Czasopismo Psychologiczne
Psychological Journal

Treść, obliczanie i eksternalizm*[1]

Christopher Peacocke
Magdalen College, Oxford

CONTENT, COMPUTATION AND EXTERNALISM

Artykuł ten zrodził się z chęci usunięcia pewnej narzucającej się niespójności. Między rozpowszechnioną koncepcją obliczania a wiarygodnym ujęciem wyjaśniania psychologicznego, rozpatrywanego w kontekście celów psychologii i faktycznego sposobu jej uprawiania, wydaje się istnieć niespójność. Zamierzam pokazać, że niespójność tę można usunąć przez przyjęcie i zastosowanie odmiennej koncepcji obliczania.

Ta koncepcja alternatywna wyjątkowo dobrze nadaje się do udzielenia odpowiedzi na pewnego rodzaju pytania dotyczące wyjaśniania. Będę dowodził, że alternatywnej koncepcji obliczania nie da się sprowadzić do rozpowszechnionego pojmowania tego fenomenu. Koncepcja alternatywna jest również nieodzowna dla właściwego rozwinięcia i obrony całego projektu obliczeniowego w psychologii. Rozpocznę od wyłożenia owej narzucającej się niespójności.

1. NIESPÓJNOŚĆ

Jedna z koncepcji obliczania mówi, że własności semantyczne reprezentacji nie mogą mieć wpływu na obliczanie. *Locus classicus* dla prezentacji tego podejścia jest artykuł Jerry'ego Fodora z 1980 r. pt. *Solipsyzm metodologiczny jako strategia badawcza w psychologii poznawczej*. „Zakładam, że procesy obliczania mają charakter zarówno symboliczny, jak i formalny" (Fodor, 1991, s. 486). Zdaniem Fodora operacje formalne nie muszą być syntaktyczne, aczkolwiek „operacje syntaktyczne są odmianą operacji formalnych przez fakt, że być syntaktycznym to nie być semantycznym. Operacje formalne są operacjami określanymi bez odniesienia do takich własności semantycznych reprezentacji, jak np. prawda, referencja i znaczenie" (tamże). „Jeśli procesy mentalne są formalne, to mają one dostęp jedynie do własności formalnych takich reprezentacji środowiska, jakie dostarczane są przez zmysły. Stąd też nie mają one dostępu do własności semantycznych tych reprezentacji, w tym także do [...] własności bycia reprezentacjami środowiska" (tamże, s. 488). Fodor bynajmniej nie jest odosobniony w swym podejściu do obliczania. Podejście to przyjmowane jest milcząco w większości prac z tej dziedziny. Nie zakłada ono bynajmniej innych poglądów Fodora na temat treści, ani też jego poglądów dotyczących języka myśli. Na pierwszy rzut oka jest całkowicie spójne jednoczesne przyjmowanie trzech następujących tez (aczkolwiek nie twierdzę, że jest to poprawna kombinacja): niektóre stany intencjonalne są realizowane przez niezdaniowe stany sieci koneksjonistycznej; niektóre sekwencje stanów tej sieci należy uznać za obliczenia; własności semantyczne nigdy nie odgrywają roli ani w wyjaśniających ani w wyjaśnianych stanach wyjaśnienia obliczeniowego.

* *Content, computation and externalism*. *„Mind and language"*, 9 (1994), ss. 303-335. ã Basil Blackwell Ltd.

[1] Wiele skorzystałem z cennych uwag Martina Daviesa, członków naszego wspólnego seminarium w Oksfordzie w trymestrze wiosennym 1993, uczestników spotkania w CREA w Paryżu, recenzenta i redaktorów *Mind and Language*, uczestników spotkania SOFIA, które odbyło się w Lizbonie w 1994 r., a zwłaszcza komentatorów mojego wystąpienia: Daniela Adlera i Josefa Toribio, oraz dyskutantów: Neda Blocka, Paula Boghossiana, Tylera Burge'a, Manuela Garcii Carpintero, Jerry'ego Fodora, Paula Horwicha, Briana Loara, Jerry'ego Katza, Pierre'a Jacoba i Jaegwona Kima.

W tym miejscu pojawia się sprzeczność. Gdyby prawidłowa była niesemantyczna koncepcja obliczania, to z pewnością większość współczesnej psychologii narażona byłaby na olbrzymie niedopasowanie celów do środków. Bez wątpienia większość teoretycznych rozważań psychologicznych zmierza do wyjaśnienia intencjonalnych, treściowych własności podmiotu. Wydaje się, na przykład, że mamy do czynienia z wyjaśnieniami tego, w jaki sposób intencjonalna własność doświadczenia polegającego na reprezentowaniu jakiegoś obiektu jako mającego określony kształt jest wynikiem obliczenia opartego na dwuwymiarowej informacji o ruchu tego obiektu. Mamy też do czynienia z obliczeniowymi wyjaśnieniami tego, w jaki sposób osoba słyszy jakieś zdanie jako mające takie a nie inne znaczenie; z obliczeniowymi wyjaśnieniami tego, dlaczego osoba wybiera taki a nie inny sposób postępowania; itd. Gdyby jednak niesemantyczna koncepcja obliczania była trafna, to formułowanie obliczeniowych wyjaśnień tych własności intencjonalnych byłoby obarczone zasadniczym błędem.

Według niesemantycznej koncepcji obliczania, wyjaśnienie obliczeniowe faktu, że dana osoba znalazła się w jakimś stanie intencjonalnym, polega na wyjaśnianiu, za pomocą takiej czy innej procedury obliczeniowej, jednego stanu niesemantycznego drugim stanem niesemantycznym. Ten drugi stan to podstawa lub realizacja wyjaśnianego stanu intencjonalnego (albo też to, co konstytuuje ów stan). Ale jeśli wyjaśnieniu podlegają tylko własności niesemantyczne, to co z wyjaśnieniem własności intencjonalnych? Wygląda na to, że przy niesemantycznej koncepcji obliczania wyjaśnić można jedynie niesemantyczne cechy stanów intencjonalnych. Gdyby warunki wyjaśniające wyjaśnienia obliczeniowego odwoływały się do treści, mielibyśmy wciąż pewną możliwość manewru – być może warunki wyjaśniające mogłyby zapewnić odpowiednie własności relacyjne wymagane, aby stan intencjonalny miał taką treść, jaką ma. Ale niesemantyczna koncepcja obliczania, zdaniem której warunki wyjaśniające są także niesemantyczne, wyklucza również ten wariant. Według tej koncepcji internalnie zindywidualizowane warunki wyjaśniające nie są w stanie wyczarować kompleksu relacji niesyntaktycznych, wymaganych by stan intencjonalny posiadał określoną treść.

To prawda, że niesemantyczne wyjaśnienia obliczeniowe, podobnie jak wszystkie inne wyjaśnienia, presuponują odpowiednie warunki wstępne. Kiedy jednak obliczanie jest rozumiane wyłącznie niesemantycznie, to owe warunki dotyczyć będą, jak się wydaje, jedynie takich spraw jak to, czy procesory reagujące na własności formalne działają prawidłowo itp. Tego typu warunki wstępne nie są w stanie uporać się z fundamentalnym problemem, który polega na tym, że koncepcja niesemantyczna nie wyjaśnia intencjonalnych własności stanu intencjonalnego. Ale kłopot nie w tym, że nie docenia się tu faktu, iż wyjaśnienia obliczeniowe to wyjaśnienia autentycznie empiryczne, a nie konstrukcje aprioryczne. Można całkowicie spójnie opowiadać się za wyjaśnieniem stanów treściowych przez inne stany, które bądź same są treściowe, bądź co najmniej presuponują rozmaite warunki niesyntaktyczne, a równocześnie uznawać, że owe wyjaśnienia psychologiczne muszą mieć charakter empiryczny i nie mogą być wymyślone *a priori*[2].

Jednakże można na to odpowiedzieć, że do wyjaśnienia intencjonalnych własności stanów mentalnych wystarczy, by obliczany stan posiadał własności syntaktyczne, dla których istnieje odpowiednia semantyka. Jeśli wyjaśnimy własności syntaktyczne, to czy dołączenie semantyki nie wyjaśni nam własności semantycznych obliczanego stanu? Aby na to odpowiedzieć, rozważmy przykład: obliczanie znaczenia określonego zdania *s*, będącego składową eksternalnego języka publicznego, na podstawie znaczeń jego części. Według niesemantycznej koncepcji obliczania, zostaje obliczony stan końcowy posiadający określoną złożoną, internalną własność syntaktyczną. Przy określonej semantyce dla tej składni, stany posiadające ową złożoną własność obliczonego stanu końcowego mają następującą treść: owo eksternalne zdanie *s* znaczy, iż *p*. Czy można powiedzieć, że z chwilą dołączenia semantyki do obliczanych

[2] Chociaż obliczeniowe wyjaśnienia działania w kategoriach stanów intencjonalnych stanowią mniej rozwiniętą część tej dyscypliny wiedzy, podobne uwagi można również poczynić pod adresem tego rodzaju wyjaśniania. Wyjaśnieniu psychologicznemu podlegają relacyjne, środowiskowe własności działania; a o tym, jakie własności podlegają wyjaśnieniu decydują treści wyjaśniających stanów intencjonalnych. Według niesemantycznej koncepcji wyjaśniania, stany, które wyjaśniają obliczeniowo nie mają dostępu do tych treści. Mamy więc analogiczny problem: w jaki sposób te stany wyjaśniające mogą wyjaśnić działania, których relacyjne opisy są ustalane na podstawie tych treści.

internalnych własności syntaktycznych, obliczanie niesemantyczne wyjaśnia rozumienie przez osobą zdania *s* jako znaczącego, iż *p*? Jeden z powodów, dla którego tak nie jest polega na tym, że owo wyjaśnienie nie korzysta z żadnych informacji o znaczeniu części składowych zdania s. Według koncepcji niesemantycznej, stany początkowe, na podstawie których zostaje obliczony stan końcowy, nie są w ogóle charakteryzowane w wyjaśnieniu jako dostarczające jakichkolwiek informacji na temat znaczenia części składowych zdania. Może więc – powie ktoś – powinniśmy je wzbogacić o charakterystyki semantyczne, mówiące, że rozmaite wyrażenia, które występują w s mają określone znaczenia? Jednakże po dokonaniu tych dwóch uzupełnień trudno będzie bronić tezy, że mamy tu do czynienia z czysto niesemantyczną koncepcją obliczania, a to z dwóch powodów.

Powód pierwszy to ten, że wygląda na to, iż jeden stan treściowy obliczamy na podstawie innego stanu treściowego i że obliczanie to potwierdza treściowe kontrfaktyczne okresy warunkowe („Gdybyśmy przyjęli, że słowo to ma inne znaczenie, to również całemu zdaniu należałoby przypisać inne znaczenie"). Powód drugi jest taki, że tożsamość poszczególnych własności syntaktycznych wymienionych w niesemantycznym wyjaśnieniu obliczeniowym nie ma dla wyjaśnienia treściowego żadnej mocy eksplanacyjnej. Istnieje, jak się zdaje, wyraźny sens, w którym dwie osoby mające różne języki myśli mogą obliczać te same znaczenia dla zdania s na podstawie tych samych informacji o jego częściach i czyniąc to w taki sam sposób z punktu widzenia poziomu treściowego. I odwrotnie, dwie osoby dokonujące pod względem syntaktycznym takiego samego obliczenia, mogą obliczać zupełnie różne znaczenia dla zdania s, na podstawie bardzo różnych, pod względem semantycznym, informacji na temat znaczeń jego części składowych.

Uwagi te skłaniają do refleksji natury ogólniejszej. W obliczu proponowanego wyjaśnienia faktu treściowego, zawsze powinniśmy sobie zadać następujące pytanie: „Czy proponowane warunki wyjaśniające mogłyby wystąpić również wtedy, gdyby zjawiska podlegające wyjaśnieniu miały inne treści?" Jeśli odpowiedź brzmi „tak", to wyjaśnienie jest w najlepszym razie niepełne. Dotyczy to zarówno proponowanych wyjaśnień tego, w jaki sposób podmiot znalazł się w danym stanie treściowym, jak i proponowanych wyjaśnień, dlaczego istnieją określone prawdopodobne związki między stanami intencjonalnymi. Nakłada to pewne ograniczenie zarówno na poszczególne wyjaśnienia psychologiczne zjawisk treściowych, jak i na filozoficzne ujęcie natury tychże wyjaśnień.

Nasza dotychczasowa argumentacja nie odwoływała się do żadnych określonych, mocnych tez na temat natury treści, z wyjątkiem bezwyjątkowej presupozycji, że własności semantyczne wykraczają poza własności syntaktyczne. Dotychczasowy wywód obowiązuje nawet przy internalistycznej koncepcji treści. Jeśli nawet własności semantyczne są indywidualizowane internalnie, tak że treść jest „wąska", to własności semantyczne i własności syntaktyczne nadal pozostają różnymi własnościami internalnymi i tych pierwszych nie da się wyjaśnić odwołując się do czysto syntaktycznych wyjaśnień obliczeniowych. Stałe logiczne są niekiedy uznawane za wyrażenia, do których stosuje się wąska koncepcja treści. Jeśli ich treść jest wynikiem jakiejś określonej roli we wnioskowaniu, to treść ta nie może być zdeterminowana wyłącznie przez ich własności syntaktyczne. To samo można powiedzieć o pojęciu bramki typu „i", o ile ma ono być scharakteryzowane semantycznie. Bramkę typu „i" można określić w relacji do pewnych przyporządkowań wartości 0 i 1 węzłom, z którymi się ona łączy. Nie jest to jeszcze charakterystyka semantyczna – gdyby bowiem przyporządkowanie wartości 1 danemu węzłowi miało znaczenie semantyczne fałszu a nie prawdy, bramka typu „i" funkcjonowałaby semantycznie jak alternatywa (węzeł na wyjściu wskazuje na fałsz wtedy i tylko wtedy, gdy oba węzły na wejściu wskazują na fałsz). Krótko mówiąc, zasada, według której składnia nie może determinować semantyki obowiązuje dla wszystkiego, co nosi znamiona treści. Wyzwanie, w postaci sformułowanej dotychczas w tym artykule, odnosi się do każdej bez wyjątku teorii gotowej w ogóle stosować pojęcie treści.

Dla eksternalistów w kwestii treści, do których się zaliczam, widoczny rozziew między zasobami niesemantycznej koncepcji obliczania a wyjaśnianymi stanami intencjonalnymi jest rzecz jasna jeszcze większy[3]. Eksternalista doda jeszcze do poprzedniego wywodu spostrzeżenie, że wyjaśnie-

[3] Na temat eksternalizmu w kwestii treści zob. na początek Putnam, 1975 oraz Burge, 1979.

nia oparte wyłącznie na stanach syntaktycznych mogą wyjaśnić jedynie stany zindywidualizowane internalnie, gdy tymczasem stany intencjonalne, które mają być wyjaśnione nie są zindywidualizowane internalnie.

