AVANT, wol. XI, nr 3
ISSN: 2082-6710 avant.edu.pl
DOI: 10.26913/avant.2020.03.26

Koherencja drogowskazem prawdy? Spójność jako źródło błędnych reprezentacji

Komentarz do książki Krystyny Bieleckiej *Błądzę, więc myślę. Co to jest błędna reprezentacja?*

Paweł Gładziejewski
Katedra Kognitywistyki
Uniwersytet Mikołaja Kopernika w Toruniu
pawel.gladz@gmail.com

Przyjęto 8 stycznia 2020; zaakceptowano 22 kwietnia 2020; opublikowano 28 lipca 2020

Abstrakt

W niniejszym artykule próbuję rozwinąć i uogólnić Krystyny Bieleckiej koherencyjną teorię rozpoznawania błędnych reprezentacji. Próbuję pokazać, że koncepcja opisana w książce *Błądzę, więc myślę* (2019) może być wstępem do szerszej historii. Rozważam hipotezę, zgodnie z którą koherencja odgrywa zasadniczą rolę w procesie ewaluacji reprezentacji umysłowych, a przez to w ich nabywaniu i odrzucaniu. System poznawczy aktywnie maksymalizuje koherencję reprezentacji, działając zgodnie z zasadą Koherencji jako Drogowskazu Prawdy (zasadą KDP). Oparte na niespójności wykrywanie błędów – główny przedmiot zainteresowania koncepcji Krystyny Bieleckiej – stanowi tylko jedną z form takiego wykorzystania koherencji. Wskazuję kilka potencjalnych kontrprzykładów dla zasady KDP. Chodzi o koncepcje i modele procesów poznawczych (dotyczące percepcji, kontroli ruchu, pamięci epizodycznej i neutralizowania dysonansu poznawczego), w świetle których maksymalizacja koherencji systematycznie prowadzi do wytwarzania błędnych reprezentacji. Argumentuję, że te kontrprzykłady są tylko pozorne i nie uderzają w zasadę KDP jako podstawę „higieny epistemicznej" systemu poznawczego.

Słowa kluczowe: epistemologia; koherencja; pamięć epizodyczna; przetwarzanie predykcyjne; reprezentacje mentalne; reprezentacjonizm

1. Wstęp

Najlepszym powodem do komplementowania innego filozofa lub filozofki jest sytuacja, w której praca tej osoby wpłynęła na nasze własne myślenie, prowadząc je na nowe, ciekawe i produktywne ścieżki. Lektura świetnej książki Krystyny Bieleckiej *Błądzę, więc myślę* (2019) właśnie tak na mnie wpłynęła (zakładając, rzecz jasna, że potrafię w miarę poprawnie oceniać ścieżki własnego myślenia i rozpoznawać błędy w tych ocenach). Niniejszy artykuł jest inspirowany tą lekturą. Jego przedmiotem jest rola koherencji w procesie ewaluacji reprezentacji przez system poznawczy. Przedstawiona tu propozycja może zostać potraktowana jako próba uogólnienia i rozwinięcia teorii błędu reprezentacyjnego przedstawionej w *Błądzę, więc myślę*.

Chcę rozważyć hipotezę, że proces nabywania i odrzucania reprezentacji przez system poznawczy podlega pod pryncypium maksymalizacji koherencji tych reprezentacji. Systemy poznawcze aktywnie redukują niepewność co do stanów świata (w tym własnych stanów) w taki sposób, by nabywane reprezentacje zwiększały ogólną koherencję ich modeli środowiska. Zgodnie z takim ujęciem, oparte na koherencji rozpoznawanie błędu reprezentacyjnego – proces stojący w centrum zainteresowań Krystyny Bieleckiej – jest tylko jednym przykładem szerszego procesu. Posługując się koncepcją koherencji Paula Thagarda (2000), spróbuję pokazać, że maksymalizacja koherencji nie będzie oznaczać tylko odrzucania reprezentacji, które są niespójne z innymi reprezentacjami posiadanymi przez system poznawczy. Poszukiwanie spójności będzie również determinować bardziej „aktywne" procesy nabywania i wyboru reprezentacji, tak, aby cały system powiązanych reprezentacji posiadanych przez system był możliwie koherentny.

Po co systemowi poznawczemu spójność? Rozważę epistemologiczną odpowiedź na to pytanie, zgodną w duchu z koncepcją Krystyny Bieleckiej. Odpowiedź ta głosi, że koherencja jest drogowskazem prawdy. Dbanie o spójność własnego modelu świata nie tylko pozwala unikać błędnych reprezentacji (to lekcja z *Błądzę, więc myślę*), ale też sprzyja nabywaniu poprawnych lub prawdziwych reprezentacji, czy też utrzymywaniu określonej, pożądanej proporcji reprezentacji poprawnych do błędnych (por. Goldman, 1986)[1].

