinterpretujemy jako należące do jednego świata, w którym występować one będą jako odrębne punkty czasowe (możliwie bliskie), tak, iż każda z sytuacji jest lub zostanie zaktualizowana. Rozważmy więc teraz przypadek, który przedstawia Wlodek Rabinowicz, w interpretacji jednoświatowej. Posiadam aktualną, bazową preferencję $p_1 = 4$ za podróżą rowerem. Chcąc wydać sąd moralny, powinienem postawić się w sytuacji Jana i Marii i wyobrazić sobie, że owe hipotetyczne sytuacje, podobne pod względem istotnych własności uniwersalnych do sytuacji rzeczywistej, staną się w moim życiu realne. Wyobrażam więc sobie, że zaraz po rozpoczęciu realizacji mojej preferencji p_1, gdy tylko wsiadłem na rower, zostanę przeniesiony w czasie i zamienię się rolą najpierw z Janem, a następnie z Marią, którzy czekają na profesora cyklistę. Muszę założyć, że niechybnie ich sytuacje staną się moim udziałem. Jeśli w pełni wejdę w ich skórę, uzyskam zaktualizowaną warunkową preferencję $awp_2 = -2$ oraz $awp_3 = -3$. W trakcie przeprowadzania eksperymentu mój stan preferencji, składający się z preferencji bazowych i zaktualizowanych preferencji warunkowych, wygląda więc: (+4, –2, –3). Wiedząc zatem, że na pewno dotknie mnie działanie, które podjąłam w czasie t_1, gdy będę w t_2 i t_3, racjonalnie jest zaniechać jazdy rowerem. Rezultat taki jest zgodny z utylitaryzmem preferencji[1].

Modyfikacje sposobu opisywania eksperymentu myślowego przedstawionego pierwotnie w *Myśleniu moralnym* nie były Hare'owi obce. Choć w książce tej podążył on drogą, która bliższa jest omawianej w kolejnej

[1] Powyższa interpretacja zakłada oczywiście, iż racjonalny podmiot nie dyskontuje przyszłości oraz że tożsamość osobowa nie jest wyznaczona w pełni przez własności uniwersalne, które są naszym udziałem. Założenia te jednak zgodne są z teorią Hare'a.

części tego artykułu interpretacji Zeno Vendlera, zdawał sobie sprawę z możliwości odmiennego formowania eksperymentów myślowych (Hare 2001, s. 129–130). Interpretację podobną do przedstawionej powyżej zaprezentował w artykule będącym odpowiedzią na krytykę I. Perssona (Hare 1989b, s. 171–177). Hare zaproponował tam, aby wyobrazić sobie, że w pewnym świecie możliwym doświadczymy serii „żywych snów", w których znajdziemy się w skórze wszystkich zaangażowanych w daną decyzję osób. W innej wersji powinniśmy wyobrazić sobie kartezjańskiego demona, który sprawia, iż doświadczymy serii przeżyć, które są ich udziałem. Gdy uznamy, że doznania takie (sny czy halucynacje) z pewnością staną się naszym doświadczeniem, uzyskamy zaktualizowane preferencje warunkowe, które wchodzą w konflikt z preferencją bazową. Dzięki temu eksperymentator będzie musiał traktować sytuacje hipotetyczne tak, jakby miały stać się rzeczywistymi.

Śmiem twierdzić, że w rozwijanym przez Hare'a uzasadnieniu utylitaryzmu to, jakich heurystycznych historii użyjemy, aby pomóc naszej wyobraźni traktować sytuacje hipotetyczne jako aktualne lub które zaktualizują się w przyszłości, nie gra istotnej roli o tyle, o ile spełnia wymagania logiki języka moralnego (przede wszystkim uniwersalizowalności). Narracja stanowiąca o przebiegu eksperymentu myślowego powinna być na tyle skuteczna, aby uzmysłowić nam równą wagę zarówno sytuacji rzeczywistych, jak i hipotetycznych[2]. To, że niektóre procedury eksperymentowania wydają się nam mniej kontrowersyjne, wynika jedynie z naszych przyzwyczajeń. Zalecenie „postaw się na miejscu innych osób" nie budzi naszych wątpliwości w takim stopniu, jak „wyobraź sobie serię żywych snów lub halucynacji", gdyż pierwsze znamy z życia rodzinnego i często stosujemy w celach wychowawczych. Hare w *Myśleniu moralnym* użył tej znanej metafory ze względu na to, że jesteśmy z nią oswojeni (1988a, s. 284).