Chciałbym teraz rozwinąć nieco jedną szczególną koncepcję natury stanów zindywidualizowanych eksternalnie. Robię to nie po to, by dalej argumentować na rzecz rozziewu między niesemantyczną koncepcją obliczania a dowolnym wyjaśnieniem stanów intencjonalnych (argumentacja ta została już przedstawiona). Rozwinięcie to wprowadzam z dwóch powodów: (a) określa ono dalsze ograniczenia, jakim podlega dobre wyjaśnienie empiryczne stanów intencjonalnych, (b) pozytywne rozwiązanie, które będę zalecał, odwołuje się faktycznie do tej samej ogólnej koncepcji, którą ilustruje podana niżej eksternalistyczna charakterystyka natury stanów intencjonalnych.

Z wyjątkiem bardzo szczególnych okoliczności, tym, co podlega wyjaśnieniu w zwykłym wyjaśnieniu psychologicznym nie jest ruch ciała. Wyjaśnieniu podlega raczej fakt, że ma miejsce ruch ciała o pewnych własnościach relacyjnych wiążących się z mniej lub bardziej rozległymi okolicznościami zewnętrznymi – fizycznymi, psychicznymi i społecznymi. Dany ruch ciała posiada nieskończenie wiele własności relacyjnych. Określony ruch ręki może być jednocześnie: ułożeniem ręki między słońcem a oczyma, tak by spowodować cień; gestem mającym na celu zwrócenie uwagi przyjaciela; fragmentem sygnalizacji semaforowej; itd. Konkretne wyjaśnienie psychologiczne może wyjaśnić pojawienie się jakiegoś zdarzenia posiadającego niektóre spośród tych własności relacyjnych, nie wyjaśniając innych. Charakterystyczne dla różnych wyjaśnień jest to, że potwierdzenie znajdują odmienne kontrfaktyczne okresy warunkowe. Gdy dane wyjaśnienie rzeczywiście wyjaśnia własność relacyjną zdarzenia, jaką jest ułożenie przez podmiot ręki między słońcem a oczyma, to – przy zachowaniu innych okoliczności – potwierdzony zostanie następujący kontrfaktyczny okres warunkowy: gdyby słońce znajdowało się w innym położeniu, ręka podmiotu ułożyłaby się inaczej. Jeśli wyjaśnianym zjawiskiem jest natomiast nadawanie określonego sygnału semaforowego, to – przy zachowaniu innych okoliczności – powyższy kontrfaktyczny okres warunkowy nie zostanie potwierdzony. Potwierdzone zostaną za to inne kontrfaktyczne okresy warunkowe.

To, co odnosi się do wyjaśnień za pomocą stanów intencjonalnych, odnosi się również, odpowiednio, do niektórych wyjaśnień stanów intencjonalnych. Oto jeden szczególnie dobitny przykład: kiedy w wyniku pewnej mojej wypowiedzi dochodzicie do jakiegoś przekonania, to kluczową rolę odgrywa tu na ogół sens mojej wypowiedzi. Z wyjątkiem szczególnych przypadków, poszczególne wypowiadane przeze mnie słowa nie odgrywają tu swoistej roli – wszelkie inne słowa, które w waszym rozumieniu znaczyłyby to samo, doprowadziłyby do ukształtowania tych samych przekonań. To samo dotyczy innych względnie wewnętrznych własności wypowiedzi, takich jak wysokość czy natężenie (w pewnych granicach!). Natomiast posiadanie przez zdanie pewnego znaczenia jest własnością wysoce relacyjną. Podobne uwagi można by sformułować pod adresem wiedzy uzyskiwanej przez percepcję.

Wysoce prawdopodobna staje się następująca teza ogólna: tożsamość dowolnego stanu mającego treść intencjonalną jest przynajmniej częściowo konstytuowana przez fakt, że w odpowiednich okolicznościach może on wyjaśnić relacyjne własności obiektów i zdarzeń eksternalnych, albo też być przez nie wyjaśniony. Zakładam, że ta ogólna teza jest trafna. A jeśli jest trafna, to zadowalające będzie tylko takie obliczeniowe wyjaśnienie faktu, iż ktoś znalazł się jakimś stanie intencjonalnym, które jest zarazem wyjaśnieniem stanu posiadającego te specyficzne możliwości wyjaśnienia relacyjnych własności zdarzeń.

Korzyści płynących z wyjaśnienia za pomocą eksternalnie zindywidualizowanych stanów nie można osiągnąć przez zwykłe uzupełnienie wyjaśnienia odwołującego się wyłącznie do stanów zindywidualizowanych internalnie. Weźmy jakieś zdarzenie scharakteryzowane internalnie w kategoriach konstytuujących go ruchów ciała i rozważmy wyjaśnienie tak scharakteryzowanego zdarzenia odwołujące się do internalnych stanów podmiotu. Te stany internalne mogą być stanami neurofizjologicznymi, „syntaktycznymi", a nawet stanami o wąskiej treści (jeśli takowe istnieją). Wyobraźmy sobie teraz, że uzupełniamy to wyjaśnienie prawdziwym twierdzeniem na temat pewnych relacji środowiskowych, w których pozostaje owo wyjaśniane zdarzenie. Czy tak uzupełnione wyjaśnienie sprowadza się do wyjaśnienia relacji środowiskowych, w których pozostaje wyjaśniane zdarzenie? Bynajmniej nie. Gdyby mogło ono od-

grywać rolę takiego wyjaśnienia, to wyjaśnienie jakiegoś zdarzenia przy jednym z jego opisów byłoby wyjaśnieniem go przy wszystkich jego opisach, a to jest posunięcie zbyt daleko idące. Zdarzenie dające się wytłumaczyć jako wskazywanie przeze mnie palcem Gwiazdy Polarnej nie musi dać się wytłumaczyć jako wskazywanie przeze mnie palcem jakiegoś dalej położonego, odległego ciała niebieskiego w Drodze Mlecznej. Nie musi dać się wytłumaczyć, nawet jeśli z powodów wypływających z praw Gwiazda Polarna i owo dalej położone ciało znajdują się zawsze w tym samym kierunku. Dalsze niezbędne rozwinięcie tej koncepcji wyjaśniania własności relacyjnych można znaleźć w literaturze przedmiotu[4].

W świetle napięcia między potrzebą wyjaśnienia semantycznych własności stanów intencjonalnych a czysto niesemantyczną koncepcją obliczania, można jedynie sympatyzować z reakcją Stephena Sticha na klasyczny artykuł Fodora: „bezkompromisowa akceptacja paradygmatu obliczeniowego pociąga za sobą odrzucenie reprezentacyjnej teorii umysłu" (Stich, 1991). Gdyby niesemantyczne wyjaśnienia obliczeniowe były jedynymi możliwymi rodzajami wyjaśnień, istniałyby uzasadnione powody do utraty nadziei na wyjaśnienie własności specyficznie intencjonalnych. Istnieje jednak inna koncepcja wyjaśnienia obliczeniowego, która pozwoli nam wypośrodkować między poglądem, że psychologia obliczeniowa opiera się na błędzie, z jednej strony, a radykalnym sceptycyzmem Sticha, z drugiej strony.

2. OBLICZANIE ZE WZGLĘDU NA TREŚĆ

Gdy charakteryzujemy jakiś system jako, na przykład, obliczający znaczenie zdania na podstawie jego opisu syntaktycznego, albo też jako obliczający kształt przedmiotu na podstawie dwuwymiarowych informacji o jego ruchu, to intuicyjnie pragniemy powiedzieć, że zdarzenie bądź stan o jednej treści reprezentacyjnej zostaje wyjaśniony przez wystąpienie wcześniejszego zdarzenia lub stanu o innej treści oraz że wyjaśnienie to jest zgodne z określoną regułą. Intuicję tę można jeszcze bardziej uwyraźnić przez wprowadzenie pojęcia uwzględniającego treść opisu obliczeniowego pary zdarzeń (lub określonych stanów). Opis taki powinien zawierać trzy elementy:

(a) wyszczególnienie własności treściowej pierwszego zdarzenia (lub stanu);
(b) wyszczególnienie własności treściowej drugiego zdarzenia (lub stanu);
(c) wyszczególnienie reguły treściowej, która stwierdza, jak obliczyć własność treściową daną w (b) z własności treściowej danej w (a). To trzecie wyszczególnienie jest opisem algorytmu treściowego.

Warunki (a) – (c) można w prosty sposób uogólnić by objąć nimi przypadek, w którym treść jakiegoś zdarzenia lub stanu jest obliczana z treści kilku różnych zdarzeń lub stanów.

Pojęcie treści, które pojawia się w tych charakterystykach jest niemal w pełni ogólne. Treścią stanu może być coś z poziomu odniesienia, np. wartość określonej wielkości fizycznej bądź położenie osi symetrii jakiegoś przedmiotu. Ale równie dobrze może być to coś z poziomu sensu. Nie jest też konieczne, aby w poprawnym treściowym opisie obliczeniowym danej pary zdarzeń obie treści należały do tego samego typu ogólnego. Przeciwnie, całkiem możliwe, że istnieją ważne przypadki, w których muszą one należeć do różnych typów. Przykładem czegoś takiego jest sytuacja, w której pojęcie, pod które coś wydaje się podpadać zostaje obliczone na podstawie własności i wielkości z poziomu referencji.

Coś, co podpada pod treściowy opis obliczeniowy może też podpadać (i będzie podpadać) pod inne, nietreściowe opisy. Tak więc własności treściowe i własności nietreściowe mogą być własnościami tego samego procesu. Warto też zaznaczyć, że z podanych charakterystyk wcale jednoznacznie nie wynika, że każda para zdarzeń lub stanów podpadająca pod treściowy opis obliczeniowy musi w jakiś sposób być związana z wyrażeniami języka myśli. Aby coś takiego wykazać należałoby podać odrębny, mocny argument.

Gdyby własności treściowe mogły wyjaśniać jedynie inne własności treściowe i gdyby same mogły być wyjaśniane wyłącznie przez inne wła-

[4] Zob. Peacocke, 1993; natomiast w kwestii wcześniejszych omówień eksplanandów relacyjnych patrz Burge, 1986; Hornsby, 1986 oraz Peacocke, 1981. „Zależne od przedmiotu" podejście do sensów jednostkowych, zaprezentowane w ostatniej z prac, nie musi być składnikiem obecnego poglądu. Zob. Peacocke, 1993, § 2.

sności treściowe, to uwzględniające treść opisy obliczeniowe były eksplanacyjnie odcięte od nietreściowych własności świata. Jeśli chcemy uniknąć tej niemiłej konsekwencji, to musimy zaakceptować dwa rodzaje przypadków mieszanych. Opis pary zdarzeń (lub stanów) w danym systemie jest przypadkiem mieszanym pierwszego rodzaju, gdy zawiera trzy elementy:

(a) nietreściową własność pierwszego zdarzenia;

(b) treściową własność zdarzenia drugiego;

(c) sformułowanie ogólnej reguły stwierdzającej, w jaki sposób własności treściowe, w tym ta wyszczególniona w (b), zostają w tym systemie wyjaśnione przez własności nietreściowe, w tym tą wyszczególnioną w (a).

Opis sposobu, w jaki w percepcji wzrokowej pierwsze treści są obliczane na podstawie nietreściowych własności obrazu na siatkówce będzie przypadkiem mieszanym pierwszego rodzaju.

Opis przypadku mieszanego drugiego rodzaju stanowi, naturalnie, odbicie lustrzane powyższej sytuacji: zdarzenie pierwsze zostaje opisane jako posiadające własność treściową i wyjaśnia ono nietreściową własność zdarzenia drugiego. Przypadki mieszane drugiego rodzaju są konieczne, by mogły być wyjaśnienia obliczeniowe, które kończą się treściowymi wyjaśnieniami działania.

Proponuję przyjąć, że wyjaśnienia zawarte w prawdziwych, treściowych, obliczeniowych opisach zdarzeń są szczególnymi przypadkami ogólnego typu, który ustaliłem w Wyjaśnianiu eksternalistycznym (Peacocke, 1993) i który został powyżej naszkicowany. Są to przypadki szczególne, w których jeden eksternalnie zindywidualizowany stan wyjaśnia inny. W przytoczonych wcześniej przykładach wyjaśnienia za pomocą zwykłych stanów intencjonalnych, mamy do czynienia z niepsychologicznym, środowiskowo zidentyfikowanym eksplanandum (lub z czymś, co zawiera takie eksplanandum), jak np. z ruchem w kierunku linii znajdującej się między oczyma podmiotu a słońcem. Dla obecnych celów określmy je mianem „eksplananda pierwotne". Są to takie eksplananda, których status jako eksternalnie zindywidualizowanych nie zależy od ich relacji do jakiegoś innego eksternalnego lub eksternalnie zindywidualizowanego stanu. Po prostu one same dotyczą środowiska. Po stronie warunków wyjaśniających można by równie dobrze wprowadzić pojęcie eksplanansu pierwotnego.

Stan eksternalnie zindywidualizowany niekoniecznie musi być stanem zindywidualizowanym przez swe relacje do pierwotnych eksplanandów lub eksplanansów. Może to być stan drugiego stopnia, zindywidualizowany przez swe relacje do stanów pierwszego stopnia, które z kolei są zindywidualizowane przez swe relacje do pierwotnych eksplanandów lub eksplanansów. (Postawy propozycjonalne drugiego stopnia mogą mieć taki właśnie charakter). Eksternalnie zindywidualizowany stan może też być stanem trzeciego stopnia, zindywidualizowanym przez swe relacje do takich stanów drugiego stopnia; i tak dalej. Wystarczającym warunkiem eksternalnej indywidualizacji jakiegoś stanu jest to, aby znajdował się on gdzieś w tej hierarchii stopni.

Treściowe wyjaśnienie obliczeniowe jest na ogół wyjaśnieniem eksternalnie zindywidualizowanego stanu treściowego przez stan o takim samym charakterze. Kiedy zarówno stan wyjaśniający, jak i stan wyjaśniany mają charakter treściowy, wyjaśnienie potwierdza szczególny rodzaj kontrfaktycznych okresów warunkowych, w których poprzednik i następnik mają charakter eksternalistyczny. Gdyby system nie znajdował się w pierwszym, treściowym stanie, to nie znalazłby się w drugim, obliczonym stanie treściowym. (Pomijam tu kwestię przedeterminowania i funkcji wielo-jednoznacznych). Z uwagi na algorytm wykorzystany w obliczaniu potwierdzenie uzyska także kontrfaktyczny okres warunkowy mówiący, że gdyby system znajdował się w jakimś innym uprzednim stanie treściowym, to znalazłby się następnie w jakimś innym stanie treściowym. Kontrfaktyczne okresy warunkowe tego typu nie wyszczególniają żadnego wewnętrznego, internalnie zindywidualizowanego stanu, w którym znalazłby się organizm, gdyby poprzednik był prawdziwy. W istocie jedną i tę samą własność kontrfaktyczną występującą w tych okresach warunkowych można przypisać zgodnie z prawdą organizmom o dość różnych reprezentacjach internalnych, pod warunkiem, że ich relacje eksternalne są takie, że uzasadniają te same opisy treściowe.