[1] Za Krystyną Bielecką przyjmuję, że prawda jest formą korespondencji pomiędzy reprezentacją a tym, co reprezentowane. Zakładam, że kwalifikowane w kategoriach wartości logicznej mogą być reprezentacje o charakterze zdaniowym lub propozycjonalnym. Używany tu termin „poprawność" jest rozszerzeniem pojęcia korespondencyjnej prawdy, mającym na celu objęcie reprezentacji, których korespondencja z tym, co reprezentowane, nie może być naturalnie ujęta w kategoriach klasycznie rozumianej wartości logicznej. Na przykład problematyczne jest powiedzieć o mapie kartograficznej, że jest prawdziwa; może ją jednak kwalifikować jako bardziej lub mniej poprawną, w zależności od tego, jak wiernie jej struktura odzwierciedla (relatywnie do pewnego założonego poziomu szczegółowości) strukturę reprezentowanego terenu. Wydaje się, że wiele reprezentacji umysłowych bardziej przypomina reprezentacje strukturalne w rodzaju mapy niż zdania lub sądy (Gładziejewski i Miłkowski, 2017). Ich korespondencja z rzeczywistością powinno się ujmować w kategoriach (nie)poprawności, a nie prawdy lub fałszu. Zasada, o której mowa w tekście głównym, mogłaby zatem być bardziej adekwatnie, lecz niezgrabnie, nazwana „zasadą Koherencji jako Drogowskazu Prawdy lub Poprawności".

Teza, że dbanie o koherencję jest dbaniem o poprawność lub prawdziwość reprezentacji, okazuje się jednak problematyczna w świetle niektórych teorii kognitywistycznych oraz odkryć empirycznych. Wskażę przypadki, gdzie poprawnie działający system poznawczy generuje błędne reprezentacje, a za fakt ten odpowiadają właśnie próby uspójnienia reprezentacji. Inaczej mówiąc, zabieganie o koherencję w sposób systematyczny, nieakcydentalny przynajmniej niekiedy prowadzi na manowce.

Postaram się skrupulatnie rozważyć, czy wskazane przypadki rzeczywiście są pełnoprawnymi kontrprzykładami dla pryncypium koherencji jako drogowskazu prawdy. Proponowana przeze mnie odpowiedź będzie negatywna. Będę argumentował, że w omówionych rzekomych kontrprzykładach związki koherencji z fałszem są niewinnym produktem ubocznym ograniczeń mechanizmów poznawczych lub wynikają z faktu, że funkcją danego mechanizmu w ogóle nie jest formowanie poprawnych reprezentacji.

Struktura artykułu jest następująca. W drugiej sekcji krótko streszczam propozycję Krystyny Bieleckiej i próbuję ją rozwinąć. Omawiam tam pewien sposób rozumienia koherencji i wskazuję, jak maksymalizacja tak rozumianej koherencji może kształtować proces rewizji reprezentacji przez system poznawczy. W sekcji trzeciej proponuję zasadę Koherencji jako Drogowskazu Prawdy (zasadę KDP), która stanowić ma epistemologiczne pryncypium uzasadniające zabieganie o spójność reprezentacji przez systemy poznawcze. Wskazuję tam też szereg potencjalnych kontrprzykładów dla zasady KDP. W czwartej sekcji pokazuję, że kontrprzykłady te są jedynie pozorne i nie zagrażają tezie, że dbanie o spójność jest narzędziem budowania poprawnych modeli środowiska.

2. Koherencja i jej rola w ewaluacji reprezentacji

2.1. Koherencyjna teoria błędu i pojęcie koherencji

Nie ma reprezentacji bez możliwości błędnej reprezentacji[2]. Zdanie to stało się truizmem wśród filozofów podejmujących problem natury reprezentacji i treści intencjonalnej. Istnieje też niemal konsensus, że naturalistyczna koncepcja treści – o ile w ogóle jest osiągalna – powinna mieć komponent teleologiczny, odwołujący się do roli funkcjonalnej, jaką reprezentacja odgrywa dla systemu poznawczego. Błąd reprezentacyjny to, mówiąc z grubsza, użycie reprezentacji w sposób niezgodny z jej funkcją (na przykład uruchomienie przez żabę struktury pełniącej funkcję detektora much w odpowiedzi na pojawienie się trzmiela). Choć koncepcja Krystyny Bieleckiej wpisuje się w ten ogólny trend teoretyczny, idzie ona o krok dalej. Inspirując się pracami Marka Bickharda (1999, 2004), Krystyna Bielecka zauważa, że systemy poznawcze powinny dysponować dostępem do faktu, iż niektóre ich reprezentacje są błędne (na przykład do faktu, że obiekt w otoczeniu został błędnie rozpoznany jako mucha). Tylko w ten sposób możliwe jest na przykład uczenie się na błędach.

[2] W pracy tej przyjmuję bez argumentów, że reprezentacje wewnętrzne istnieją i są ważnym narzędziem eksplanacyjnym dla kognitywistów. Jest to pogląd obecnie poddawany w wątpliwość (zob. np. Chemero, 2014; Hutto i Myin, 2012). W artykule tym przyjmuję reprezentacjonizm, ponieważ jest on centralnym założeniem pracy Krystyny Bieleckiej. Ponadto zakładam prawdziwość tego stanowiska, ponieważ najprawdopodobniej jest ono prawdziwe (zob. np. Gładziejewski, 2015).