Wlodek Rabinowicz powyższą modyfikację eksperymentów myślowych Hare'a poddał jednak krytyce (Rabinowicz, Strömberg 1996, s. 103–106). Wysunął on kilka argumentów, choć w tym miejscu rozważę jedynie dwa najbardziej istotne. Pierwszy zarzut głosi, że Hare nie powi-

[2] Należy pamiętać, że eksperymenty myślowe Hare'a mają charakter wyłącznie instrumentalny: służą na drodze racjonalnej decyzji do osiągnięcia uniwersalnej nadrzędnej preskrypcji. Należy je więc w ten sposób formułować, aby jak najskuteczniej mogły one pełnić swoją funkcję.

nien decydować się na powyższą interpretację, bowiem wymaga ona od eksperymentatora symulowania, iż nie posiada on wiedzy co do tożsamości numerycznej osób zaangażowanych w decyzje i, w konsekwencji, nie wie, która sytuacja jest rzeczywista, a która czysto hipotetyczna. Takie założenie dla Hare'a jest jednak kłopotliwe: po pierwsze, jest ono niezgodne z jego przekonaniem, iż moralność zakłada pełnoinformacyjną teorię racjonalności oraz, po drugie, powoduje, iż rezultat osiągnięty przez eksperymentatora nie jest zgodny z jego pierwotnymi ambicjami wykazania, iż myślenie moralne ma praktyczny, motywacyjny charakter. Jeśli bowiem, po pierwsze, nie akceptuje opisanego powyżej rozwiązania Rawlsa-Harsanyiego ze względu na niepotrzebne symulowanie niewiedzy, nie powinien również akceptować zakładania niewiedzy co do aktualności bądź hipotetyczności rozważanych sytuacji. Albo aprobuje się pełnoinformacyjną teorię racjonalności, albo się ją odrzuca. Po drugie, w omawianej interpretacji eksperymentator nie zrealizuje ambicji, które, w mniemaniu Hare'a, czynią jego projekt lepszym od podejścia Rawlsa. Jeśli bowiem zaczyna się od udawania, to na udawaniu się kończy (*pretense in*, *pretense out*), ambicje Hare'a sięgały jednak dalej. Eksperymenty myślowe, które zakładają symulowanie niewiedzy, mogą nie dostarczać podmiotowi praktycznych racji, aby wykonać zalecane działanie, gdy powróci on ze świata wyobraźni do rzeczywistości. Z tego względu Hare nie powinien (jak pierwotnie tego nie czynił) przyjmować pozorowania jako elementu moralnego namysłu. Jak pisze Rabinowicz:

> 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 (Rabinowicz 2010, przypis 7).

Richard Hare, zdaniem Rabinowicza, nie może zakładać zatem równego traktowania sytuacji hipotetycznych z rzeczywistą. Taka interpretacja stanowiska Hare'a nie jest jednak trafna. Co najmniej w kilku miejscach *explicite* zauważa on bowiem, że uniwersalizacja wymaga, abyśmy w eksperymencie myślowym sytuacje hipotetyczne traktowali tak, jakby były one aktualne albo miały się zaktualizować (Hare 1984, s. 84–85; 1989, s. 171–177; 2001, rozdz. 6.4.). Wynika to z jego założeń metaetycznych: uniwersalizacja narzuca konieczność ignorowania stałych indywiduowych, takich jak nazwy osób, miejsc, indeksy aktualności i możliwości światów (Saja 2008, s. 154–163,