Wszystkie powyższe uwagi odnoszą się wprost do treściowych obliczeń subpersonalnych w systemie wzrokowym. Algorytm służący do obliczania głębi na podstawie specyfikacji dwóch obrazów może stanowić odpowiedni składnik obliczeniowego wyjaśnienia percepcji głębi przez dwa różne

organizmy, których mentalne reprezentacje głębi różnią się dość znacznie pod względem ich bardziej wewnętrznych własności. Podobne uwagi można poczynić pod adresem wyjaśnienia rozumienia języka. Weźmy przykład, w którym można przyjąć, że poprzednik ma charakter eksternalistyczny. Tyler i Marslen-Wilson wykazali, że w zdaniu rozpoczynającym się od „*If you walk too near the runway, landing planes...*", wyraz „*landing*" zostanie odczytany raczej jako przymiotnik aniżeli rzeczownik odsłowny. Odwrotnie natomiast będzie w przypadku zdań rozpoczynających się od „*If you've been trained as a pilot, landing planes...*" (Marslen-Wilson i Tyler, 1987). Niewiarygodne wydaje się przypuszczenie, że wyjaśnienie tego ma coś wspólnego z poszczególnymi słowami w poprzedniku zdania, rozpatrywanymi niezależnie od ich znaczenia. Można oczekiwać, że podobny efekt wystąpiłby w każdym zdaniu języka angielskiego, w którym poprzednik posiadałby te same cechy znaczeniowe; a znaczenie jest na pewno pojęciem eksternalistycznym. Innego przykładu dostarcza zjawisko eliminacji wieloznaczności w eksperymentach dotyczących słyszenia dychotycznego (Lackner i Garrett, 1972). W eksperymentach tych znaczenie zdań słyszanych w kanale, na którym nie jest skupiona uwaga zniekształca interpretację wieloznacznego zdania słyszanego w kanale, na którym skupiona jest uwaga. Każde inne zdanie o identycznym znaczeniu oddziaływałoby podobnie, tak jak i wiele sposobów wypowiadania zdań słyszanych faktycznie w kanale, na którym nie jest skupiona uwaga. W tych przypadkach mamy do czynienia z wyjaśnieniem jednego stanu treściowego przez inny stan treściowy oraz z potwierdzonymi kontrfaktycznymi okresami warunkowymi o charakterze eksternalistycznym.

Czym innym jest przytaczanie przykładów, czym innym zaś skonstruowanie ogólnej koncepcji subpersonalnego, obliczeniowego wyjaśnienia treściowego. Ogólna koncepcja powinna zawierać przynajmniej następujące elementy. Po pierwsze, będziemy potrzebować jakiegoś opisu zasad rządzących prawidłowym przypisywaniem treści subpersonalnym stanom obliczeniowym. Niejasność w tej kwestii zawsze znajdzie odbicie w niejasności co do wagi wyjaśnień związanych z tymi stanami. Po drugie, potrzebna nam jest ogólne, pozytywne ujęcie charakterystycznych cech obliczeniowego wyjaśnienia treściowego. Powinien temu towarzyszyć, po trzecie, opis celów, które osiągnąć można jedynie dzięki wyjaśnieniom opartym na obliczeniach treściowych oraz podanie racji, dlaczego tak jest. Podejmę się tych zadań po kolei, pierwszego w części następnej, a pozostałych w częściach 4 i 5.

3. PRZYPISYWANIE TREŚCI SUBPERSONALNYCH

Przypisywanie treści z poziomu personalnego przez psychologię postaw propozycjonalnych podlega nadrzędnemu wymogowi uczynienia podmiotu tych askrypcji zrozumiałym (Davidson, 1984; Grandy, 1973; McDowell, 1986). Z tym ogólnym stwierdzeniem zgodziliby się teoretycy różniący się skądinąd co do tego, czy wymóg ten dałoby się jeszcze bardziej sprecyzować, a jeśli tak, to w jaki sposób. Czy analogiczny wymóg nadrzędny istnieje w przypadku przypisywania treści subpersonalnych? Wiarygodny wariant tego wymogu wygląda następująco:

> prawidłowe askrypcje treści stanom subpersonalnym podlegają faktom dotyczących relacyjnych (środowiskowych) własności wyjaśnianych zdarzeń oraz kontrfaktycznym okresom warunkowym dotyczącym relacyjnych własności zdarzeń, które owe askrypcje wyjaśniałyby w rozmaitych okolicznościach kontrfaktycznych.

Przez powiązanie subpersonalnych askrypcji treści z relacyjnymi własnościami zdarzeń, wymóg ten nadaje treści subpersonalnej charakter eksternalistyczny.

Zdarzeniem, którego własności relacyjne pomaga wyjaśnić stan o treści eksternalistycznej, może być działanie – przypadek eksplanandu pierwotnego. Ale wyjaśniane zdarzenie może również polegać na przejściu systemu subpersonalnego w stan, który sam posiada treść eksternalistyczną. W takim przypadku wymóg nadrzędny działa rekursywnie.

Czym jest relacja podlegania *answerability* wspomniana w sformułowaniu tego wymogu? Załóżmy, że jakiś zbiór stanów, w tym stanów subpersonalnych, reprezentuje łącznie świat jako mający określoną postać. Wobec tego gdy podmiot tych stanów posiada pewien cel, stany te sprawiają, iż odpowiednie staje się określone zachowanie,

scharakteryzowane eksternalnie poprzez jego relacje do świata. W tej sytuacji odpowiednie może być także wejście w szereg innych, eksternalnie zindywidualizowanych stanów. W pierwszym, bardzo niedokładnym przybliżeniu i pomijając wszelkiego rodzaju szczegóły, o którym będzie mowa poniżej, można najogólniej powiedzieć, co następuje: treść przypisywana stanom subpersonalnym musi być treścią, która czyni odpowiednimi własności relacyjne wyjaśnianych przez te stany zdarzeń oraz która czyni odpowiednimi własności relacyjne zdarzeń, które byłyby przez te stany wyjaśniane w okolicznościach kontrfaktycznych.

Przypisywanie treści stanom subpersonalnym zgodnie z tym wymogiem nadrzędnym musi być również ograniczone, by tak rzec, z góry. Należy odrzucić zbiór askrypcji, który jest bardziej zróżnicowany niż mogą to uzasadnić prawdy relacyjne wymienione w wymogu nadrzędnym. Jest to naturalne uogólnienie tego, co w książce *Zmysły i treści* (Peacocke, 1983) nazwałem wymogiem spoistości (*Tightness Constraint*).

Stany mentalne, o jakich mowa w potocznej psychologii postaw propozycjonalnych są zindywidualizowane eksternalnie. Zwróciliśmy też uwagę na wiarygodność tego, iż owe potoczne stany są (przynajmniej) częściowo zindywidualizowane przez ich moc wyjaśniania relacyjnych własności działań. Jeśli tak jest, to wyjaśnianie przez treściowe stany subpersonalne może obejmować wyjaśnianie postaw propozycjonalnych. Kiedy tak się dzieje, treść będąca treścią stanu z poziomu personalnego jest obliczana z treści niektórych stanów subpersonalnych. W takich przypadkach wyjaśniające stany subpersonalne będą się również przyczyniały, na mocy przechodniości, do wyjaśnienia tego, co wyjaśniają stany z poziomu personalnego. Oczywiście nie prowadzi to do eksplanacyjnej jałowości stanów z poziomu personalnego. Wręcz przeciwnie, wyjaśnienie stanów końcowych tego łańcucha przez stany początkowe opiera się na tym, że wywołują one stany z poziomu personalnego.

Dotychczasowy opis wymogu nadrzędnego może sprawiać wrażenie skoncentrowanego na przypisywaniu treści elementom struktury organizmu znajdującym się na wyjściu, z pominięciem treści subpersonalnych obliczanych w procesach percepcyjnych. Ale w rzeczywistości, z przyczyn zasadniczych, obie te rzeczy idą w parze. Załóżmy, że jakąś czynność można wytłumaczyć przy opisie relacyjnym „zabranie pięćdziesięciopensówki z szufladki z monetami". Wytłumaczenie tej czynności w taki sposób wiąże się z prawdziwością pewnych kontrfaktycznych okresów warunkowych: przy zachowaniu innych okoliczności, podmiot zabrałby pięćdziesięciopensówkę nawet gdyby znajdowała się ona w innym miejscu w szufladce; itd. Jeśli prawdziwość tego zestawu kontrfaktycznych okresów warunkowych nie ma graniczyć z cudem, to podmiot musi dysponować jakąś percepcyjną zdolnością identyfikowania pięćdziesięciopensówek; a to, że znajduje się on w stanie percepcyjnym umożliwiającym identyfikację pięćdziesięciopensówek jest czymś, dla czego odpowiednie jest treściowe wyjaśnienie obliczeniowe. Jest to uwaga natury dość ogólnej i odnosi się ona do wszystkich percepcyjnie zidentyfikowanych własności, relacji, rodzajów i wielkości, które mają swój udział w relacyjnej charakterystyce przy której dana czynność jest wyjaśniana. Nie można zrozumieć zestawu stanów wyjaśniających zachowanie przy opisach odwołujących się do środowiska, jeśli część tych stanów, w odpowiednich okolicznościach, sama nie podlega wyjaśnieniu w kategoriach pewnych warunków środowiskowych.

Wcale nie jest oczywiste, że wymóg nadrzędny ogranicza wyjaśnianie za pomocą stanów o treści subpersonalnej do przypadków, w których ostatecznym celem wyjaśniania jest coś, co znajduje się na poziomie psychologii postaw propozycjonalnych lub jej charakterystycznych eksplanandów. Wydaje się rzeczą oczywistą, że istnieją eksplananda relacyjne odpowiednie do tego, by powiązać je ze stanami treściowymi, chociaż nie dotyczą one spraw, które zwykle, czy kiedykolwiek, podlegają intencjonalnej kontroli. Przykładem pierwszego rodzaju – zwykle nie podlegającym kontroli intencjonalnej – byłoby wyjaśnienie, w jaki sposób stojący organizm utrzymuje równowagę. Niewątpliwie wchodzą tu w grę eksplananda o charakterze relacyjnym, ponieważ chodzi tu o relacje organizmu do pionu grawitacyjnego. Wcale nie jest oczywiste, że istnieje zasadnicza racja a priori, aby w tym przypadku nie mogłyby być odpowiednie subpersonalne wyjaśnienia obliczeniowe. Nic nie stoi na przeszkodzie, by organizm dwunożny zawierał podsystem obliczający, czy linia między jego stopami znajduje się grawitacyjnie poniżej jego środka ciężkości, a jeśli nie, to gdzie się

znajduje (i jak wobec tego ma się on zachować). Mówiąc ogólniej, w przypadku mechanizmu posiadającego środowiskowo zindywidualizowaną zdolność, jest możliwe, że odpowiednim wyjaśnieniem posiadania tej zdolności jest obliczenie treściowe. Przede wszystkim przychodzą tu na myśl mechanizmy będące wytworem doboru naturalnego. Ale nawet w przypadku mechanizmu będącego pierwotnie wytworem przypadkowej mutacji musi istnieć wyjaśnienie trwałej, środowiskowo zindywidualizowanej zdolności.

W myśl bronionego przeze mnie wymogu nadrzędnego, stan treściowy, który jest rzetelnie wytwarzany w określonym typie środowiska, wcale nie musi prawidłowo reprezentować tego środowiska. Zgodnie z wymogiem nadrzędnym, można prawomocnie przypisywać treści w sposób, który prowadzi do nieprawidłowego obliczenia wielkości, pod warunkiem, że przypisanie wielkości prawidłowej doprowadziłoby do niewyjaśnionych rozbieżności między działaniami podmiotu a opisami, jakich należałoby się spodziewać, gdyby treść została obliczona prawidłowo. Błędy w percepcji odległości lub kierunku będą miały swoje konsekwencje, zwłaszcza dla relacyjnych własności działań, które są intencjonalne przy odpowiednich przestrzennych opisach uwzględniających odległość lub kierunek. Tego rodzaju przypadki możemy nazwać wybiegającym w przód potwierdzeniem przypisania błędnej wielkości.

Równie dobrze mogą też być wsteczne potwierdzenia niespełnionej treści w wytwarzaniu działania. Ktoś może usłyszeć trudny do wymówienia zwrot i nie móc go samemu wymówić. Gdybyśmy mu przypisali skuteczny zamiar wypowiedzenia pewnego zdania, musielibyśmy dojść do wniosku, że źle usłyszał wypowiedziane w jego obecności zdanie. Ale takie przypuszczenie musiałoby zostać podważone przez rozmaite inne czynności relacyjne, które ta osoba mogłaby wykonać, jak jej zdolność zapisania usłyszanego zdania, powiązanie jego składników z innymi usłyszanymi wypowiedziami itd. A zatem wymóg nadrzędny jest raczej subpersonalnym odpowiednikiem zasady z poziomu personalnego, mówiącej że powinniśmy dążyć do maksymalizacji zrozumiałości, a nie zasady życzliwości (mówiącej, że powinniśmy maksymalizować poprawność).

4. ODPOWIADANIE NA PYTANIA „JAK"

Przechodzę teraz do drugiego z wymienionych pod koniec części 2 zagadnień, a mianowicie do charakterystycznych cech obliczeniowych wyjaśnień treściowych. W tej i następnej części zajmiemy się przede wszystkim zagadnieniami z filozofii wyjaśniania. Czytelnicy specjalizujący się w naukach kognitywnych, którzy mają mniej do czynienia z takimi problemami, powinni przejść bezpośrednio do części 6.

Zdarza się, że ludzie uprawiający daną naukę zmuszeni są odpowiedzieć na pytanie następującego rodzaju. Wiedzą już, że pewna rzecz posiada daną własność i chcą się teraz dowiedzieć, jak to się dzieje, że owa rzecz może tę własność posiadać. Nazwijmy tego rodzaju pytania pytaniami „jak". Oto przykłady takich pytań: „Jak organizm ludzki jest w stanie zapobiegać gromadzeniu się we krwi zbędnych produktów przemiany materii?"; „Jak to się dzieje, że osoba jest w stanie zrozumieć zdanie, z którym nigdy wcześniej się nie zetknęła?" Pytanie „jak" musi posiadać odpowiedź, jeśli własność, której dotyczy nie ma być jakimś cudem. Potrzeba znalezienia odpowiedzi jest szczególnie nagląca w przypadku własności zakładających relacje rozważanego obiektu do innych rzeczy.

Pytania o postaci „Jak to się dzieje, że określony typ organizmu może się znajdować w stanach o takiej to a takiej treści?" tworzą jeden z podzbiorów pytań „jak". Nadal będę zakładał słuszność argumentów na rzecz poglądu, że treść jest czymś eksternalnie zindywidualizowanym, w tym sensie, iż dany stan mentalny dlatego ma taką a nie inną treść, że pozostaje on w pewnych złożonych relacjach z innymi rzeczami, własnościami i relacjami w środowisku podmiotu tego stanu. Jeśli tak, to pytania o postaci: „Jak to się dzieje, że określony typ organizmu może się znajdować w stanach o takiej to a takiej treści?" są, mówiąc bardziej szczegółowo, pytaniami „jak" dotyczącymi stanów relacyjnych wchodzących w grę organizmów. Uwaga ta stosuje się nie tylko do percepcji, lecz także do rozumienia języka, przekonań, intencji, pragnień, emocji oraz wszelkich innych stanów o treściach zindywidualizowanych eksternalnie.