Potrzebujemy zatem teorii wyjaśniającej, jak systemy poznawcze rozpoznają własne błędne reprezentacje. Po groźbą regresu *ad inifinitum*, koncepcja rozpoznania błędu powinna unikać twierdzenia, że system może przyjąć „boski" punt widzenia, z którego porównuje swoją reprezentację z samą rzeczywistością. Autorka *Błądzę, więc myślę* broni koncepcji, zgodnie z którą do rozpoznania błędu wystarczy porównanie reprezentacji z inną reprezentacją (tu i dalej por. Bielecka, 2019, s. 207–215)[3]. Jeśli wykryta zostanie niespójność między tymi reprezentacjami (na przykład między wzrokową reprezentacją rzekomej muchy a dotykową reprezentacją trzmiela, pochodzącą z receptorów rozlokowanych na żabim języku), jedna z nich może zostać odrzucona jako fałszywa. Kryterium rozstrzygającym, która z reprezentacji powinna zostać odrzuca, jest szacowana rzetelność reprezentacji (na przykład reprezentacja pochodząca z modalności dotykowej jest traktowana jako bardziej rzetelna niż reprezentacja wzrokowa). W ujęciu Krystyny Bieleckiej, błąd reprezentacyjny nie musi być sam (meta)reprezentowany; wystarczy, że system poznawczy ma dyspozycję do odpowiedniego reagowania na powstające błędy.

Aby rozszerzyć koncepcję autorki *Błądzę, więc myślę*, chciałbym przyjrzeć się bliżej samemu pojęciu koherencji. Jak sądzę, uwidoczni to, że teoria Krystyny Bieleckej stanowi fragment potencjalnie szerszej opowieści.

Niesprzeczność czy brak niespójności między reprezentacjami (ich treściami) wystarczy do zachodzenia koherencji tylko w bardzo słabym znaczeniu tego terminu. Niesprzeczny zbiór zupełnie niepowiązanych ze sobą zdań spełnia taki słaby warunek koherencji. Aby jednak mieć pełniejszy obraz znaczenia koherencji w poznaniu, musimy wziąć pod uwagę związki treściowe między reprezentacjami. Inaczej mówiąc, powinniśmy myśleć o całym systemie czy sieci bardziej lub mniej „pasujących" do siebie reprezentacji, powiązanych relacjami takimi jak na przykład wynikanie czy wyjaśnianie. Poziom koherencji zbioru czy systemu reprezentacji w takim silniejszym ujęciu może być wyznaczany nie tylko przez brak niespójności, ale także przez liczbę i siłę związków (inferencyjnych, ale być może też asocjacyjnych) między reprezentacjami należącymi do tego zbioru czy systemu (por. BonJour, 1985).

Jak sądzę, ta ogólna intuicja znajduje owocne doprecyzowanie w pracy Thagarda (tu i dalej por. Thagard, 2000, s. 15–40). Propozycja ta nie jest, rzecz jasna, jedyną filozoficzną próbą scharakteryzowania natury koherencji (zob. Olsson, 2003). Jednak ujęcie Thagarda służyć ma rozjaśnianiu roli koherencji w kognitywistycznych modelach zjawisk poznawczych. Jest to zbieżne z koncepcją przedstawioną w *Błądzę, więc myślę*, która w zamierzeniu jest naturalistyczna (por. Bielecka, 2019, s. 16–23). Dlatego właśnie skorzystam z propozycji Thagarda jako *prima facie* bliskiej ogólnym filozoficznym założeniom Krystyny Bieleckiej.

Thagard charakteryzuje koherencję w kategoriach spełniania ograniczeń nakładanych na zbiór złożony z rzeczy posiadających treści semantyczne. Mówiąc bardzo nieformalnie, w ujęciu tego autora „dostajemy punkty" za sposób, w jaki akceptujemy i odrzucamy reprezentacje. Gromadzimy punkty, kiedy akceptujemy te reprezentacje, które w ogólnym rozrachunku łączą się z innymi w bogatą sieć relacji (zachodzą tu pozytywne ograniczenia). Dostajemy też

[3] Mówiąc precyzyjniej, teoria autorki *Błądzę, więc myślę* dopuszcza również rozpoznanie błędu przez porównanie reprezentacji z informacją pozbawioną własności semantycznych (Bielecka, 2019, s. 209).

punkty, jeśli odrzucamy reprezentacje, które w ogólnym rozrachunku nie pasują do tych akceptowanych przez nas (wiążą je z nimi negatywne ograniczenia). Im więcej uzyskamy punktów, tym bardziej koherentny jest nasz zbiór reprezentacji.

Spróbujmy teraz wyrazić tę myśl bardziej precyzyjnie. Weźmy pod uwagę jakiś skończony zbiór *E*, złożony z elementów $\{e_i\}$, np. sądów, zdań czy reprezentacji mentalnych. Jeśli między jakimiś dwoma elementami *E*, e_1 i e_2, zachodzi relacja wynikania, wyjaśniania, analogii, pozytywnej asocjacji itp., powiemy o e_1 i e_2, że spełniają pozytywne ograniczenie, to znaczy, że są koherentne. Przy założeniu, że akceptujemy e_1, powinniśmy zaakceptować również e_2, i *vice versa* (koherencja jest relacją symetryczną). Jeśli odrzucamy e_1, powinniśmy odrzucić e_2, i *vice versa*. Jeśli jednak e_1 oraz e_2 są wzajemnie sprzeczne, niespójne ze sobą, powiązane negatywną asocjacją itp., powiemy o e_1 i e_2, że spełniają negatywne ograniczenie, to znaczy, że nie są koherentne. Przy założeniu, że akceptujemy e_1, będziemy chcieli odrzucić e_2, i *vice versa*.