182–197). Nie oznacza to jednak, jak sugeruje Rabinowicz, iż w związku z tym powinien on zaakceptować przedstawione przez Perssona rozwiązanie problemu braku konfliktu w duchu Harsanyiego. Dla Hare'a bowiem wymóg ignorowania wiedzy o tożsamości numerycznej osób i sytuacji ma istotne uzasadnienie: wynika z cech języka moralnego. Stanowi on konieczne dla etyki ograniczenie nałożone na racjonalność praktyczną, którą najtrafniej ujmuje teoria pełnoinformacyjna. Rozwiązanie Harsanyiego natomiast zbyt daleko od niej odbiega. Podczas gdy interpretacja Hare'a wymaga ignorowania wyłącznie wiedzy o stałych indywiduowych (hipotetyczności lub aktualności sytuacji), rozwiązanie Harsanyiego zakłada dodatkowo bezstronność (równość) traktowania rozważanych sytuacji oraz niepewność zajścia sytuacji hipotetycznych (33% szans). W konsekwencji, choć myślenie moralne według Hare'a ma charakter bezstronny, bezstronność ta jest *wynikiem* namysłu moralnego, nie zaś *założeniem*, które, budując procedurę eksperymentu, powinniśmy wziąć pod uwagę. Hare nie akceptował metodologii refleksyjnej równowagi jako przejawu nieuzasadnionego intuicjonizmu. Pragnął on bowiem ufundować etykę normatywną z pominięciem założeń etycznych (Hare 1989a, s. 209; Saja 2008, s. 178–197). Podsumowując, interpretacja Rawlsa-Harsanyiego niepotrzebnie zwiększa niewiedzę oraz w punkcie wyjścia przesądza pewne kwestie moralne.

Drugi główny zarzut, który Rabinowicz postawił jednoświatowej interpretacji eksperymentów moralnych, głosi, iż jeżeli sądy moralne mamy formować w oparciu o eksperymenty myślowe w tak dziwacznych światach możliwych, jak postuluje Hare w odpowiedzi Perssonowi (np. światach żywych snów czy halucynacji), aby spełnić wymóg uniwersalności, nasza preskrypcja powinna obowiązywać również w innych dziwacznych światach możliwych $w_1 - w_n$ (np. takich, w których sekwencyjnie będziemy sto razy z rzędu zajmowali pozycję studenta, który czeka na profesora cyklistę) (Hare 1996, s. 106). Jednak

> ilość możliwości, dla których musi on [eksperymentator – K.S.] wydać zalecenie, szybko eksploduje i nie jest wcale jasne, dlaczego jego uniwersalne zalecenie dotyczące wszystkich tych możliwości powinno się pokrywać z tym, które wybrał on dla świata *w* (Hare 1996, s. 106).

Trudność ta generowana jest jednak przez nieporozumienie dotyczące zakresu światów możliwych, które nakazodawca powinien wziąć pod uwagę. Pewne światy możliwe są bowiem relewantne, inne natomiast nie. Ponieważ

eksperyment myślowy jest metodą rozstrzygnięcia konkretnego problemu moralnego ujętego jedynie poprzez własności uniwersalne, nasze zalecenie nie musi być spójne ze wszystkimi światami możliwymi, tak jak konkretna reguła moralna, choć uniwersalna, nie musi być ogólna i dotyczyć każdej sytuacji. Z tego względu Hare rozróżnił ogólność od uniwersalności: preskrypcja może być uniwersalna, choć bardzo szczegółowa – odnosić się jedynie do wąskiej klasy światów możliwych[3]. Świat w_1, w którym sto razy z rzędu przeżywać będę doświadczenia oczekującego studenta, przy założeniu, że tylko raz będę w skórze profesora, nie jest istotnym światem ze względu na rozważany problem, który dotyczy profesora i dwóch studentów. Gdyby jednak studentów czekało stu, możliwy świat w_1 okazałby się istotny.

Jednoświatowa interpretacja Zeno Vendlera

Innym sposobem interpretacji Hare'a eksperymentu myślowego wejścia w skórę innych, które respektuje wymóg traktowania sytuacji hipotetycznych tak jak aktualnych, jest, omawiana przez Rabinowicza, interpretacja Zeno Vendlera. Jest ona rozwinięciem drogi, którą podążył Hare w *Myśleniu moralnym.* Jak słusznie zauważa Rabinowicz, nie powoduje ona problemu braku konfliktu. Związane jest to z faktem, że traktuje ona zarówno sytuacje hipotetyczne, jak i rzeczywiste jako takie, które z pewnością zrealizują się w życiu eksperymentatora. Według Rabinowicza rodzi ona jednak inne, bardziej istotne kłopoty. Przyjrzyjmy się jej więc bliżej.