Teoria DNA i jego zdolności dostarcza odpowiedzi na następujące pytanie „jak": „Jak to się

dzieje, że zasady genetyki mendlowskiej stosują się do ludzi?"[5]. Warto się zastanowić nad tym przykładem, ponieważ pod pewnymi względami przejawia on strukturalne podobieństwo do przypadku psychologicznego, a zarazem pod innymi względami jest strukturalnie odmienny. Posiadać jedną z własności zidentyfikowanych przez genetykę mendlowską to posiadać własność wysoce relacyjną. Posiadanie przez jakąś osobę genu recesywnego dla rudych włosów („czynnika" recesywnego dla rudych włosów, jak powiedziałby Mendel) wiąże się z relacjami tej osoby do: koloru włosów, innych genów (lub czynników), rodziców i empirycznie możliwych potomków. Teoria Mendla określa dokładnie, na czym ta relacyjna własność polega. Pierwszą rzeczą, na którą należy zwrócić uwagę jest to, że zadowalająca odpowiedź na pytanie: „Jak człowiek może posiadać recesywny gen dla rudych włosów?" musi się odwołać do czegoś posiadającego własności relacyjne, które z kolei są wystarczające do wytłumaczenia relacyjnych własności konstytutywnych dla posiadania genu recesywnego dla rudych włosów. Utożsamienie genu z pewną sekwencją na molekule DNA tylko dlatego jest dobrą odpowiedzią na pytanie „jak", że przyjmuje się istnienie pewnego normalnego dla funkcjonowania DNA środowiska chemicznego, a mianowicie środowiska chemicznego, w którym pewna sekwencja wywrze, w odpowiednich okolicznościach, określony wpływ przyczynowy na rozwój człowieka o rudych włosach.

Generalnie odpowiedź na pytanie „jak" dotyczące jakiejś własności relacyjnej obiektu tylko wtedy będzie zadowalająca, gdy odniesie się ona na jakimś etapie bądź do stanów o podobnie relacyjnym charakterze, bądź do stanów, co do których można założyć, że posiadają pewne własności relacyjne. Jeśli stany wymienione w psychologii subpersonalnej mają dać odpowiedź na pytania „jak" dotyczące własności treściowych, to stany wyjaśniające muszą pozostawać w pewnych relacjach eksternalnych. Próba odpowiedzi na pytanie „jak" dotyczące stanów relacyjnych, która by nie wymieniała stanów relacyjnych, byłaby w najlepszym razie niepełna; a właśnie to było niepokojące w niesemantycznej koncepcji obliczania, o której była mowa w części 1. Proponowane wyjaśnienie cechy dziedzicznej odwołujące się do własności molekuł DNA byłoby dalece niezadowalające, gdyby nie zachodziły owe zakładane normalne chemiczne warunki środowiskowe. Jaki bowiem, bez tych normalnych warunków, istniałby związek między własnościami DNA a własnościami organizmu, który się w rezultacie rozwinął?

Sławne trzy poziomy opisu zaproponowane przez Marra także można uznać za zawierające rozróżnienie między tym, co podlega obliczaniu a sposobem obliczania (Marr, 1982, rozdz. 1). Prawidłowy opis algorytmu, na drugim poziomie u Marra, stanowi odpowiedź na następujące pytanie „jak": „Jak obliczana jest funkcja wyszczególniona na poziomie pierwszym?" W ciągu ostatnich kilku lat zaproponowana przez Marra koncepcja trzech poziomów podlegała rozmaitym atakom, lecz żaden z tych, które znam nie podważył rozróżnienia między obliczaną funkcją a stosowanym do jej obliczenia algorytmem. Zwłaszcza nie należy sądzić, że twierdzenie o jedno-wieloznacznej relacji między funkcją obliczaną (poziom 1) a algorytmem, za pomocą którego dokonywane jest obliczenie (poziom 2), implikuje tezę epistemologiczną, iż zawsze możemy odkryć, bez zbadania algorytmu z poziomu drugiego, która funkcja podlega obliczaniu. Patricia Churchland i Terrence Sejnowski (1990, s. 368) piszą:

> Zdaniem Marra, poziom wyższy jest niezależny od poziomów względem niego niższych i dlatego problemy obliczeniowe można analizować niezależnie od rozumienia algorytmu, za pomocą którego dokonuje się obliczenia, a problem algorytmiczny można rozwiązać niezależnie od rozumienia fizycznej implementacji.

Są dowody na to, że krytykowany przez Churchland i Sejnowskiego pogląd rzeczywiście był głoszony przez Marra. Marr faktycznie napisał, że „zapewne łatwiej przyjdzie nam zrozumieć algorytm rozumiejąc naturę rozwiązywanego problemu aniżeli badając mechanizm (i hardware), w którym jest on osadzony" (Marr, 1982, s. 27). Jednakże Marr mógł i powinien był podkreślać wagę wszystkich trzech poziomów opisu bez opowiadania się za tezą epistemologiczną, którą przyjmuje w wypowiedziach takich, jak wyżej zacytowana. Akceptacja trzech poziomów

[5] Bardzo klarowne omówienie wstępne elementów genetyki mendlowskiej istotnych dla obecnego wywodu zawarte jest w Kitcher, 1982.

opisu nie zobowiązuje do akceptacji przejścia od nieepistemicznego do epistemicznego sensu „niezależności". W przypadku konkretnego procesu odkrycie prawidłowego opisu na poziomie 1, bez wcześniejszego zbadania obu właściwych dla tego procesu poziomów niższych, może się okazać niemożliwe. Tak czy inaczej, kiedy już opisy te zostaną odkryte, można powiedzieć, że zarówno algorytm, jak i jego implementacja odpowiedzą na pytanie, jak system oblicza funkcję uchwyconą w charakterystyce z poziomu 1. Nie wydaje się to być bardziej problematyczne niż to, że ktoś nie odkrył, jaką funkcję pełnią nerki zanim nie zbadał przeprowadzanych przez nerki szczegółowych procesów chemicznych[6].

O ile dobrze sformułowane pytanie musi zawierać operatywny opis własności, która ma być wyjaśniona, o tyle opis może być operatywny nawet wtedy, gdy nie dostarcza maksymalnie pełnej analizy tego, co to znaczy posiadać daną własność. Spełnieniem jednego marzenia byłaby sytuacja, w której dla każdego typu treściowego stanu mentalnego posiadalibyśmy pełną, konstytutywną teorię określającą, jakie własności i relacje powinien posiadać podmiot, aby się znajdować w danym stanie (pełna, konstytutywna teoria byłaby teorią wyprowadzalną z bogatych i zupełnych teorii typu stanu i tożsamości wchodzących w grę treści). Spełnieniem kolejnego marzenia byłaby sytuacja, w której udałoby się nam podać subpersonalne wyjaśnienie psychologiczne tego, jak organizm może posiadać własności i relacje zidentyfikowane przez ową pełną, konstytutywną teorię. Ale i bez osiągnięcia takiej zupełności możliwy jest eksplanacyjny postęp. Faktycznie nie jest nawet pewne, czy osiągnięcie takiej zupełności ma być możliwe, nawet jako idea regulatywna. Już sama charakterystyka rodzajów treści związanych z danym stanem mentalnym, wraz z określeniem niektórych ich wzajemnych relacji oraz relacji do rozmaitych innych stanów, jest czymś, wobec czego można postawić dobrze zdefiniowane pytanie „jak" i uzyskać na nie odpowiedź. Wydaje się na przykład oczywiste, że postęp w psychologii percepcji odbywa się dzięki teoriom, które z góry zakładają, iż stany percepcyjne posiadają treści określonych rodzajów i przechodzą do udzielenia odpowiedzi na rozmaite pytania empiryczne, w tym na pytania „jak", które pojawiają się w przypadku tych stanów. To wszystko jest możliwe bez pełnej, konstytutywnej, filozoficznej koncepcji treści percepcyjnej. Nawet jeśli nie ma niczego takiego, jak pełna koncepcja, to nadal jest rzeczą pożądaną, by dla każdej konstytutywnej cechy treści posiadać wyjaśnienie tego, jak organizmy danego rodzaju zdolne są wchodzić w wymagane przez nią relacje.

Odpowiem teraz na zarzut, że to, co powiedziałem jest niespójne. Podkreślałem, że w wyjaśnieniu obliczeniowym ze względu na treść stany eksternalistyczne wyjaśnia się za pomocą stanów eksternalistycznych i powiedziałem co nieco na temat cech charakterystycznych takiego wyjaśnienia. Jako przykłady treściowego wyjaśnienia obliczeniowego podałem obliczeniowe wyjaśnienia doświadczenia percepcyjnego, zwracając uwagę, że wyjaśnienia te w przypadku percepcji wzrokowej wychodzą od własności obrazu na siatkówce. Ale stany siatkówki nie są stanami eksternalistycznymi; co więcej, podobne uwagi odnoszą się też do innych obliczeniowych wyjaśnień percepcji (i rozumienia języka). Wydaje się zatem, że we wszystkich tych przypadkach będziemy mieli do czynienia z pierwszym, eksternalnie zindywidualizowanym stanem treściowym, który nie jest wyjaśniany przez stan eksternalistyczny. Toteż bądź fałszywa jest zasada, że stany eksternalistycznie należy wyjaśniać za pomocą stanów eksternalistycznych, bądź, wbrew pozorom, nawet najwcześniejsze stany w tego typu wyjaśnieniach należy traktować jako stany eksternalnie zindywidualizowane. (Ta druga alternatywa przypomina zalążkowe stanowisko gibsonowskie). Podobne uwagi odnoszą się, *pari passu*, do działań.

Niemniej jednak na zarzut ten jest lepsza odpowiedź, odpowiedź mająca pewną motywację teoretyczną. Każde treściowe wyjaśnienie obliczeniowe zakłada jakieś normalne środowisko

[6] Powinienem jednak zaznaczyć, że w pełni zgadzam się z Churchland i Sejnowskim, że „idea jakoby w istocie istniał jeden poziom implementacyjny jest uproszczeniem" (s. 369). Kwestionowano również, czy zawsze musi istnieć możliwość skonstruowania opisu procesu obliczania na marrowskim poziomie 1 (zob. Boden, 1988, s. 227). Kiedy jednak interesuje nas obliczanie treściowe, to jest prawdą konieczną, iż tam gdzie mamy do czynienia z obliczaniem funkcji treściowej, tam też istnieje funkcja w ramach ekstensji, która jest obliczaną funkcją treściową. Wszędzie gdzie mamy do czynienia z wyjaśnieniem obliczeniowym, pojawi się pozwalające na poprawną odpowiedź pytanie „jak", na które odpowiedzią jest prawidłowe sformułowanie związanego z tym procesem algorytmu.

dla organizmu posiadającego stany obliczeniowe (a przynajmniej zakłada jego normalność pod pewnymi względami). Było od dawna dostrzegane i podkreślane przez obrońców eksternalizmu w kwestii treści (zob. Davies, 1986). Trudno uniknąć takiego założenia, jeśli ostatecznym celem treściowego opisu obliczeniowego jest przyczynienie się do wyjaśnienia tego, że organizm znajduje się w stanach treściowych psychologii postaw propozycjonalnych. Istnienie takiego samego normalnego środowiska zakłada też psychologia postaw propozycjonalnych. O ile organizm znajduje się w założonym normalnym środowisku, o tyle w przypadku percepcji wzrokowej pierwszy treściowy, eksternalnie zindywidualizowany stan wytworzony przez stany siatkówki będzie pozostawał w określonych relacjach środowiskowych. Tak więc to wyjaśnienie nie wymaga w żaden sposób, aby ów stan treściowy pojawił się bez względu na to, jakie jest normalne środowisko organizmu. Analogicznie, stan treściowy będzie posiadał taką własność, że w normalnym środowisku organizmu będzie on wyjaśniany przez środowiskowe stany rzeczy oraz, przy odpowiednich okolicznościach dodatkowych, będzie on mógł te stany rzeczy wyjaśniać. Jak zawsze, i zgodnie z zasadami przypisywania treści subpersonalnej przedstawionymi w części 3, poprawna treść przypisywana pierwszemu stanowi treściowemu będzie zależała nie tylko od tego, jakie warunki środowiskowe ten stan wywołały, lecz także od natury innych procesów, którym daje on lub może dawać początek oraz od skutków środowiskowych tych procesów.

Można by na to odpowiedzieć, że przy założeniu o normalnym środowisku, łańcuch stanów całkowicie internalistycznych będzie mógł w pewnym sensie wyjaśnić wystąpienie eksternalnie zindywidualizowanego, treściowego stanu psychologii postaw propozycjonalnych. Należy się oczywiście zgodzić, że pewne zarzuty modalne, na które stanowisko takie byłoby inaczej narażone, zostają zneutralizowane przez założenie o normalnym środowisku. Jednakże wyjaśnienie za pomocą takiej sekwencji stanów internalistycznych byłoby także o wiele mniej ogólne niż wyjaśnienie za pomocą stanów treściowych. Treściowe wyjaśnienie obliczeniowe posiada własność znaną również z innych przypadków, która polega na tym, że stosuje się ono do rozmaitych, posiadanych przez odmienne organizmy stanów internalistycznych. Na tym zresztą polega zapowiadana różnica między treściowym wyjaśnieniem obliczeniowym a wyjaśnieniami genetycznymi opartymi na DNA. Tych drugich nie można stosować bez względu na biochemiczną strukturę organizmu. I choć (jak mniemam) mogłaby w zasadzie istnieć genetyka bardziej abstrakcyjna, posługująca się pojęciem „treści instrukcyjnej" genu i nie odwołująca się do jego określonego składu chemicznego, to jednak wyjaśnienia w sposób charakterystyczny odwołujące się do DNA funkcjonują de facto na niższym niż ten poziomie.