Mając zbiór *E*, możemy wyznaczyć zbiór *C* złożony z par elementów *E*, $\{(e_i, e_j)\}$. *C* dzieli się na podzbiory *C*+ i *C*-. *C*+ obejmuje pary elementów *E* powiązane pozytywnymi ograniczeniami, a *C*- obejmuje pary elementów *E* powiązane negatywnymi ograniczeniami. Każdemu typowi ograniczenia odpowiada pewna waga, *w*.

Możemy teraz powiedzieć, na czym polega maksymalizacja koherencji *E*. Chcemy wiedzieć, które elementy *E* powinniśmy zaakceptować, a które odrzucić, jeśli pragniemy być możliwie spójni. Inaczej mówiąc, chcemy podzielić *E* na podzbiór *A*, złożony z elementów, które akceptujemy, oraz rozłączny z nim podzbiór *O*, złożony z elementów, które odrzucamy. Nasze zadanie polega zatem wyznaczeniu *A* i *O* w taki sposób, by w maksymalny sposób zrealizować następujące warunki:

- Jeśli para (e_i, e_j) należy do *C*+, to e_i należy do *A* wtedy i tylko wtedy, gdy e_j należy do *A*.
- Jeśli para (e_i, e_j) należy do C-, to e_i należy do *A* wtedy i tylko wtedy, gdy e_j należy do *O*.

Każdemu możliwemu podziałowi *E* na *A* i *O* odpowiada wartość *W*, będąca sumą (ważonych) spełnionych przez ten podział ograniczeń. Wartość *W* mierzy poziom koherencji takiego podziału. Inaczej mówiąc, mierzy ona, w jakim stopniu zbiór *A* zawiera reprezentacje „pasujące" do siebie nawzajem, a zbiór *O* zwiera reprezentacje „niepasujące" do tych z *A*. Istnieje pewien maksymalnie koherentny podział zbioru *E*, to znaczy taki, dla którego wartość *W* jest największa. Każdy inny podział jest tym bardziej koherentny, im bliżej mu do tej maksymalnej wartości[4].

[4] Przedstawiona tu krótka rekonstrukcja z konieczności pomija wiele subtelności propozycji Thagarda. Warto na przykład zaznaczyć, że wyznaczenie wszystkich możliwych podziałów *E*, aby wybrać ten o maksymalnym *W*, będzie możliwe tylko dla trywialnie małych zbiorów reprezentacji. Wraz z powiększaniem zbioru *E*, czas potrzebny do odnalezienia maksymalnie koherentnego podziału będzie rósł szybciej niż wielomianowo. Dlatego Thagard (2000, s. 25–39) omawia w swojej pracy kilka obliczalnych algorytmów, pozwalających aproksymować maksymalnie koherentny podział.

2.2. Koherencja służy nie tylko filtrowaniu błędów

Prześledźmy teraz konsekwencje idei, że reprezentacje środowiska posiadane przez system poznawczy tworzą *zbiór* czy *system* powiązanych elementów, który może być bardziej lub mniej koherentny w sensie zaproponowanym przez Thagarda. Jeśli tak jest, potencjalne nowe elementy tego zbioru – nowe reprezentacje – mogą być oceniane przez pryzmat tego, czy dodanie ich zwiększa, czy zmniejsza wartość *W* dla całego zbioru. Jak sądzę, takie spojrzenie pozwala na dwa sposoby rozwinąć i uogólnić koncepcję Krystyny Bieleckiej. Zwróćmy uwagę, że przypadki rozpoznania błędu omawiane przez autorkę *Błądzę, więc myślę* na ogół obejmują pary niespójnych reprezentacji, zwłaszcza percepcyjnych[5]. Omawiane tu ujęcie koherencji pozostawia możliwość szacowania poziomu koherencji dla par reprezentacji. Jednak pokazuje ono zarazem, że możliwa jest ewaluacja reprezentacji przez pryzmat tego, jak „pasuje" ona do całego zbioru czy systemu powiązanych reprezentacji.