W sposobie rozumienia eksperymentów myślowych, który rozwinął Zeno Vendler, przypadki $s_1 - s_3$ powinniśmy potraktować jako trzy różne, subiektywne aspekty tej samej obiektywnej sytuacji. Nie zmieniamy więc ról, nie zamieniamy się miejscami, lecz patrzymy z różnych perspektyw na tę samą sytuację (Vendler 1988, s. 181). Powoduje to, iż nasz eksperyment myślowy dotyczy nie różnych światów możliwych, lecz jednego świata aktualnego, nie kilku sytuacji, lecz jedynie sytuacji aktualnej. Dzięki temu podmiot, dokonujący namysłu moralnego, traktuje je wszystkie jako równie istotne.

[3] O różnicach między uniwersalnością a ogólnością oraz relewantnych światach możliwych pisałem w: Saja 2008, s. 154–155, 191–192.

Vendler, aby móc wyjaśnić możliwość radykalnej zmiany perspektywy patrzenia na jedną sytuację, której mamy dokonać, zaproponował rozróżnienie osoby – istoty posiadającej konkretne własności – od „ja" – które ma transcendentalny charakter, może przybierać różne treści, jest pustą ramą świadomości. Jak pisze Rabinowicz: „idąc tym tropem, nie tylko wydaje się znikać nasz pierwotny problem. Wydaje się również, iż nie potrzebujemy już zasady warunkowej troski własnej (*Principle of Conditional Self-Endorsement!*)" (Rabinowicz, Strömberg 1996, s. 119). Formułując jednak zarzut, podnosi on argument, iż zasada warunkowej refleksji jest konieczna, aby uzyskać wymagany przez Hare'a rezultat, oraz, choć jest ona konieczna, nie tylko nie stosuje się do interpretacji Vendlera, ale zastosowana być nie może. Jak pisze:

> 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 [warunkowej refleksji – K.S.] 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 (Rabinowicz 2010).

> To, co „zmienia pozycję" w takim eksperymencie [w interpretacji Vendlera – K.S.], to nie ja, osoba, którą jestem, lecz *tylko* [podkr. – K.S.] „transcendentalne ja", aby użyć kantowskiej terminologii Vendlera – *jedynie* [podkr. – K.S.] rama świadomości, która w zasadzie może być wypełniona dowolną treścią. [...] 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 mi 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 (Rabinowicz 2010).

Hare, aby móc umożliwić wyjaśnienie frazy „w pełni postawić się na miejscu innej osoby" bez popadnięcia w zarzut, iż jest to metafizycznie niemożliwe ze względu na leibnizjańską zasadę tożsamości, stał na stanowisku, iż ja jako osoba, do której odnosi się moje imię i nazwisko, nie mogę być w pełni na miejscu innej osoby, pozostając sobą. Tożsamość osoby Krzysztofa Saji wyznaczona jest bowiem przez deskryptywne cechy uniwersalne, które, jeśli zostaną radykalnie zmienione, spowodują przerwanie tej tożsamości. Nie dotyczy to jednak pojęcia „ja", które nie jest powiązane z żadną istotą (Hare 2001, s. 153). Swojego znaczenia nie czerpie ono z odnoszenia się jedynie do cech deskryptywnych. Pojęcie „ja" ma przede wszystkim charakter preskryptywny: „ja" to każdy strumień świadomości, o który się troszczę, na spełnieniu preferencji którego mi zależy. Można powiedzieć, że „ja" jest to samoświadomość, której granice wyznacza zakres mojej troski własnej. Hare'a analiza znaczenia pojęcia „ja", której użył on do wyjaśnienia radykalnych eksperymentów myślowych ze zmianą ról, Vendler rozszerzył o aspekt metafizyczny. Zaproponował traktowanie „ja" jak kantowskie transcendentalne „ja", pustą ramę świadomości.