Dotychczas rozważaliśmy, jak dalece treściowe wyjaśnienia obliczeniowe są w stanie odpowiedzieć na pytania o to, jak to się dzieje, że organizm znajduje się w określonym, eksternalnie zindywidualizowanym stanie. Ale treściowe wyjaśnienie obliczeniowe może się też przyczynić w sposób zasadniczy do odpowiedzi na inny rodzaj pytań „jak". Pytania tego drugiego rodzaju mają następującą postać: „Jak taki to a taki stan reprezentuje rzetelnie i poprawnie (lub nawet niepoprawnie) w takich to a takich okolicznościach, w ramach normalnego dla danego organizmu środowiska?" Tutaj wyjaśnieniu podlega własność poprawności (lub niepoprawności), która jest relacyjną własnością stanu relacyjnego. Obliczenia treściowe z natury dobrze nadają się do tego, by dostarczać takich wyjaśnień. Załóżmy, że każde z treściowych obliczeń cząstkowych jakiegoś procesu stanowi poprawną metodę obliczenia wielkości (lub własności, lub czegokolwiek), którą ten krok cząstkowy oblicza na podstawie dowolnych wartości wejściowych. W takim przypadku w ostatnim kroku procesu wielkość (lub własność, lub cokolwiek) obliczona zostanie w sposób poprawny. Analogicznie, jeśli któryś krok jest niepoprawny, to wiemy dlaczego, w określonych okolicznościach, system wygeneruje niepoprawny pod względem treści stan końcowy (pomijając sytuacje wzajemnego znoszenia się błędów). Rozumienie, jakiego takie obliczenia treściowe mogą dostarczyć w kwestii tego, dlaczego wynik jest poprawny lub niepoprawny jest specyficzne dla wyjaśnienia odwołującego się do stanów treściowych, gdyż opiera się w sposób istotny na treściowym charakterze poszczególnych obliczeń cząstkowych. Opis obliczenia, który pomijałby treść stanów obliczeniowych, nigdy nie doprowadziłby do takiego rozumienia. Dokładnie to samo można powiedzieć o semantycznych opisach obliczeń cząstkowych w ramach złożonego algorytmu arytmetycznego.

Tylko dysponując opisami semantycznymi możemy wyjaśnić, dlaczego dostarcza on w sposób rzetelny poprawnej odpowiedzi na pytania dotyczące wartości pewnej funkcji arytmetycznej.

5. OBIEKTYWNOŚĆ, NIEREDUKOWALNOŚĆ I WYJAŚNIENIE

Przyjmuje się powszechnie, że proces może być obliczaniem tylko wtedy, gdy ma jakiś prawdziwy opis treściowy. Możliwość obliczania, w którym nie ma mowy o tym, co jest obliczane i na jakiej podstawie, jest dla nas niezrozumiała. Ten powszechnie akceptowany pogląd dobrze wyrazili Churchland i Sejnowski: „w najogólniejszym sensie, system fizyczny może być uważany za system obliczeniowy tylko wtedy, gdy jego stany fizyczne można rozpatrywać jako reprezentacje stanów pewnych innych systemów i gdzie przejścia między jego stanami da się wyjaśnić jako operacje na reprezentacjach" (Churchland i Sejnowski, 1992, s. 62).

Ten wymiar semantyczny nie jest bynajmniej trywialnym, definicyjnym dodatkiem. Przeciwnie, powodem jego wprowadzenia był pierwotny sens i cel pojęcia obliczania. Choć Turing opisywał swoje maszyny w kategoriach czysto formalnych i mechanicznych, miały one znaczenie dla teorii obliczalności właśnie dlatego, że można przyjąć, iż symbol „1" odnosi się do liczby 1, zaś zestawienie na taśmie maszyny posiada pewien sens. Rozpatrując każdą z maszyn Turinga jako źródło funkcji arytmetycznej, opieramy się na wymiarze semantycznym. Z wyników dotyczących maszyn Turinga możemy wyciągać wnioski w sprawie funkcji arytmetycznych w obrębie liczb naturalnych tylko dzięki temu wymiarowi referencjalnemu. Nie jest prawdą jakoby pierwotne, kluczowe pojęcie obliczania miało charakter czysto syntaktyczny.

Ale zgoda na to, że każde obliczanie musi mieć jakiś prawdziwy opis treściowy nie przesądza jeszcze o obiektywności, nieredukowalności czy wadze eksplanacyjnej treściowego poziomu opisu. Posiadanie każdej z tych własności przez treściowe obliczenia było kwestionowane w taki czy inny sposób. Przejdę teraz do omówienia tych spraw.

Churchland i Sejnowski sami uważają, że przyjmowana charakterystyka obliczania łączy się z tezą, że to, czy coś jest obliczaniem, czy też komputerem, zależy od naszych zainteresowań. „Uznajemy coś za komputer, ponieważ, i tylko wtedy gdy, jego wejścia i wyjścia można użytecznie i systematycznie interpretować jako reprezentujące uporządkowane pary pewnej funkcji, która nas interesuje" (tamże, s. 65). Ze względu na to, co nazywają oni „wymiarem zrelatywizowanym do zainteresowań", obliczanie nie jest, ich zdaniem, rodzajem naturalnym[7]. Pogląd, który tutaj przedstawiam jest diametralnie różny. Przypisywanie treści subpersonalnych podlega środowiskowym kontrfaktycznym okresom warunkowym. Na poziomie najbardziej podstawowym, askrypcje te podlegają prawdziwości rozmaitych kontrfaktycznych okresów warunkowych, zawierających to, co w części 2 nazwałem pierwotnymi eksplanandami i eksplanansami. Prawdziwość askrypcji treści subpersonalnej jakiemuś stanowi o tyle mogłaby być zrelatywizowana do zainteresowań, o ile prawdziwość tychże kontrfaktycznych okresów warunkowych byłaby zrelatywizowana do zainteresowań. W przypadku eksperymentu ze słyszeniem dychotycznym, jeden z potwierdzanych okresów warunkowych miałby postać: „Gdyby w kanale, na którym nie jest skupiona uwaga usłyszane zostały różne słowa o tym samym znaczeniu, osoba badana tak samo zinterpretowałaby zdanie usłyszane w kanale, na którym skupiona jest uwaga". Nie rozumiem, w jaki sposób prawdziwość tego kontrfaktycznego okresu warunkowego miałaby być zrelatywizowana do zainteresowań. To samo dotyczy leżącej u podstaw owego okresu warunkowego relacji wyjaśniającej, której wskaźnikiem jest jego prawdziwość.

Dlaczego twierdzenie z zależności obliczania od zainteresowań jest tak kuszące. Być może wypływa to z następującej myśli: „nawet przetaki lub młockarnie mogłyby być komputerami, ponieważ każde z tych urządzeń kategoryzuje dane wejściowe i gdyby ktoś miał czasu w nadmiarze, mógłby odkryć funkcję opisującą zachowanie wejście-wyjście" (tamże, s. 66). Jednak w obu tych przykładach pojęcie treści reprezentacyjnej niczego nie wyjaśnia. Jeśli w przypadku przetaka lub

[7] Ściśle mówiąc, powiadają oni, że komputery nie tworzą rodzaju naturalnego. Gdyby jednak nie dało się tego twierdzenia rozciągnąć na zjawisko obliczania, jego znaczenie byłoby ograniczone. Kontekst tego twierdzenia pojawiającego się w ogólnej dyskusji na temat obliczania w *The Computational Brain* (1992) sugeruje bez wątpienia, że przypisuje się mu znaczenie ogólniejsze.

młockarni powiemy, że (na przykład) tego rodzaju kształty i wielkości reprezentują same siebie, to przypisując te treści przestrzenne i traktując procesy je uwzględniające jako obliczenia, z natury rzeczy nie zwiększylibyśmy siły wyjaśniającej, w sposób, który nie byłby możliwy bez takiego rzekomego przypisania treści[8]. To prawda, że właściwie o każdym konkretnym przedmiocie można powiedzieć, że realizuje on taką bądź inną funkcję typu wejście – wyjście. Fakt ten w niczym jednak nie zagraża koncepcji obliczania jako teoretycznie ważnego rodzaju ogólnego, ponieważ w przypadkach, które pragniemy wykluczyć, przypisywanie treści jest albo zabiegiem eksplanacyjnie redundantnym, albo zbyt daleko idącym. W autentycznym wyjaśnianiu obliczeniowym ani jedna, ani druga sytuacja nie ma miejsca[9].

Stanowisko, które bronię przemawia na rzecz nieredukowalności treściowego, subpersonalnego wyjaśniania obliczeniowego do wyjaśniania neurofizjologicznego. Wstępny argument za nieredukowalnością polega na tym, że wyjaśnienia obliczeniowe mogą wchodzić w skład wyjaśnienia zawierającego relacyjne eksplanandum dotyczące relacji podmiotu lub systemu do środowiska. Wyjaśnienia neurofizjologiczne wchodzą w skład wyjaśnień zawierających eksplananda dotyczące własności cielesnych lub ruchów, a nie ich relacji do środowiska. Nieredukowalność wyjaśnień obliczeniowych do wyjaśnień neurofizjologicznych bynajmniej nie kłóci się z tezą, że istnieje jakiegoś rodzaju zależność między wyjaśnieniami obliczeniowymi a neurofizjologicznymi. Te z proponowanych wyjaśnień obliczeniowych faktów dotyczących człowieka, które nie mają podstaw w neurofizjologii, są nie do przyjęcia. Jeśli jednak eksplananda obliczeniowe różnią się od neurofizjologicznych, to nic nie wskazuje na to, by przypadek ten dało się upodobnić do takiej redukcji makro – mikro, której przykładem jest redukcja praw opisujących zachowanie gazów do mechaniki cząstek. Zjawiska wyjaśniane przez prawa opisujące zachowanie gazów – konkretne przypadki ogólnej relacji między ciśnieniem, objętością i temperaturą – dają się również wyjaśnić przez teorię redukującą (plus twierdzenia identycznościowe czy „prawa łączące"). Ale nawet opis jakiegoś stanu jako czegoś, co reprezentuje umiejscowienie czegoś w układzie współrzędnych, którego centrum jest głowa – ten rodzaj treści reprezentacyjnej, który stanowi chleb powszedni dla podejścia obliczeniowego – jest opisem środowiskowym[10]. Analogicznie, obliczeniowe wyjaśnienie faktu, iż dany system znajduje się w takim to a takim stanie może wchodzić w skład wyjaśnienia szerszego, z eksplanandum mówiącym, że podmiot sięga w pewnym zrelatywizowanym do głowy kierunku (by, na przykład, wyłączyć urządzenie emitujące dźwięk).

Czy odwołuję się tu do nadmiernie wygórowanego standardu redukcji? Na pewno przedstawione przed chwilą rozumowanie wymaga rozwinięcia, ponieważ faktycznie istnieją inne dobre przypadki redukcji, w których nie jest wymagana ścisła identyczność eksplanandów teorii podlegającej redukcji z niektórymi eksplanandami teorii redukującej. Eksplananda mechaniki Newtona nie zawierają żadnej relatywizacji do jakiegoś układu odniesienia, gdy dotyczą relacji czasowych, natomiast taka relatywizacja jest stałym elementem eksplanandów szczególnej teorii względności.

[8] W przytoczonym przez Churchland i Sejnowskiego przykładzie suwaka logarytmicznego, przyrząd ten traktowany jest jako komputer ze względu na ludzką intencję używania go w określony sposób.

[9] Przynajmniej w tym punkcie zgadzam się całkowicie z Churchland i Sejnowskim. „W przeciwieństwie do systemów, które zazwyczaj nazywamy komputerami, *modus operandi* niektórych urządzeń jest taki, że wystarczy podać wyjaśnienie czysto przyczynowe, bez odwołania się do tego, że coś jest obliczane lub reprezentowane" (s. 68).

[10] Stanowisko Churchland i Sejnowskiego często jawi się jako redukcjonistyczne: autorzy ci jako „negatywną wersję" hipotezy roboczej swojej książki podają hipotezę, że „jest rzeczą wysoce nieprawdopodobną, iż własności emergentne są własnościami, których nie można wyjaśnić za pomocą własności z niższego poziomu [...]; albo że są one w jakimś sensie nieredukowalne, przyczynowo *sui generis*, albo, jak zwykli mawiać filozofowie, «nomologicznie autonomiczne», czyli, z grubsza biorąc, «nie będące częścią reszty nauki»" (s. 2). Z drugiej jednak strony, w swoim omówieniu konkretnych przykładów wysuwają oni często twierdzenia, które akcentują właśnie zwolennicy nieredukowalności. I tak o roli neuronów, których zachowanie da się wyjaśnić jako obliczanie współrzędnych, których centrum jest głowa, na podstawie położenia bodźca na siatkówce i umiejscowienia gałki ocznej w głowie, piszą oni: „Wiedza o tym, że niektóre neurony posiadają taki profil reakcji, który powoduje, iż inne neurony reagują w pewien sposób, może być przydatna [...] ale sama w sobie niewiele nam mówi o roli, jaka odgrywają te neurony w posiadanej przez zwierzę zdolności widzenia. Musimy ponadto wiedzieć, co reprezentują rozmaite stany tych neuronów i jak takie reprezentacje mogą poprzez interakcje neuronalne być przekształcane w inne reprezentacje" (s. 68).

A jednak szczególna teoria względności wyjaśnia przybliżoną poprawność praw Newtona dotyczących obiektów poruszających się z prędkością o wiele mniejszą niż prędkość światła. Pojawia się zatem pytanie: czy istnieją takie parametry, po ustaleniu których, albo po odpowiednim zrelatywizowaniu których, można prawomocnie powiedzieć, iż wyjaśnienie obliczeniowe redukuje się do neurofizjologicznego?

Do określonego neurofizjologicznego wyjaśnienia ruchu ciała lub własności cielesnej moglibyśmy dodać dwie rzeczy. Pierwsze uzupełnienie polegałoby na określeniu ogólnych zasad przypisywania treści stanom obliczeniowym, o ile można je wyraźnie sformułować. Uzupełnienie drugie polegałoby na określeniu, dla każdego ze stanów wymienionych w danym wyjaśnieniu neurofizjologicznym, ich odpowiednich relacji środowiskowych oraz relacji, w jakich pozostają one do innych odpowiednich stanów. Dokonawszy tego podwójnego uzupełnienia, moglibyśmy przypuszczalnie wyprowadzić eksplanandum pozostające w takiej samej relacji do środowiska, co konkluzja wyjaśnienia obliczeniowego. Takie wyprowadzenia pozwalają odwzorować wyjaśnienia neurofizjologiczne w wyjaśnienia obliczeniowe, korzystając z zasad przypisywania treści obliczeniowych stanom subpersonalnym oraz z relacyjnych, środowiskowych własności określonych stanów neuronalnych. Czy dołączenie zasad przypisywania można uznać za analogiczne do „reguł przypisywania ciśnienia i temperatury", wiążących ciśnienie i temperaturę z wielkościami molekularnymi? W końcu przecież każda skuteczna redukcja będzie miała jakąś postać praw łączących. Czy zatem dołączenie reguł przypisywania można upodobnić do owego bliższego nam zjawiska?

Fundamentalna różnica polega na tym, że reguły przypisywania treści stanom obliczeniowym wiążą te treści nie (lub nie tylko) ze stanami neurofizjologicznymi, tj. stanami teorii rzekomo redukującej, lecz także ze sprawami środowiskowymi. Dlatego właśnie przypadek ten nie jest podobny do innych przypadków redukcji. Również dlatego potrzebne jest uzupełnienie drugie, które zawiera określenie pewnych relacji środowiskowych do poszczególnych stanów neurofizjologicznych wymienionych w danym wyjaśnieniu. To drugie uzupełnienie nie byłoby konieczne, gdyby zasady przypisywania treści odwoływały się jedynie do spraw neurofizjologicznych. W istnieniu dla każdego wyjaśnienia obliczeniowego takiego relacyjnie uzupełnionego, neurofizjologicznego wyjaśnienia lepiej byłoby widzieć rozwinięcie tego, na czym polega nieredukcyjna zależność wyjaśnienia obliczeniowego od tego, co neurofizjologiczne. W rzeczywistości, po wprowadzeniu obu uzupełnień staje się jasne, że dla eksplanandum stanu obliczeniowego nie ma znaczenia identyczność określonego stanu neurofizjologicznego. Każdy inny stan neurofizjologiczny byłby równie dobry, pod warunkiem że pozostawały w relacjach odpowiednich do tego, by posiadać tę samą treść.