Ponadto zauważmy, że dla autorki *Błądzę, więc jestem* ocena koherencji jest pewnego rodzaju filtrem pozwalającym systemowi poznawczemu odrzucić te reprezentacje, które w ogólnym rozrachunku nie są koherentne z innymi reprezentacjami; to znaczy reprezentacje, których dodanie do zbioru skutkuje obniżeniem wartości *W*. Przykłady opisywane przez Krystynę Bielecką dotyczą sytuacji, gdy nowa reprezentacja jest powiązana negatywnym ograniczeniem z inną, bardziej rzetelną reprezentacją (i nie towarzyszą temu żadne „równoważące" pozytywne ograniczenia z innymi reprezentacjami). Jednak wartość *W* wzrasta, jeśli *dodajemy* do zbioru reprezentacje powiązane z innymi elementami *pozytywnym* ograniczeniem (czy też reprezentacje, dla których stosunek pozytywnych i negatywnych ograniczeń jest taki, że ogólnie dodanie tych reprezentacji do systemu zwiększa wartość *W*). Jeśli założyć zatem, że organizm jest zainteresowany maksymalizowaniem koherencji reprezentacji, to zwiększanie wartości *W* przez nową potencjalną reprezentację powinno sprawiać, iż zostanie ona nabyta. To znaczy, że system poznawczy nabędzie tę reprezentację. Tym samym maksymalizacja koherencji może nie tylko odgrywać „pasywną" rolę epistemicznego ochroniarza przed fałszywymi reprezentacjami, ale też „aktywnie" kształtować proces nabywania reprezentacji – kształtować sposób, w jaki system poznawczy redukuje swoją niepewność co do stanów środowiska.

Co do zasady, *dictum* „Bądź spójny!" może wpływać na ten proces na co najmniej cztery sposoby.

Po pierwsze, organizmy poruszają się w świecie niepewności. Oznacza to, że niejednokrotnie system poznawczy będzie musiał decydować między kilkoma alternatywnymi (i wzajemnie niespójnymi) reprezentacjami, chociażby między kilkoma możliwymi interpretacjami niejednoznacznego sygnału zmysłowego. W takiej sytuacji może zostać wybrana ta spośród alternatywnych reprezentacji, która jest najbardziej spójna z istniejącym już modelem środowiska. Na przykład w sytuacji rywalizacji obuocznej preferowana będzie ta interpretacja sceny wzrokowej, która wchodzi w odpowiednie relacje probabilistyczne/inferencyjne z istniejącym

[5] Trzeba zaznaczyć, że jest to celowe uproszczenie stosowane przez autorkę (por. Bielecka, 2019, s. 212–213). Sugeruję jednak, że ta idealizacja może być zbyt daleko posunięta. Dopiero mówienie o całych systemach czy repertuarach powiązanych reprezentacji w pełni pokazuje znaczenie koherencji w procesie nabywania reprezentacji.

modelem świata; w istocie system poznawczy będzie oscylował pomiędzy dwiema takimi interpretacjami (por. Hohwy, Roepstorff i Friston, 2008).

Po drugie, potrzeba maksymalizacji koherencji powinna czasem promować zachowania eksploracyjne, których celem jest odkrycie struktury środowiska poprzez uzupełnienie istniejącego systemu reprezentacji (por. Daw, O'Doherty, Dayan i in., 2006). Pomyślmy o szczurze, który dysponuje allocentryczną mapą środowiska, określającą między innymi położenie dwóch dobrych miejsc żerowania. Jeśli szczur „nie wie", czy jedno z tych miejsc jest dostępne z drugiego miejsca – lub brakuje mu reprezentacji ścieżki optymalnie łączącej te dwa miejsca – może zaangażować się w działanie eksploracyjne, służące redukcji niepewności co do tego stanu rzeczy. Uzupełnienie luki w mapie może zwiększyć ogólny poziom koherencji tej mapy, wskazując relację zachodzącą między do tej pory nie powiązanymi ze sobą reprezentacjami.

Po trzecie, nabywanie nowych, zwiększających koherencję reprezentacji może odbywać się poprzez wyciąganie konsekwencji z posiadanych już reprezentacji. System poznawczy może na przykład uruchamiać wewnętrzne symulacje kontrfaktycznych sytuacji, aby oszacować zachowanie świata w różnych możliwych scenariuszach (por. np. Ullman, Spelke, Battaglia i Tenenbaum, 2017). Rezultat tego rodzaju symulacji może zostać dodany do istniejącego repertuaru, powiększając jego ogólny poziom koherencji.

Po czwarte, zwróćmy uwagę, że zachodzenie niespójności pomiędzy dwiema reprezentacjami wcale nie musi prowadzić systemu poznawczego do odrzucenia jednej z nich. Patrząc z bardziej holistycznej pespektywy, reprezentacje mogą być weryfikowane nie pojedynczo, ale jako system (por. Quine, 2000). Na przykład, pochodzący z percepcji „kandydat" na nową reprezentację może zostać aktywnie „wpasowany" do ogólnego systemu poprzez dodanie hipotez pomocniczych, usuwających jego niespójność z wcześniejszymi reprezentacjami. Procedura taka wcale nie musi być nieracjonalna i kompletnie *ad hoc*. Jest to dobry sposób, aby ochronić dobrze ugruntowane modele środowiska przed pochopnym obaleniem.

3. Koherencja drogowskazem prawdy? Potencjalne kontrprzykłady

Przyjmijmy, że niektóre procesy związane z nabywaniem i rewizją reprezentacji przebiegają w taki sposób, aby zwiększać ogólny poziom koherencji tych reprezentacji (nawet jeśli norma maksymalizacji koherencji nie jest sama eksplicytnie reprezentowana w systemie). Dążenie do koherencji możemy potraktować jako normę epistemiczną, pod którą podpadają przynajmniej niektóre procesy poznawcze.