Powyżej zacytowana krytyka, dotykająca interpretacji Vendlera, swe źródło, jak mniemam, czerpie z nieporozumienia dotyczącego, z jednej strony, zakładanej przez Hare'a teorii tożsamości osobowej, z drugiej zaś ambicji, które miał on względem swojego argumentu. Profesor Rabinowicz bowiem, odpowiadając na retoryczne pytanie „o kogo się troszczę, gdy troszczę się o *siebie*?", odpowiada na nie inaczej, niż, moim zdaniem, uczyniłby to Hare. Co więcej, odpowiedź, iż troszczę się o Krzysztofa Saję – osobę, którą jestem – wcale nie jest, jak sugeruje on, oczywista[4]. Czy nie należałoby bowiem odpowiedzieć: nie troszczę się o osobę – Krzysztofa Saję – zbiór deskryptywnych własności – lecz o „ja", kimkolwiek jestem lub będę – o coś, co nie jest wyznaczone jedynie przez naturalne własności podlegające zmianom, ale je transcenduje? Troszczę się o „ja", które normowane jest przez zakres mojej troski (Hare 2001, s. 96–99). Stwierdzenie, iż nie

[4] Powyższa teoria znaczenia pojęcia „ja" zdaje się wyjaśnić wiele zawiłości związanych z tożsamością osobową, które generowane są przez czysto deskryptywną teorię znaczenia „ja". Przy jej użyciu możemy np. wyjaśnić przekonania religijne, iż po mojej śmierci Bóg wymierzy mi sprawiedliwość, choć w przyszłym życiu nie będę już taki sam (np. będę duchowy, bez wad itd.), czy przeświadczenie, że powinienem się troszczyć o interesy starca, którym będę za 60 lat. Nie chcę przez to powiedzieć, że uważam teorię tę za prawdziwą, lecz że jej błędność nie jest wcale oczywista.

troszczymy się przecież o jakąś „ramę świadomości" zwodzi nas więc, gdyż zakłada inną teorię znaczenia „ja" i sugeruje, że „ja" to cechy Krzysztofa Saji. Odpowiedź ta zwodzi nas również dlatego, iż uprzedmiotowia „ja", czyniąc je, *de facto* za Vendlerem (Hare 1988a, s. 280–281), ale wbrew intencjom Hare'a, „pustą ramą świadomości". Kto zaś może troszczyć się o pustą ramę? Pojęcie „ja" ma jednak dwie główne cechy. Po pierwsze, zdaniem Hare'a, w pojęciu „ja" zawiera się owa możliwość transcendencji poza moje aktualne cechy i choć „ja", w moim potocznym użyciu, wypełnione jest pewną treścią (nie jest puste), treść ta nie determinuje poprawnego użycia pojęcia „ja". Po drugie, pojęcie „ja" posiada preskryptywny charakter: na mocy znaczeń, tam, gdzie jestem „ja", tam jest troska własna. Z tego powodu w omawianej interpretacji eksperymentów myślowych, wbrew tezie Rabinowicza, troska własna gra olbrzymią rolę, wyznaczając granice poprawnego użycia słowa „ja". Rozważając rzeczywistą sytuację z trzech różnych aspektów $a_1 - a_3$, jeśli „ja" patrzę z tych punktów widzenia, to „ja", na mocy znaczenia tego pojęcia, posiadam preferencje, które są z tymi perspektywami związane. Ponieważ jednak, aby dojść do końcowego stanu preferencji, nie muszę rozważać odmiennych sytuacji, nie istnieje potrzeba zastosowania zasady *warunkowej* refleksji. Perspektywy $a_1 - a_3$ dotyczą bowiem sytuacji aktualnej i nie mają charakteru warunkowego. W ich trakcie realizuję jedynie zasadę tkwiącą w znaczeniu pojęcia „wiedzieć" oraz w trosce własnej: jeśli obecnie (w pełni) wiem, że aktualnie posiadam preferencję na korzyść A, to posiadam tę preferencję na korzyść A. Eksperymenty myślowe w jednoświatowej interpretacji Vendlera polegają więc na wyobrażeniu sobie, że jest się w skórze innych osób w ten sposób, iż patrząc na sytuację rzeczywistą z ich perspektywy jako „ja", nie tyle abstrahuję od tożsamości innych osób, ile, jak mówi Vendler, pomijam ją jako nieistotną (Vendler 1988, s. 181). Jestem tylko „ja", który poszerza swoją świadomość o inne perspektywy patrzenia, analogicznie do tego, jak czynię, gdy, dokonując racjonalnego egoistycznego wyboru, przypominam sobie lub obrazuję nowe fakty, które są wynikiem wyobrażeniowego spojrzenia z innej perspektywy. Wiedza o tych nowych faktach, jeśli wiem, że dotyczą *mnie*, zmienia moje stany preferencji. W tej interpretacji, podsumowując, nie korzysta się z zasady warunkowej refleksji, a jedynie z niedeskryptywnej analizy pojęcia „ja", które analitycznie związane jest z troską własną.