Istnieje i inne stanowisko, takie, które nie musi się odwoływać do redukcjonizmu neurofizjologicznego, lecz mimo to stawia pod znakiem zapytania doniosłość treściowej koncepcji obliczania. Stanowisko to wyrażają dwie tezy: (a) te same korzyści, które płyną z wyjaśnienia obliczeniowego odwołującego się do treści eksternalistycznych można osiągnąć dołączając do wyjaśnienia obliczeniowego odpowiednie uzupełnienia nie odwołujące się do treści eksternalistycznych; (b) takie podejście należy preferować, ponieważ prowadzi ono do celu wyjaśnień obliczeniowych, jakim jest odsłonięcie mechanizmów poznania. Pogląd ten nie jest zobowiązany do przyjęcia redukcjonizmu neurofizjologicznego, ponieważ może on utrzymywać, że istnieje pewien rodzaj treści, która nie ma charakteru eksternalistycznego, a która znacznie wykracza ponad poziom neurofizjologiczny. Ten rodzaj treści może znaleźć zastosowanie w wyjaśnieniach obliczeniowych, które – jak głosi twierdzenie (a) – dają się odpowiednio uzupełnić.

Pogląd, na który składają się tezy (a) i (b) jest zbliżony do stanowiska bronionego przez Frances Egan. Pisze ona (1992, s. 447), że:

> wyjaśnienie interakcji organizmu ze środowiskiem nie jest głównym celem teorii obliczeniowych; takie wyjaśnienia pojawiają się dopiero wtedy, gdy daną teorię obliczeniową uzupełni się o dalsze założenia dotyczące normalnego środowiska, w którym opisane mechanizmy poznawcze są rozwijane.

Egan pisze również, że jej podejście „zależy częściowo od poglądu, że celem takich teorii jest charakterystyka mechanizmów leżących u pod-

łoża naszych rozmaitych zdolności poznawczych, a ponadto, że do tego celu najodpowiedniejsze są teorie, w których taksonomizuje się stany indywidualistycznie" (ss. 444-5). Być może stanowisko Egan jest jedynie zbliżone, a nie identyczne, do stanowiska, na które składają się tezy (a) i (b), ponieważ zamierzeń zwolenników obliczania treściowego nie można w pełni opisać jako „wyjaśnienie interakcji organizmu ze środowiskiem". Być może uzupełnienie ułatwi wyjaśnienie tej interakcji, ale tak naprawdę chodziło o coś innego – o wyjaśnienie tego, że organizm znajduje się w stanach, których indywidualizacja zakłada relacje do środowiska. Niemniej jednak linia rozumowania jest bardzo zbliżona i chciałbym rozważyć, na ile racje przedłożone przez Egan można wykorzystać do podważenia rozwijanych przeze mnie argumentów. Zajmę się po kolei tezą (a), a potem (b).

Zarzut, że korzyści eksplanacyjne płynące z treściowych obliczeń eksternalistycznych można osiągnąć po prostu uzupełniając opis obliczania, który nie ma charakteru eksternalistycznego, jest w ramach tej dziedziny subpersonalnej dokładnym odpowiednikiem odpowiedzi na argumenty eksternalistyczne w odniesieniu do stanów psychologii ludowej. Kiedy zwolennik eksternalizmu w kwestii stanów opisywanych przez psychologię ludową argumentuje, że stany te wyjaśniają działania jedynie o tyle, o ile scharakteryzowane są w kategoriach środowiskowych, internaliści skłaniają się do odpowiedzi, że jedynymi stanami naprawdę eksplanacyjnymi są stany eksternalistyczne. Twierdzą oni, że stany internalistyczne wyjaśniają indywidualistycznie opisane ruchy ciała. Aby zatem dojść do wyjaśnienia relacyjnych własności działania, przy których obstaje eksternalista, wystarczy uzupełnić indywidualistyczny opis działania informacjami o okolicznościach środowiskowych tego działania[11].

Replika na odpowiedź, jakoby wystarczyło dokonać uzupełnienia, by sprostać wchodzącej tu w grę potrzebie, jest taka w psychologii subpersonalnej, jak w psychologii ludowej z poziomu personalnego. Została on już podana wcześniej: przytaczając wyjaśnienie jakiegoś zdarzenia przy jednym z jego opisów i uzupełniając owo wyjaśnienie informacją, że to samo zdarzenie podpada pod inny opis, nie uzyskuje się tym samym wyjaśnienia tego zdarzenia przy tym innym opisie. Rozważmy na przykład wyjaśnienie obliczeniowe tego, że organizm znajduje się w stanie reprezentującym znajdujący się przed nim obiekt materialny jako mający określony kształt i kolor. W kontekście innych postaw podmiotu, gdy wszystko przebiega dobrze, to wyjaśnienie obliczeniowe może pomóc w obszerniejszym wyjaśnieniu, dlaczego podmiot ujmuje znajdujący się przed nim obiekt o owym kształcie i kolorze. Gdyby ten sam obiekt był tysięcznym przedmiotem schodzącym z fabrycznej taśmy produkcyjnej, nie będziemy tak samo dysponowali obszerniejszym wyjaśnieniem, którego częścią jest historia obliczeniowa, tj. obszerniejszym wyjaśnieniem następującego faktu: tego, że podmiot uchwycił tysięczny przedmiot schodzący z taśmy produkcyjnej (tutaj oczywiście deskrypcja określona „tysięczny przedmiot" na węższy zasięg niż zawierające tę deskrypcję zdanie rozpoczynające się od „że"). W szczególności, nie zostały potwierdzone konieczne kontrfaktyczne okresy warunkowe. Nie jest prawdą, że gdyby jakiś inny obiekt, być może znajdujący się w innym położeniu w stosunku do podmiotu, był tysięcznym przedmiotem schodzącym z taśmy, to podmiot uchwyciłby ten przedmiot. Uchwyciłby natomiast jakiś inny przedmiot, być może znajdujący się w innym położeniu w stosunku do niego, gdyby posiadał on ten sam kształt i kolor[12].

Część (b) rozważanego tu konkurencyjnego stanowiska stwierdza, że tylko internalnie zindywidualizowane stany obliczeniowe są w stanie osiągnąć cel, jakim jest identyfikacja mechanizmów leżących u podstaw naszych rozmaitych zdolności poznawczych. Jeśli jakiś mechanizm z definicji traktowany jest jako nie zindywidualizowany w kategoriach treściowych, to wówczas bezpośrednią prawdą aprioryczną jest to, że treściowe obliczenia nie są mechanizmami. W pełni popieram ideę, że powinniśmy dążyć do

[11] Zob. omówienie tej odpowiedzi w *Externalist Explanation* (Peacocke, 1993, ss. 208-209).

[12] „Ale załóżmy, że rozważamy własność nomologicznie koekstensywną z koniunkcją kształtu i koloru. W takiej sytuacji przytoczony właśnie argument nie ma zastosowania". Zgoda – w tym wypadku powinniśmy się raczej odwołać do zasad rządzących przypisywaniem treści subpersonalnych, omówionych w części 3. Sprawdzian kontrfaktyczny jest tylko przybliżoną wskazówką, że spełnione są wymogi narzucane przez te zasady.

zidentyfikowania beztreściowych mechanizmów umożliwiających poznanie. Nie wynika z tego jednak, że eksternalistyczne obliczenia treściowe nie odgrywają także charakterystycznej tylko dla nich roli eksplanacyjnej. Nie podważa to również wcześniejszych argumentów na rzecz tezy, że czysto indywidualistyczne stany nie mogą odgrywać tych ról.

Gdyby termin „wyjaśnienie mechaniczne" ograniczyć do stanów nie określanych jako treściowe, to wyjaśnienie treściowe i tak mogłoby być wyjaśnieniem mechanicznym w innym, ważnym sensie. W potocznym, intencjonalnym wyjaśnianiu z poziomu personalnego, które dotyczy działań lub stanów mentalnych człowieka, zakładamy cały szereg zdolności poznawczych człowieka. Zakładamy jego zdolność spostrzegania, rozumowania, zapamiętywania, rozumienia, by wymienić tylko kilka. Nawet jeśli wyjaśnienie subpersonalne odwołuje się do pojęcia treści – co, jak usiłowałem dowieść, jest nieodzowne – nie zakłada ono tych zdolności poznawczych z poziomu personalnego. Stara się je raczej wyjaśnić. Psycholog obliczeniowy na pewno nie zakłada zwyczajnych zdolności percepcyjnych, kiedy dowodzi, na przykład, że fenomeny grupowania percepcyjnego można wyjaśnić obliczając minimalne inwariantności translacyjne w prezentowanej scenie. Zdolność jakiegoś mechanizmu subpersonalnego do przeprowadzenia tego typu obliczenia zostanie wyjaśniona w kategoriach działania pewnego algorytmu. Na całe to przedsięwzięcie nałożony zostanie, lub powinien zostać nałożony, warunek, według którego byłoby rzeczą całkiem nieuprawnioną, gdyby w jakimś momencie wyjaśnienie opierało się na tym, że system subpersonalny „zauważył pewne zgrupowanie". Istnieje zatem pośredni obszar między, z jednej strony, wyjaśnieniami intencjonalnymi z poziomu personalnego, w których zakłada się zdolności poznawcze, a czysto mechanicznymi wyjaśnieniami beztreściowymi, z drugiej strony. To właśnie na tym pośrednim obszarze funkcjonują treściowe wyjaśnienia obliczeniowe. Zawężenie terminu „mechaniczny" do wyjaśnień beztreściowych nie spowoduje zniknięcia tego pośredniego obszaru.

Mimo tych wszystkich argumentów nadal można odnosić silne wrażenie, że całkowita historia przyczynowa nadal sprowadza się do tej, którą zawiera syntaktyczny opis procesu obliczeniowego i że z natury rzeczy proces syntaktyczny pomija jakiekolwiek własności semantyczne. Wydaje mi się, że wrażenie to wypływa z nieprzemyślanego użycia pojęcia „całkowitej historii przyczynowej". Jeśli weźmiemy pod uwagę sekwencję określonych zdarzeń, w rodzaju wręczania kilku monet w zamian za produkt, to trudno jest twierdzić, że znajduje tu swe zastosowanie pojęcie jedynej historii przyczynowej opisującej tę sekwencję. Sekwencja ta ma opis fizyczny, opis intencjonalny, opis ekonomiczny i wiele innych. Może będziemy w stanie nadać sens pojęciu jedynej historii przyczynowej, jeśli zrelatywizujemy je do opisów jakiegoś określonego rodzaju. Moglibyśmy wtedy odróżnić jedyną całkowitą historię przyczynową, opartą na opisach neurofizjologicznych, od jedynej całkowitej historii przyczynowej opartej na opisach intencjonalnych, a obie te historie od jedynej całkowitej historii przyczynowej opartej na opisach ekonomicznych, itd. Jednakże stosowalność tych zrelatywizowanych pojęć jedynej całkowitej historii przyczynowej nie podważa w żaden sposób koncepcji obliczania treściowego. Jedyna całkowita historia sekwencji zdarzeń na poziomie syntaktycznym niczego nam rzeczywiście nie powie na temat treści. Jest to całkowicie spójne z jedyną historią przyczynową na poziomie treściowym, która korzysta z pojęcia obliczania.

W literaturze filozoficznej dokonuje się słusznego rozróżnienia między, z jednej strony, szeroko rozumianymi epistemicznymi teoriami wyjaśniania a szeroko rozumianymi teoriami „ontycznymi" czy „metafizycznymi", z drugiej strony[13]. Epistemiczne teorie wyjaśniania głoszą, że to, co powoduje, iż coś jest wyjaśnieniem da się w pełni wyłuszczyć w kategoriach stanów wiedzy lub przekonań, lub też tego, co uważa się za informatywne lub konwersacyjnie stosowne[14]. Natomiast teorie nieepistemiczne, „ontyczne" głoszą, że autentyczne wyjaśnienie zawsze musi zawierać jakieś pojęcie przyczynowości, prawa lub kontrfaktycznych okresów warunkowych, których – jak się utrzymuje – nie sposób sprowadzić do pojęć epistemicznych. Jest to bardzo szeroka kategoria; dwóch myślicieli może się radykalnie różnić co do natury praw, przyczynowości lub kontrfaktycznych okresów warunkowych, mimo

[13] Więcej na temat tego rozróżnienia można znaleźć w Ruben, 1993.

[14] Stąd też dla celów obecnych rozważań pragmatyczne podejścia do wyjaśniania zaliczam do „epistemicznych".

że obaj są teoretykami nieepistemicznymi. Zarówno zwolennicy regularnościowych teorii przyczynowości, jak i ci, którzy odrzucają teorie czysto regularnościowe mogą być teoretykami nieepistemicznymi. Wspominam o tym rozróżnieniu po to, by postawić następujące pytanie: czy poczynione przeze mnie uwagi dotyczące wagi treściowego wyjaśniania obliczeniowego wymagają przyjęcia epistemicznego poglądu na wyjaśnienie?

Choć prezentowana przeze mnie koncepcja obliczania treściowego jest do przyjęcia dla zwolennika epistemicznego podejścia do wyjaśniania, koncepcja ta nie wymaga akceptacji tego podejścia. To prawda, że bardzo łatwo można zacząć formułować niektóre z moich ostatnich tez w postaci uwag na temat wiedzy lub informacji, jak wtedy, gdy utrzymuję, że wyjaśnienie neurofizjologiczne „nic nam nie mówi" o środowiskowych własnościach wyjaśnianych ruchów. Jeśli jednak przyjmujemy, że wyjaśnienie nie jest kwestią w pierwszym rzędzie epistemiczną, to sformułowań tych nie należy traktować jako fundamentalnych, a jedynie jako przeciwstawiające eksplananda treściowego wyjaśnienia obliczeniowego eksplanandom wyjaśnień nietreściowych.