Epistemologowie często uzasadniają obowiązywanie różnych norm epistemicznych, odwołując się do faktu, że podporządkowanie tym normom sprzyja nabywaniu prawdziwych przekonań; „dobre" normy w epistemologii to na ogół normy sprzyjające prawdzie. Jest to pogląd często podzielany przez zarówno zwolenników internalizmu (por. BonJour, 1985, s. 7–8), jak i eksternalizmu (Goldman, 1986) w epistemologii.

Chcę wykorzystać tę ideę, aby zrozumieć rolę koherencji w procesach poznawczych. Proponuję uznać, że zachowywanie spójności reprezentacji jest przydatne systemom poznawczym jako sposób tworzenia *poprawnych* modeli środowiska. Dbanie o to, by reprezentacje

„pasowały" do siebie jako elementy koherentnego systemu, sprzyja temu, by reprezentacje te były prawdziwe lub aby utrzymana była odpowiednia proporcja reprezentacji poprawnych względem niepoprawnych. Jest to istota zasady Koherencji jako Drogowskazu Prawdy (KDP).

Zasada KDP jest inspirowana koherencyjnymi koncepcjami uzasadnienia epistemicznego w epistemologii. Potencjalnie dziedziczy ona problemy tych koncepcji, związane między innymi z regresem uzasadnień *ad infinitum* czy możliwością istnienia wielu alternatywnych, wewnętrznie równie spójnych, ale wzajemnie niekompatybilnych systemów reprezentacji (por. Olsson, 2003). Nie będę tutaj podejmował próby odpowiedzi na takie potencjalne zarzuty. Zwrócę tylko uwagę na fakt, że problemy koherencjonizmu w sprawie uzasadnienia epistemicznego są często rozwiązywane kompromisowo – poprzez wzbogacenie koherencyjnej koncepcji o aspekt czy komponent o charakterze fundacjonistycznym (por. Haack, 1993; Olsson, 2003; por. także zasadę *data priority* w: Thagard, 2000). Na ogół filozofowie wskazują jakąś wyróżnioną subklasę sądów, przekonań czy reprezentacji, które są uzasadnione w inny sposób niż (jedynie) poprzez koherencję z innymi sądami, przekonaniami czy reprezentacjami. Sądzę, że analogiczny ekumeniczny ruch w kierunku fundacjonizmu możliwy jest do wykonania w także w kontekście kognitywistycznym (Gładziejewski, 2017). Na obecne potrzeby zaznaczę jednak tylko, iż zasada KDP nie ma oznaczać lub implikować, że koherencja jest *jedynym* drogowskazem prawdy lub że dbanie o spójność *wystarcza* systemowi poznawczemu, aby zapewnić poprawność reprezentacji.

Chcę skupić się na innym potencjalnym problemie z zasadą KDP. Chodzi o przykłady zjawisk poznawczych, które wydają się tę zasadę bezpośrednio obalać – przykłady, w których spójność nie jest drogowskazem prawdy. Mówiąc inaczej, chcę wskazać zjawiska poznawcze, w których proces nadawania spójności systemowi wewnętrznych reprezentacji prowadzi system poznawczy – w sposób systematyczny, nieakcydentalny – do formowania niepoprawnych reprezentacji.

Oczywiście można by po prostu wskazać zaburzenia poznawcze, w których utrzymywanie koherencji systematycznie prowadzi do błędów. Osoby z anozognozją czy somatoparafrenią potrafią uporczywie bronić swoich urojeń przed kontrargumentami, usuwając wewnętrzne sprzeczności poprzez dodawanie hipotez ad *hoc*. Osoby przechodzące epizod psychotyczny potrafią stworzyć systematycznie fałszywy, lecz wewnętrzne spójny system przekonań, oparty na jednym centralnym urojeniu prześladowczym. Jednak przykłady takie obejmują dysfunkcje mechanizmów poznawczych. Zwolennik zasady KDP mógłby po prostu stwierdzić, że pryncypium to obowiązuje dla poprawnie działających systemów poznawczych i psychopatologie nie mogą go obalić. Właśnie dlatego poniższa lista obejmuje jedynie przypadki, w których system poznawczy działa w sposób poprawny, to znaczy jego wewnętrzne mechanizmy funkcjonują zgodnie ze swoim projektem czy funkcją[6].

[6] Kierując się sugestią Krystyny Bieleckiej (2019, s. 207), przyjmuję pluralistyczne podejście do atrybucji funkcjonalnych. Uważam, że funkcje mechanizmów poznawczych (lub ich komponentów) mogą być ugruntowane zarówno w historii adaptacyjnej, jak i w procesie podtrzymywania szerszego systemu w stanie dalekim od równowagi termodynamicznej z otoczeniem.

Przykład 1: Iluzje percepcyjne jako wynik wnioskowania percepcyjnego

Pierwszy przykład nawiązuje do tradycyjnego już w psychologii ujęcia percepcji zmysłowej jako nieuświadamianego wnioskowania (Gregory, 1980). Zgodnie z tym podejściem, stany percepcyjne są hipotezami na temat przyczyn sygnałów zmysłowych. Percepcja jest zatem wnioskowaniem do najlepszego wyjaśnienia, z wykorzystaniem wiedzy uprzedniej. Najnowsze odmiany tej koncepcji postulują, że proces wnioskowania aproksymuje wnioskowanie bayesowskie, które z kolei realizowane jest przez minimalizację błędu predykcyjnego (tak zwane przetwarzanie predykcyjne, por. Hohwy, 2013). Mówiąc z grubsza, system poznawczy buduje model statystyczny środowiska, który wykorzystuje, aby przewidywać nadchodzące strumienie sygnałów zmysłowych (i który optymalizuje pod wpływem błędów predykcyjnych).