Czy do powyższego rozwiązania nie można jednak zgłosić zarzutu, iż zakłada ono zbyt mocną przesłankę, która pociąga za sobą niemożność

istnienia empatii bez sympatii? Argument taki wysunął Rabinowicz wobec zmodyfikowanej przez Petera Dietscha zasady warunkowej refleksji: „O ile wiem, co w rzeczywistej sytuacji preferowałbym z perspektywy kogoś innego, muszę posiadać odpowiednią preferencję (ten sam znak, taka sama siła)" (Rabinowicz 2010). Wydawać się może, iż przedstawiona powyżej interpretacja zakłada ową zmodyfikowaną zasadę. Jeśli tak jednak jest, to bazuje ona na nieprawdziwym i nieakceptowanym przez Hare'a założeniu (Hare 1988, s. 230; 2001, s. 99; Gibbard 1988, s. 62). Możemy wszakże, wbrew owej tezie, jak słusznie stwierdza Rabinowicz, znać emocje i przekonania innych, nie podzielając ich samemu. Należy jednak zauważyć, iż zmodyfikowaną przez Dietscha zasadę warunkowej refleksji odczytać można na dwa sposoby i jedynie pierwszy stanowi problem dla naszego wywodu. Interpretacja, którą w swej krytyce zakłada Rabinowicz, głosi:

> O ile w pełni wiem, co w rzeczywistej sytuacji preferuje *ktoś inny*, muszę posiadać korespondującą preferencję (ten sam znak, taka sama siła).

Drugi sposób jej rozumienia, który stanowi faktyczną przesłankę omawianej interpretacji, nie powoduje jednak nieprzychylnych dla nas konkluzji:

> O ile w pełni wiem, co w rzeczywistej sytuacji *ja* bym preferował z innej perspektywy (np. perspektywy, w której realnie znajduje się kogoś inny), muszę posiadać korespondującą preferencję (ten sam znak, taka sama siła).

Sformułowanie to nie podważa możliwości istnienia empatii bez sympatii. Empatia jest bowiem zdolnością odczuwania stanów psychicznych *innych osób*: wczuciem się w ich preferencje i pragnienia (empatia emocjonalna), umiejętnością przyjęcia ich sposobu myślenia i przekonań (empatia poznawcza). Druga interpretacja nie podlega więc powyższemu zarzutowi, gdyż nie dotyczy ona empatii – nie dotyczy *innych osób, tylko pojęcia „ja"*. Jest ona jedynie wyrazem troski własnej i znaczenia pojęcia 'wiedzieć'.

Omawiane rozumienie procesu myślenia moralnego, jako eksperymentu myślowego polegającego na poszerzeniu mojej świadomości o nowe preferencje związane z różnymi perspektywami patrzenia na daną rzeczywistą sytuację, doprowadza nas do poszukiwanego przez Hare'a rezultatu: do stanu

preferencji, który w wyniku racjonalnego wyboru może być odpowiednio zbalansowany. Rezultat takiej refleksji będzie zgodny z utylitaryzmem.