Niemniej jednak połączenie obecnego podejścia do wyjaśnienia obliczeniowego z podejściem ogólnie nieepistemicznym narzuca pewne ograniczenia. Jeśli przyjmujemy tę kombinację poglądów, nie możemy przyjmować tezy, że eksplananda czasowych stanów rzeczy przyjmują postać: „określone zdarzenie *e* zachodzi" i zarazem utożsamiać zdarzenie zajścia określonego ruchu ciała ze zdarzeniem, jakim jest wskazywanie przez daną osobę na określony dom. Gdybyśmy tak uczynili, podważylibyśmy moje wcześniejsze rozróżnienie między przypadkiem, w którym mamy do czynienia z wyjaśnieniem określonej własności relacyjnej zdarzenia a przypadkiem, w którym nie mamy z takim wyjaśnieniem do czynienia. Do podejścia obecnego lepiej pasuje traktowanie eksplanandum jako czegoś na wzór faktu, jako czegoś przybierającego na ogół postać „x_1 ... x_n pozostają w relacji R w chwili t". Podejście to przystaje również lepiej do moich uwag na temat wyjaśniania zjawisk treściowych, poczynionych w części 1. Rozsądnie byłoby twierdzić, że określone zdarzenie w postaci wytworzenia przez jakiś mechanizm reprezentacji mentalnej posiadającej określone własności syntaktyczne jest identyczne z wytworzeniem reprezentacji mentalnej posiadającej określone własności semantyczne – podobnie jak wypowiedzenie pewnego zdania w określonej sytuacji może być identyczne ze zdarzeniem wygłoszenia asercji o określonej treści. Gdyby potraktować eksplanandum danego wyjaśnienia psychologicznego po prostu jako zajście określonego zdarzenia, można byłoby potraktować opisy semantyczne stanów wyjaśniających to zdarzenie jako eksplanacyjnie nieistotne dla tego zdarzenia i słusznie uważać, że eksplanacyjnej mocy syntaktycznych opisów procesu niczego nie brak. Ale gdy przejmiemy, że eksplanandum ma postać „x_1 ... x_n pozostają w relacji R w chwili t", taka deflacyjna reakcja zostaje wykluczona. Eksplanandum mówiące, że zaszło zdarzenie o pewnej własności syntaktycznej jest różne od eksplanandum mówiącego, że zaszło zdarzenie o pewnej własności semantycznej. W przypadku tego ostatniego eksplanandum, odpowiednie będą treściowe warunki wyjaśniające (pomijając przypadki „mieszane"). Dla miłośników sformułowań w kategoriach logiki filozoficznej, powiedzielibyśmy tak: w prawdziwych stwierdzeniach na temat tego, co wyjaśnia, iż dana osoba znajduje się w stanie treściowym, charakterystyki stanów jako stanów treściowych występują w obrębie zasięgu „wyjaśnia", a nie poza jego obrębem.

Gdybyśmy natomiast potraktowali eksplanandum czasowe po prostu jako zajście określonego zdarzenia, i gdybyśmy jednocześnie utożsamili ruch ciała w konkretnej sytuacji z wyjaśnianą czynnością, groziłoby nam, że zostalibyśmy zmuszeni, chcąc nie chcąc, do przyjęcia epistemicznego stanowiska w sprawie eksplanacyjnego wymiaru wyjaśnień obliczeniowych. Rozważmy, na przykład, znaną teorię Lewisa, że wyjaśnić zdarzenie to dostarczyć informacji o jego historii przyczynowej (Lewis, 1986). Teoria ta sama przez się nie jest epistemicznym ujęciem wyjaśniania. Ale jeśli ruch ciała i ułożenie przez osobę ręki między swymi oczyma a słońcem są identyczne (a takie na pewno się wydają, przynajmniej dla mnie), to ich historie przyczynowe, rozumiane jako drzewa genealogiczne zdarzeń przyczynowo je poprzedzających, są również identyczne. Stąd też informacje o historii przyczynowej jednego z tych zdarzeń są informacjami o historii przy-

[15] Uwaga ta dotyczy nie tylko wyjaśniania treściowego, lecz także wyjaśniania eksternalistycznego w ogóle (Peacocke, 1993).

czynowej drugiego z nich[15]. Można więc odnieść wrażenie, że wyjaśnienie treściowe mogłoby mieć odrębne znaczenie tylko w obszarze epistemicznym. Jeśli chcemy uniknąć tej konsekwencji, w ramach z gruntu nieepistemicznego podejścia do wyjaśniania, musimy odróżnić eksplananda.

Niekiedy odróżniamy nawet między eksplanandami, które są koniecznie równoważne. Rozróżnień takich możemy dokonywać nie tylko w ramach epistemicznego poglądu na wyjaśnianie. Wyjaśnienie, dlaczego maszyna drukuje określoną formułę, różni się od wyjaśnienia, dlaczego drukuje on formułę noszącą gödlowski numer *n*, nawet jeśli bycie tą formułą i posiadanie tego numeru gödlowskiego są koniecznie równoważne. Wyjaśnieniem tego, że maszyna wydrukowała tę formułę może być to, że zastosowała ona do jakichś innych formuł pewną regułę syntaktyczną, co dało ową formułę. Wyjaśnieniem wydrukowania formuły o gödlowskim numerze *n* będzie, powiedzmy, raczej to, że dokonała ona obliczenia liczbowego, co dało liczbę *n*, a następnie obliczyła, jakiej formule ona odpowiada, po czym wydrukowała tę formułę. Aby dane zdarzenie można było wytłumaczyć przy pewnym jego opisie, opis ten musi figurować w zasadzie eksplanacyjnej, która leży u podstaw określonego wyjaśnienia[16].

Poza wykazaniem, że nieepistemiczne stanowiska w kwestii wyjaśniania pozwalają na przeprowadzenie całkiem subtelnych rozróżnień, przykład przytoczony w poprzednim akapicie ilustruje jeszcze coś innego. Załóżmy, że maszyna operuje na pewnych formułach początkowych, traktowanych jako dane wejściowe, i stosuje do nich jakąś operację syntaktyczną tak, aby otrzymać inną formułę, którą następnie drukuje. Będą wówczas prawdziwe określone kontrfaktyczne okresy warunkowe, mające postać: „Gdyby dane były takie a takie inne formuły, takie a takie formuły zostałyby wydrukowane". Ale prawdziwe będą również kontrfaktyczne okresy warunkowe, mające postać: „Gdyby dane były formuły noszące takie a takie numery gödlowskie, wydrukowane zostałyby formuły noszące takie a takie numery gödlowskie". Jak z tego widać, prawdziwość określonych kontrfaktycznych okresów warunkowych dotyczących własności zdarzeń wyjaśniających nie wystarcza, aby te własności były własnościami eksplanacyjnymi. Potwierdzane kontrfaktyczne okresy warunkowe są często dobrą wskazówką tego, co jest eksplanacyjne w stosunku do czego, lecz same przez się nie są w pełni wystarczające do ustalenia relacji eksplanacyjnych, ani też, *a fortiori*, tego, co sprawia, że coś jest wyjaśnieniem[17].

6. PORÓWNANIE Z DENNETTEM

Chciałbym teraz sprecyzować charakterystykę treści subpersonalnej, która wyłania się z prezentowanego tu stanowiska, zestawiając ją z koncepcją treści subpersonalnej rozwiniętą w ważnych pracach Dennetta. Oba stanowiska zgodne są co do tego, że należy dokonać rozróżnienia między tym, na czym polega bycie systemem intencjonalnym a empirycznym wyjaśnieniem tego, w jaki sposób organizm jest systemem intencjonalnym. Zgodne są również co do tego, że jednym z charakterystycznych zadań psychologii jest właśnie wyjaśnienie, w jaki sposób organizmowi udaje się być systemem intencjonalnym (Dennett, 1987, zwłaszcza ss. 43-65). Istnieje jednak poważna rozbieżność między Dennettem a podejściem tutaj prezentowanym, dotycząca sposobu, w jaki należy pojmować treść stanów subpersonalnych oraz tego, na czym polega odpowiednie wyjaśnienie faktu, iż organizm skutecznie realizuje jakiś system intencjonalny. Różnica zdań sprowadza się do właściwego ujęcia relacji między treściami stanów subpersonalnych a treścią stanów intencjonalnych z poziomu personalnego.

Skrajna postać twierdzenia o niezależności treści z poziomu personalnego i stanów personalnych byłaby następująca:

> Treści stanów związanych z obliczeniami subpersonalnymi nie są tego samego rodzaju, co treści stanów z poziomu personalnego, ani nie można im przypisać treści, których askrypcja

[16] Może to być jedynie „pozbawiona parametrów" wersja opisu, który figuruje w zasadzie eksplanacyjnej, jeśli podstawowa zasada wyjaśniania uwzględnia parametry dla kontekstu i rozmaite obiekty znajdujące się w środowisku podmiotu.

[17] Przyznaję, że pełna obrona obliczania treściowego musi uwzględnić kwestię sprawczości stanów zawierających treści subpersonalne. Artykuł ten jest już bardzo obszerny, tak więc chciałbym tylko napomknąć, że moje stanowisko w sprawie sprawczości treści subpersonalnych byłoby analogiczne do przyjętego w Peacocke, 1993. Podejrzewam, że ci, którzy domagają się czegoś więcej, żywią pragnienie, które nigdy nie zostanie zaspokojone.

byłaby uzasadniona przez ich zdolność do wyjaśniania faktów dotyczących treści stanów z poziomu personalnego.

Jest to twierdzenie o niezależności treści subpersonalnych stanów obliczeniowych od treści z poziomu personalnego; nazwijmy je po prostu „twierdzeniem o niezależności". W pismach Dennetta znajdujemy rozmaite argumenty, które są albo argumentami na rzecz twierdzenia o niezależności, albo dostarczają materiału do konstrukcji takich argumentów. A ponieważ w sposób wyraźny zależy mi na podważeniu twierdzenia o niezależności, będę kwestionował te argumenty. Należy jednak zaznaczyć, że przypisuję Dennettowi twierdzenie o niezależności nie bez pewnych zastrzeżeń. Jest tak częściowo dlatego, że wydaje się jasne, iż niektóre z celów, do których Dennett, w *Consciousness Explained* (Dennett, 1991), wykorzystuje pojęcie stanów mających treść subpersonalną wymagają, by treści te były tego samego rodzaju, co treści stanów z poziomu personalnego. Rozważamy, na przykład, pandemonijny model, przedstawiony w rozdziale 8 *Consciousness Explained*, w celu wyjaśnienia, dlaczego osoba wypowiada coś mającego taką właśnie treść z poziomu personalnego, a nie jakąś inną. Rywalizujące ze sobą demony lub podmioty muszą mieć powiązane z nimi treści pojęciowe, o ile to zwycięstwo jednego z nich (lub jakiejś ich grupy) ma wyjaśnić wybór określonej treści wypowiedzi. Niemniej jednak w innych miejscach Dennett podaje ważne argumenty ogólne, i związany z nimi materiał, na poparcie twierdzenia o niezależności.

Trzy argumenty są kluczowe.

Argument pierwszy wychodzi od przesłanki, że „mózg jest maszyną syntaktyczną" (Dennett, 1987, s. 61). Nic nie jest w stanie dokonać tego, co jest niemożliwe, a mianowicie wydobyć własności semantyczne z cech syntaktycznych. Czy to, co powiedziałem o wyjaśnianiu za pomocą treściowego obliczania subpersonalnego koliduje z tą oczywistą prawdą?

Warto zauważyć, że jeśli argument ten stanowi zagrożenie, jest to zagrożenie o ogólniejszym zasięgu. Argument ten przemawiałby równie silnie przeciw wyjaśnieniom obliczeniowym, odwołującym się do czysto „wąskich" treści, ponieważ wąskie treści także wykraczają poza własności syntaktyczne[18]. Argument tego nie da się zasadnie wykorzystać do pokazania wyższości wyjaśnień obliczeniowych zawierających tylko wąskie treści nad wyjaśnieniami obliczeniowymi wykorzystującymi treści eksternalistyczne. Nie można go, na przykład, zasadnie wykorzystać do potwierdzenia tezy, że subpersonalna psychologia obliczeniowa jest prawomocna jedynie pod warunkiem, że korzysta wyłącznie z wąskich treści.

Proponowane przeze mnie stanowisko w żadnym wypadku nie zakłada wykonania niemożliwego zadania, jakim jest uzyskanie semantyki z syntaktyki. W obliczaniu treściowym stany semantyczne są rezultatem albo innych stanów semantycznych, albo – w przypadkach mieszanych pierwszego rodzaju – stanów niesemantycznych w połączeniu z założonym tłem środowiskowym (przy zachodzeniu innych warunków). Zgadzam się z tym, że nic nie może być wewnętrznie reprezentacją czegoś, jak i z tym, że w przypadku reprezentacji o treści eksternalistycznej, fakty internalne dotyczące sposobu użycia reprezentacji przez podsystem nie mogą ustalić jej treści. To właśnie dlatego, że moje stanowisko uznaje trafność tych uwag, jestem w stanie twierdzić, że charakterystyczna rola eksplanacyjna stanów treściowych musi się pod pewnymi względami różnić od roli stanów beztreściowych. Jednym z takich względów jest właśnie wyjaśnianie relacyjnie zindywidualizowanych stanów rzeczy.

Sam Dennett głosi pogląd, że mózg „mógłby być tak zaprojektowany [...] by naśladować zachowanie niemożliwego obiektu (maszyny semantycznej) czyniąc użytek z bliskich (dostatecznie bliskich) szczęśliwych zbieżności między strukturalnymi regularnościami – środowiska oraz jego własnych internalnych stanów i operacji – a typami semantycznymi (tamże, s. 61). „Zwierzę musi wiedzieć, kiedy osiągnęło cel, jakim jest znalezienie i spożycie pokarmu, lecz zadowala się wykrywaczem tarcia-w gardle-po-którym-następuje-pęcznienie żołądka [...]" (s. 61). Przypuśćmy, że ograniczymy się do tego wykrytego stanu internalnego i ustalimy złożone relacje internalne i środowiskowe, na mocy których wykrycie tego stanu jest równie dobre jak wykrycie, że organizm spożył pokarm. Wydaje mi się, że określamy wówczas relacje, na mocy których wykrywaczowi tego stanu można przypisać treść sprowadzającą

[18] Wśród takich wąskich treści znalazłyby się również treści pojęciowe (*notional*) ustalane przez „światy pojęciowe" Dennetta, 1987.

się do tego, że pokarm został spożyty. Wcześniej usiłowałem podać kilka reguł przypisywania treści stanom subpersonalnym. Reguły te są na poziomie subpersonalnym odpowiednikiem tego, co Dennett uznaje na poziomie personalnym – jego „reguł przypisywania" wchodzących w skład teorii systemu intencjonalnego, reguł rządzących przyporządkowaniem treści intencjonalnych stanom z poziomu personalnego (Dennett, 1987, s. 58). Jeśli, jak usiłowałem dowieść, semantyczne i relacyjne własności stanów późniejszych można wyjaśnić obecnością jakiegoś stanu z tą treścią dotyczącą spożycia pokarmu, to w pewnym sensie umysł jest mimo wszystko maszyną semantyczną, i to bez implikowania wykonywania czegoś niemożliwego.