Za Alexem Kieferem (2017) przyjmuję, że maksymalizacja koherencji odgrywa centralną rolę w tak rozumianych wnioskowanych percepcyjnych. Kiefer broni takiego stanowiska, odwołując się do ogólnych własności rozumowań indukcyjnych. Polegają one w jego ujęciu (które samo inspirowane jest pracą Gilberta Harmana; por. Harman, 1973) na zmianie systemu przekonań czy reprezentacji w taki sposób, aby „zwiększyć jego spójność eksplanacyjną, uczynić go bardziej kompletnym, mniej *ad hoc*, bardziej wiarygodnym" (Harman, 1973, s. 159, tłum. PG). Kiefer analizuje modele koneksjonistyczne percepcji, pokazując, że proces stopniowej minimalizacji średniego błędu predykcyjnego (energii swobodnej) w sieciach neuronowych można rozumieć jako maksymalizację koherencji kodowanego w sieci modelu statystycznego. Przekładając to na język poznawczy, bayesowska percepcja polega na wyborze (pod wpływem sygnałów zmysłowych) takiej reprezentacji percepcyjnej, która jest najbardziej spójna z rozkładami prawdopodobieństw zakodowanymi w posiadanym już przez system modelu. Zarazem zmiana samego modelu pod wpływem błędów predykcyjnych ma go uczynić bardziej spójnym z dochodzącymi do systemu wzorcami sygnałów zmysłowych.

Kluczowe znaczenie ma fakt, że ten racjonalny, maksymalizujący koherencję proces wnioskowania powinien systematycznie powodować występowanie *iluzji percepcyjnych* (które traktuję tu jako błędne reprezentacje percepcyjne). Już klasyczne inferencyjne koncepcje percepcji wskazywały właśnie ukrytą „logikę" iluzji wzrokowych jako dowód na to, że percepcja jest wnioskowaniem (Gregory, 1963; 1997). Iluzja Ponza czy iluzja towarzysząca percepcji pokoju Amesa stanowią najbardziej prawdopodobne – na gruncie wiedzy uprzedniej o stałości wielkości obiektów – wyjaśnienia osobliwych danych zmysłowych (na przykład dwa obiekty powodują obraz siatkówkowy tej samej wielkości, pomimo że wskazówki kontekstowe sugerują, iż obiekty te znajdują się w różnej odległości od obserwatora; albo obraz siatkówkowy rzutowany przez poruszającą się osobę powiększa się, chociaż wskazówki kontekstowe sugerują, że odległość pomiędzy tą osobą a obserwatorem nie zmienia się). Bardziej współczesne ujęcia bayesowskie również są wykorzystywane do wyjaśnienia iluzji wzrokowych (Geisler i Kesler, 2002; Rescorla, 2015) oraz iluzji wynikających z interakcji między modalnościami, takich, jak iluzja gumowej ręki (Hohwy, 2013) czy efekt McGurka (Magnotti i Beauchamp, 2017). Próby zachowania koherencji w systematyczny sposób prowadzą zatem mechanizmy percepcyjne do tworzenia błędnych reprezentacji.

Przykład 2: Kontrola ruchu, wnioskowanie aktywne i błędne reprezentacje precyzji proprioceptywnej

Drugi przykład również zaczerpnięty jest z koncepcji przetwarzania predykcyjnego, jednak tym razem zastosowanej do wyjaśnienia kontroli ruchu (por. Adams, Shipp i Friston, 2013; Hohwy, 2013). Teoria ta traktuje kontrolę ruchu jako proces bliźniaczy percepcji. O ile percepcja to próba dostosowania wewnętrznych predykcji do nadchodzących sygnałów zmysłowych, kontrola ruchu przypomina samospełniającą się przepowiednię. Polega ona na zmianie pozycji ciała w taki sposób, aby dopasować sygnały zmysłowe do wewnętrznych predykcji. Podczas inicjacji ruchu system przewiduje, że znajduje się w określonej pozycji, a przez to przewiduje określone pobudzenia zmysłowe. Błąd predykcyjny jest minimalizowany relatywnie do tych predykcji poprzez zmianę pozycji ciała (osiąganą dzięki aktywacji sekwencji łuków odruchów), tak aby otrzymać przewidziany zwrotny sygnał zmysłowy (zwłaszcza proprioceptywny).