Dlaczego więc mielibyśmy chcieć, jak w innym miejscu sugeruje Rabinowicz, aby zasada warunkowej refleksji również tu była zastosowana? Jak pisze:

> Można zatem wciąż utrzymywać, że zasada warunkowej troski własnej ma do spełnienia istotną rolę: jest potrzebna, jeśli mam *zatrzymać* preferencję, która właściwie należy do perspektywy kogoś innego, również wtedy, gdy zdystansuję się do niej i wycofam się do tej perspektywy, którą aktualnie zajmuję (Rabinowicz, Strömberg 1996, s. 120).

Identyczne przekonanie, jak wyrażone w powyższym cytacie, funduje również inny argument, o którym wspominałem powyżej. Jeden z zarzutów przeciwko wielosytuacyjnej interpretacji jednoświatowej głosił, iż tak rozumiane eksperymenty myślowe generują rozstrzygnięcia, które nie mają mocy praktycznej. Rabinowicz wyraził to w haśle: „od udawania się zaczyna, na udawaniu się kończy", jeśli będziemy pozorowali brak wiedzy o hipotetyczności rozważanych światów możliwych, eksperymenty myślowe, które zakładają symulowanie niewiedzy, mogą nie dostarczać podmiotowi praktycznych racji, aby wykonać zalecane działanie, gdy powróci on ze świata wyobraźni do rzeczywistości.

Nie powinniśmy jednak godzić się na takie konkluzje, bowiem wymaganie stawiane Hare'owi jest zbyt ambitne. Pamiętać bowiem należy, iż w świetle jego etyki nie musimy zakładać, że amoralista konsekwentny jest osobą nieracjonalną, gdyż postawa taka nie musi być wewnętrznie sprzeczna (Hare 2001, s. 238). Jeśli ktoś konsekwentnie nie chce używać języka moralnego, może tego nie czynić, nie obawiając się sprzeczności. Może on więc, nie będąc posądzonym o nieracjonalność, przeprowadzić moralny eksperyment myślowy, zuniwersalizować preferencje, uzyskać utylitarystyczną konkluzję i „wycofać się do tej perspektywy, którą aktualnie zajmuje". Jak wyznał Hare:

> logika nie może prowadzić nas przez cały czas od wiary w fakty pozamoralne do dyspozycji do działania. Możemy więc wybrać albo rozziew między przekonaniami dotyczącymi faktów a sądami moralnymi, tak jak to robią preskryptywiści, albo, co wolą deskryptywiści, rozziew między sądami moralnymi a dyspozycjami do działania, utrzymując ścisły związek między przekonaniami dotyczącymi faktów a sądami

> moralnymi. [...] Ponieważ nie jesteśmy deskryptywistami, nie możemy utrzymywać, że istnieje ścisły związek między przekonaniami dotyczącymi faktów a sądami moralnymi; to zaś prowadzi, jak widzieliśmy, do tego, że konsekwentnemu amoraliście pozostawia się drogę ucieczki: może uznać wszystkie nasze fakty, a mimo to nie formułować żadnych sądów moralnych (Hare 2001, s. 237).

Pomimo że konkluzja taka nie jest optymistyczna, zauważyć trzeba, iż argument Hare'a, przy założeniu prawdziwości przesłanek, wydaje się omijać omówione trudności i prowadzi do swojego celu: do namysłu moralnego zgodnego z utylitaryzmem preferencji. To, iż działając możemy zignorować moralny punkt widzenia, gdyż nie dostarcza on najsilniejszej motywacji, nie podważa omawianego uzasadnienia utylitaryzmu.