Dennett wysuwa drugi argument, bardzo groźny dla mojej koncepcji, wychodzący z założenia, że skoro reprezentacje subpersonalne nie mogą posiadać swych treści na mocy sposobu, w jaki są one wykorzystywane przez jakiegoś „użytkownika bez zobowiązań" spoza systemu, jawnie inteligentne użycie tych reprezentacji trzeba wyjaśnić przez jakieś dające się zneutralizować homunkulusy. Ale, głosi dalej argument, mechaniczne działanie tych neutralizowalnych homunkulusów polegać może wyłącznie na operacjach wewnątrzsystemowych. Oto ów argument, w innym cytacie z Dennetta dotyczącym Fodora (Dennett, 1978b, s. 102):

> [Fodor] nie zauważa, że [homunkulusy] nadal odgrywają teoretyczną rolę ustalania „tematu" i „słownika" wiadomości, które komunikują. Jeśli przedsięwzięcie Fodora można uchronić od niekoherencji – co, jak myślę, jest w zasadzie możliwe – przyjmując, że wiadomości w kodzie wewnętrznym są samozrozumiałymi reprezentacjami, to dzieje się tak za sprawą nałożenia na pojęcie internalnego systemu reprezentacji dodatkowych ograniczeń, raczej podkreślających niż eliminujących rozróżnienie między przypisywaniem przekonań i pragnień z poziomu personalnego a przypisywaniem treści operacjom wewnątrzsystemowym z poziomu subpersonalnego.

Gdyby to rozumowanie było trafne, to moje własne stanowisko stanęłoby przed kłopotliwym dylematem. Jeśli homonkulusy ustalają semantykę wiadomości subpersonalnych, to albo zostają one ostatecznie zneutralizowane, albo nie. Nie do przyjęcia jest to, aby pozostały one ostatecznie nie zneutralizowane. Ale jeśli podlegają one neutralizacji przez mechanizmy, które są czysto internalnie zindywidualizowane, jak wydaje się sugerować powyższy cytat z Dennetta, to same z siebie nie będą one odpowiednią podstawą wyjaśnień specyficznie eksternalistycznych.

Dylemat ten pomija pewną możliwość. Stan subpersonalny może posiadać treść na mocy swoich nieinteligentnych relacji do innych stanów, które są wszelako w dalszym ciągu zindywidualizowane relacyjnie. Przez „nieinteligentne" nie rozumiem tutaj „wolne od wszelkich charakterystyk treściowych". Prace samego Dennetta nauczyły nas (Dennett, 1978a), że przez „nieinteligentny" należy rozumieć „nie zakładający zdolności psychicznych, do wyjaśnienia których się przyczynia". Na przykład, w ujęciu wielu zdolności w kategoriach sieci neuronowych, od czytania po widzenie przestrzenne, obliczane reprezentacje – fonemów, głębi czy cokolwiek innego – posiadają treści, które posiadają dlatego, że są wytwarzane w pewien sposób przez określone stany, posiadające również treści reprezentacyjne o charakterze eksternalistycznym. Reprezentacje te wytwarzane są w sposób mechaniczny w sensie „obszaru pośredniego", o którym mówiliśmy wcześniej, i w żadnym razie nie zakładają zdolności podlegającej wyjaśnieniu.

Trzeci, równie groźny argument zmierza do uzasadnienia, że „przekonania i pragnienia nie są właściwymi obiektami badawczymi psychologii poznawczej. Mówiąc inaczej, teorie kognitywistyczne są, lub powinny być, teoriami poziomu subpersonalnego, na którym przekonania i pragnienia znikają, ustępując miejsca innego rodzaju reprezentacjom dotyczącym czegoś innego" (Dennett, 1978b, s. 105). Argument ten jest następujący: dwa różne obliczeniowo programy, subpersonalnie egzemplifikowane u dwóch różnych osób, Mary i Ruth, mogą oba uzasadniać przypisanie przekonania o danej treści. Przypisanie przekonania o tej właśnie treści będzie, w danym kontekście, równie dobrze wyjaśniało pewne działania Mary, jak i Ruth. Różne, subpersonalnie egzemplifikowane programy wyjaśniają różne układy danych – Dennett wspomina o różnicach między Mary i Ruth dotyczących opóźniania działania, błędów i tym podobnych rzeczy. Przypisywanie przekonań po prostu nie służy do wyjaśniania tego rodzaju danych. A zatem, taka

jest konkluzja argumentu, psychologia subpersonalna nie zajmuje się w ogóle przekonaniami i pragnieniami. Argument ten można uogólnić z przekonań i pragnień na inne stany intencjonalne z poziomu personalnego, np. na stany percepcyjne.

Moim zdaniem argument ten opiera się na wnioskowaniu, którego nie powinniśmy akceptować. Ma on taką samą postać, co następujące rozumowanie: „Podłoże tej samej własności, jaką jest bycie lepkim, wspólnej dla dwóch różnych cieczy, mogą stanowić dwie różne struktury molekularne. Te różne struktury molekularne będą wyjaśniały pewne różnice między tymi cieczami: różne reakcje na przepływ promieni rentgenowskich, różne wzorce interakcji chemicznych itp. Przypisanie lepkości po prostu nie zmierza do wyjaśnienia tak szczegółowych danych. Dlatego też określone struktury molekularne nie wyjaśniają lepkości". Moim zdaniem przejście do tego ostatniego zdania jest przykładem *non sequitur*. Fakt, że lepkość zgadza się z kilkoma różnymi układami danych szczegółowych nie może dowodzić, że struktura molekularna nie wyjaśnia lepkości; w rzeczywistości bowiem ją wyjaśnia. Wydaje mi się, że równie błędne byłoby zaproponowanie analogicznego argumentu na rzecz konkluzji: „Stany subpersonalne nie wyjaśniają treściowych stanów z poziomu personalnego".

Być może próbuję zbyt wiele wyczytać z argumentu Dennetta, i zaliczając postawy propozycjonalne do stanów po prostu przesądzam sprawę na niekorzyść jego całego stanowiska. Ale nie chodzi tu o coś, co opiera się na tendencyjnej presumpcji, że klasyfikacje psychologii ludowej dzielą stany subpersonalne w ten sam sposób, co klasyfikacje psychologii subpersonalnej. Przypuśćmy, że mówimy bardzo neutralnie, iż predykat „jest przekonany, że to a to" jest prawdziwy zarówno w przypadku Mary, jak i Jane. Niemniej jednak możemy chcieć uzyskać wyjaśnienie, w kategoriach psychologii subpersonalnej, dlaczego tak się dzieje. Teorie subpersonalne podlegają w istocie precyzyjniejszym danym niż askrypcje psychologii intencjonalnej z poziomu personalnego, ale mimo to mogą one wyjaśniać dane mniej precyzyjne, w tym spełnianie owego predykatu przez Mary i Jane.

7. DALSZE ZADANIA

Wygłosiłem szereg twierdzeń na temat natury i znaczenia treściowego wyjaśniania obliczeniowego. Istnieje wiele kierunków poszukiwań, w jakie można wyruszyć z punktu, do którego obecnie doszliśmy. Wymienię trzy. Pierwszy leży w obrębie psychologii. Powinniśmy starać się dowiedzieć o wiele więcej o zasadach przypisywania określonych treści stanom subpersonalnym. To, o czym wspomniałem wcześniej w tym artykule to jedynie bardzo ogólne warunki przypisywania treści, poparte pewną liczbą przykładów. W teorii treści konceptualnej – które to pojęcie z istoty należy do poziomu personalnego – przeprowadzamy rozróżnienie między ogólnymi ograniczeniami nakładanymi na postać, jaką powinien przyjąć opis opanowywania określonego pojęcia, z jednej strony, a konkretnymi przypadkami tej postaci dla pewnych pojęć, z drugiej strony. Podobne rozróżnienie stosuje się do treści subpersonalnych i naszym celem powinno być poszukiwanie dalszych przypadków podlegających tym ogólnym ograniczeniom. Nasze zwykłe możliwości myślenia o przedmiotach, czasie, ruchu, materii, innych umysłach, ich emocjach, intencjach i sprawstwie tworzą razem imponujący zestaw eksternalnie zindywidualizowanych zdolności i wiedzy. Wyjaśnienia, jak to możliwe, że możemy się w tych stanach znajdować, i dlaczego znajdowanie się w nich jest poprawne (lub niepoprawne), muszą zakładać rozbudowane subpersonalne systemy reprezentacji. Musimy rozważać nie tylko, czym one są, ale i dlaczego poprawne jest przypisywanie im określonych, charakterystycznych dla nich treści. Prawidłowe zrozumienie tych spraw będzie ważne nie tylko dla wyjaśnienia psychologicznego, lecz potencjalnie także dla konstytutywnych kwestii dotyczących tego, co to znaczy posiadać eksternalnie zindywidualizowane zdolności podlegające wyjaśnieniu.

Drugi kierunek, w którym warto podążać stwarza nadzieję na lepsze zrozumienie roli obliczania treściowego w sieciach koneksjonistycznych. Jest coraz bardziej prawdopodobne, że gdy naszym celem jest zrozumienie ich funkcjonowania, to musimy przypisywać treść zarówno rozległym, jak i niektórym węższym wzorcom aktywacji w

tych sieciach. Poza obszarem klasycznym, gdzie obliczanie ma własności autentycznie syntaktyczne, klasyczna teoria obliczania opracowana przez Turinga, Churcha i Kleene'a, dostarcza nam niewielu wskazówek, jak rozumieć obliczanie. Potrzebujemy teorii, która dostarczy ogólnego aparatu do opisu w kategoriach treściowych aktywności obliczeniowej sieci wykonującej określone zadanie. W procesie dochodzenia do takiej teorii być może będziemy musieli wziąć pod uwagę różne relacje między treściami, a może i rodzaje treści odmienne od tych, które pasują do klasycznych przypadków syntaktycznych. Osobiście jestem optymistą w sprawie możliwości zbudowania takiej teorii. Ale póki jej nie zbudujemy, nie możemy twierdzić, że w pełni rozumiemy naturę i znaczenie obliczania treściowego.

Trzeci kierunek możliwy do obrania z miejsca, w którym się znajdujemy, wychodzi od spostrzeżenia, że wyartykułowane tu relacje między treściowym wyjaśnieniem obliczeniowym a eksternalnymi eksplanandami psychologii są przypadkami ogólnego typu, który może być obecny również w innych naukach szczegółowych. W wielu innych naukach szczegółowych, zwłaszcza w naukach społecznych, możemy uzyskać pewną postać rozróżnienia eksternalne-internalne. Często się zdarza, że eksplananda specyficzne dla określonej nauki szczegółowej są indywidualizowane eksternalnie, odpowiednio do właściwego dla tej nauki przeprowadzenia rozróżnienia internalne--eksternalne. Kiedy tak jest, rodzi się również pytanie, jak obiekty lub podmioty z dziedziny tej nauki szczegółowej mogą się znajdować w stanach zindywidualizowanych eksternalnie, przy właściwym dla dziedziny tej nauki rozróżnieniu internalne-eksternalne? Gdy stany tej nauki szczegółowej będą zarazem stanami reprezentacyjnymi, pojawią się też pytania o wyjaśnienie ich poprawności lub niepoprawności. We wszystkich tego typu przypadkach właściwe wydaje się uogólnienie wyjaśnienia, z jakim mamy do czynienia w treściowym wyjaśnianiu obliczeniowym. Sugestia ta rodzi setki pytań. Jedyne, co mogę w tym momencie zrobić, to polecić ten sposób podejścia do niektórych ogólnych, interdyscyplinarnych problemów dotyczących natury wyjaśniania.

Przełożyli z języka angielskiego:
Helena Grzegołowska-Klarkowska, Marcin Iwanicki i Tadeusz Szubka

LITERATURA

Boden, M. (1988). *Computer Models of Mind: Computational Approaches to Theoretical Psychology*. Cambridge: University Press.

Burge, T. (1979). Individualism and the Mental. *Midwest Studies in Philosophy, 4*, 73-121.

Burge, T. (1986). Individualism and Psychology. *Philosophical Review, 95*, 3-45.

Churchland, P.S. i Sejnowski, T.J. (1990). Neural Representation and Neural Computation, W: J.E. Tomberlin (Ed.), *Philosophical Perspectives, 4: Action Theory and Philosophy of Mind*. Atascadero, CA.: Ridgeview Publishing Company.

Churchland, P.S. i Sejnowski, T.J. (1992). *The Computational Brain*. Cambridge, MA.: MIT Press.

Davidson, D. (1984). Radical Interpretation, W: *Inquiries into Truth and Interpretation*. Oxford University Press [Interpretacja radykalna, przeł. P. Józefowicz, W: D. Davidson, *Eseje o prawdzie, języku i umyśle*. Warszawa: PWN, 1992].

Davies, M. (1986). Externality, Psychological Explanation and Narrow Content. *Proceedings of the Aristotelian Society, supplementary volume 60*, 263-83.

Dennett, D. 1978a: Artificial Intelligence as Philosophy and as Psychology. W: *Brainstorms*. Montgomery, VT.: Bradford Books.

Dennett, D. (1978b). *Brainstorms*. Montgomery, VT.: Bradford Books.

Dennett, D. (1987). *The Intentional Stance*. Cambridge, MA.: MIT Press.

Dennett, D. (1991). *Consciousness Explained*. Boston: Little, Brown & Company.

Egan, F. (1992). Individualism, Computation, and Perceptual Content. *Mind, 101*, 443-59.

Fodor, J. (1991). Methodological Solipsism Considered as a Research Strategy in Cognitive Psychology. W: D.M. Rosenthal (Ed.), *The Nature of Mind*. New York: Oxford University Press.

Grandy, R. (1973). Reference, Meaning, and Belief. *Journal of Philosophy, 70*, 439-52.

Hornsby, J. (1986). Physicalist Thinking and Conceptions of Behaviour. W: P. Pettit i J. McDowell (Eds), *Subject, Thought, and Context*. Oxford University Press.

Kitcher, Philip (1982). *Abusing Science: The Case Against Creationism*. Cambridge, MA.: MIT Press.

Lackner, J. i Garrett, M. (1972). Resolving Ambiguity: Effects of Biasing Context in the Unattended Ear, *Cognition, 1*, 359-72.

Lewis, D. (1986). Causal Explanation. W: *Philosophical Papers: Volume II*, New York: Oxford University Press.

Marr, D. (1982). *Vision*. San Francisco: Freeman.

Marslen-Wilson, W. i Tyler, L. (1987). Against Modularity, [In:] J. Garfield (Ed.), *Modularity in Knowledge Representation and Natural-Language Understanding*. Cambridge, MA.: MIT Press.

McDowell, J. (1986). Functionalism and Anomalous Monism. W: E. LePore i B.McLaughlin (Eds), *Actions and Events: Perspectives on the Philosophy of Donald Davidson*. Oxford: Basil Blackwell.

Peacocke, C. (1981). Demonstrative Thought and Psychological Explanation. *Synthese, 49*, 187-217.

Peacocke, C. (1983). *Sense and Content: Experience, Thought*

and their Relations. Oxford University Press.
Peacocke, C. (1993). Externalist Explanation. *Proceedings of the Aristotelian Society, 93*, 203-30.
Putnam, H. (1975). The Meaning of ‚Meaning'. W: *Mind, Language and Reality*. Cambridge University Press [Znaczenie wyrazu „znaczenie". W: H. Putnam, *Wiele twarzy realizmu i inne eseje*, przeł. A. Grobler. Warszawa: PWN, 1998].
Ruben, D. (1993). Introduction. W: D. Ruben (Ed.), *Explanation*. Oxford University Press.
Stich, S. (1991). Paying the Price for Methodological Solipsism. W: D.M. Rosenthal (Ed.), *The Nature of Mind*. New York: Oxford University Press.