Podsumowanie

Rozważane powyżej trzy sposoby interpretowania eksperymentów myślowych Hare'a, polegających na wejściu w skórę innych, mają swoje odrębne założenia, których wartość należałoby poddać ocenie. Moim celem nie była jednak ich ewaluacja, a jedynie zasypanie luki w argumencie Hare'a, którą omawia Rabinowicz. Jak wskazywałem powyżej, problem braku konfliktu powstaje jedynie wtedy, gdy w rozważanych eksperymentach myślowych nie weźmie się pod uwagę wymogu abstrahowania od informacji dotyczących tego, które z rozważanych sytuacji są rzeczywiste, a które hipotetyczne. Jeśli podążymy za sugestią Hare'a, aby sytuacje hipotetyczne traktować jako takie, które się zaktualizują, problem braku konfliktu znika. Nie istnieje on również, jeśli sytuacje hipotetyczne potraktujemy, tak jak Vendler, jako aktualne: jako różne aspekty tej samej sytuacji. Wbrew sugestiom Rabinowicza nie wydaje się także, co starałem się uzasadnić, aby propozycja Vendlera powodowała istotne komplikacje. Wymaga ona jedynie zaakceptowania Hare'a teorii znaczenia słowa „ja". Problem braku konfliktu dotyczy więc tych interpretacji, w których eksperymenty myślowe dokonywane są na wielu możliwych światach. Implikuje to bowiem, że nie wszystkie rozważane sytuacje mogą być zrealizowane, co czyni je nieistotnymi dla eksperymentatora. Jak jednak przyznał Hare, w odpowiedzi na tekst Vendlera, jeśli mówimy o własnościach uniwersalnych, w eksperymentach myślowych faktycznie rozważamy tylko jedną sytuację (1988a, s. 286–287). Jak pisał: „Świat,

w którym byłbym Napoleonem, byłby innym światem niż obecny, jednak nie pod względem własności uniwersalnych" (1988a, s. 285). Ponieważ jednolite zalecenie powinno być uniwersalne, eksperymenty myślowe muszą rozgrywać się w tylko jednym świecie możliwym.

Literatura

Gibbard A. (1988), *Hare's Analysis of 'Ought' and its Implications*, [w:] *Hare and his Critics: Essays on Moral Thinking*, Oxford: Clarendon Press, s. 57–72.

Hare R.M. (1984), *Some Reasoning about Preferences: A Response to Essays by Persson, Feldman, and Schueler*, „Ethics", s. 81–85.

Hare R.M. (1988a), *Comments on Vendler*, [w:] D. Seanor, N. Fotion (red.), *Hare and Critics: Essays on Moral Thinking*, Oxford: Clarendon Press, s. 280–287.

Hare R.M. (1988b), *Comments on Gibbard*, [w:] D. Seanor, N. Fotion (red.), *Hare and Critics: Essays on Moral Thinking*, Oxford: Clarendon Press, s. 229–234.

Hare R.M. (1989a), *Relevance*, [w:] R.M. Hare (red.), *Essays in Ethical Theory*, Oxford: Clarendon Press.

Hare R.M. (1989b), *Universalizability and the Summing of Desires: Reply to Ingmar Persson*, „Theoria" 55 (3), s. 171–177.

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

Persson I. (1989), *Universalizability and the Summing of Desires*, „Theoria" 55 (3), s. 159–170.

Rabinowicz W. (2009), *Preference Utilitarianism by Way of Preference Change*, [w:] T. Grüne-Yanoff, S.O. Hansson (red.), *Preference Change: Approaches from Philosophy, Economics and Psychology*, Dordrecht–London: Springer, s. 185–206.

Rabinowicz W. (1989), *Hare on Prudence*, „Theoria" 55 (3), s. 145–151.

Rabinowicz W. (2010), *Utylitaryzm preferencji poprzez zmianę preferencji?*, tłum. K. Saja, „Analiza i Egzystencja" 12, s. 7–36.

Rabinowicz W., Strömberg B. (1996), *What if I were in his Shoes? On Hare's Argument for Preference Utilitarianism*, „Theoria" 62 (1–2), s. 95–123.

Saja K. (2008), *Język a utylitaryzm. Filozofia moralna Richarda M. Hare'a*, Kraków: Aureus.

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

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

THOUGHT EXPERIMENTS AND UTILITARIANISM A REPLY TO WLODEK RABINOWICZ'S *PREFERENCE UTILITARIANISM BY WAY OF PREFERENCE CHANGE?*

Summary

This article is a comment on Wlodek Rabinowicz *Utilitarianism by Way of Preference Change?*, in which I defend Richard Hare's argument for utilitarianism. I argue that Hare's role reversal thought experiments can, despite of Rabinowicz's criticism, lead to transforming interpersonal preference conflicts into intrapersonal ones, but, in order to achieve this, we need to interpret his thought experiment correctly. I distinguish three interpretation of reversing roles experiment and argue, that for at least two of them, which de facto Hare endorse, Rabinowicz criticism fails.