Opisany wyżej proces „wnioskowania aktywnego" ma być obliczeniowo ekwiwalentny wnioskowaniu percepcyjnemu (Hohwy, 2013). Oznacza to, że oba procesy są formą wnioskowania bayesowskiego, realizowanego przez minimalizację błędu predykcyjnego. Przez ekstrapolację można zatem uznać wnioskowanie aktywne za proces maksymalizacji koherencji modelu środowiska (por. także Friston, 2018). Mówiąc intuicyjnie, system przyjmuje „za dobrą monetę" pewną reprezentację samego siebie („właśnie sięgam po kubek z kawą"), a następnie dopasowuje do tej reprezentacji stan świata (sięgając po kubek z kawą) w taki sposób, aby otrzymywane dane były spójne z przyjętą hipotezą i z bardziej ogólnymi oczekiwaniami („zawsze o tej porze piję kawę").

Jak argumentuje Wiese (2016), wnioskowanie aktywne w sposób systematyczny wiąże się z wytwarzaniem fałszywych reprezentacji. *Przed* inicjacją ruchu system poznawczy otrzymuje dane proprioceptywne świadczące *przeciwko* przyjętej hipotezie (kiedy siedzi nieruchomo, otrzymuje dane niespójne z hipotezą, że sięga po kawę). Aby rozpocząć proces wnioskowania aktywnego, system musi „zawiesić wiarę" w dostępne dane zmysłowe, tak aby jego aktywność była dyktowana bardziej przez założoną hipotezę („przekonanie uprzednie"), a nie przez dane. W przetwarzaniu predykcyjnym procesem regulującym relatywny wpływ odgórnych hipotez i oddolnych sygnałów zmysłowych na przebieg procesów obliczeniowych jest *szacowanie precyzji* (Hohwy, 2013). Jeśli system poznawczy szacuje sygnały oddolne jako sygnały o niskiej precyzji (wysokiej wariancji), będzie je traktował jako mniej wiarygodne, a zatem będzie opierał swoją aktywność bardziej na uprzednim modelu i wywiedzionych z niego hipotezach. Szacowanie precyzji jest kluczowe w inicjacji ruchu. Aby wykonać ruch, system poznawczy musi z konieczności obniżyć estymowaną precyzję sygnałów zmysłowych (szczególnie proprioceptywnych), *niezależnie* od tego, jaka jest ich *rzeczywista* precyzja. Tylko w ten sposób sygnały te mogą zostać odpowiednio „wygaszone". Wytworzenie fałszywej reprezentacji precyzji proprioceptywnej jest zatem warunkiem możliwości kontroli motorycznej. Raz jeszcze przetwarzanie predykcyjne pokazuje, że maksymalizacja koherencji wewnętrznego modelu – realizowana w procesie wnioskowania aktywnego – dokonuje się kosztem poprawnych reprezentacji.

Przykład 3: Błędy pamięci epizodycznej jako skutek nadawania spójności wspomnieniom

Pamięć epizodyczna jest stosunkowo zawodną zdolnością poznawczą (tu i dalej por. De Brigard, 2014; Jagodzińska, 2008; Schacter, 2001). Wbrew subiektywnemu poczuciu weredyczności towarzyszącemu wspominaniu, nasze mentalne rekonstrukcje osobistej przeszłości są często zniekształcone; czasem w ogóle nie odpowiadają one żadnemu przeszłemu wydarzeniu. Nawet wiele spośród częściowo poprawnych wspomnień epizodycznych cechuje się specyficznymi przesunięciami perspektywy przestrzennej w stosunku do pierwotnego doświadczenia (na przykład we wspomnieniach często obserwujemy siebie z perspektywy trzecioosobowej; pole widzenia we wspomnieniach jest często szersze niż w rzeczywistości, przez co mamy wrażenie, że miejsca, których dotyczą te wspomnienia, były większe). Osoby badane często fałszywie pamiętają, że usłyszały określony wyraz w sekwencji słów, o ile jest on semantycznie powiązany ze słowami, które rzeczywiście usłyszeli. Wspomnienia świadków w sprawach sądowych często okazują się znacząco zniekształcone lub całkowicie błędne. Badani często formują całkowicie fałszywe wspomnienia – takie, które nie odpowiadają żadnemu przeszłemu wydarzeniu – jeśli otrzymają wystarczająco wiarygodne wskazówki, że określone („przypominane") wydarzenie zaszło.

Aby zrozumieć związek (przynajmniej niektórych) błędów pamięci epizodycznej z pojęciem koherencji, rozważmy następującą propozycję teoretyczną, dotyczącą etiologii tych błędów. Zacznę od sugestywnego przykładu (za: De Brigard, 2014). Wyobraźmy sobie, że jesteśmy świadkami wypadku samochodowego: granatowa Honda nie zwalnia przed znakiem „STOP" i zderza się z ciężarówką. Kiedy uczestniczymy w tym wydarzeniu, nasza uwaga selektywnie skupia się na elementach sceny, które są związane z naszym własnym przeżyciem (odległość między naszym samochodem a samochodami uczestniczącymi w kolizji, naciśnięcie hamulca, reakcja samochodów jadących za nami) lub mają własności „przyciągające" ludzki aparat uwagowy (twarz kierowcy Hondy, głośny pisk opon). Na większości elementów sceny w ogóle się nie skupiamy. Spośród tych elementów, na których skupia się nasza uwaga, tylko niektóre pozostają w pamięci roboczej na tyle długo, by stworzyć reprezentację w pamięci długotrwałej. Te elementy, które zostaną zakodowane w pamięci długotrwałej, będą obejmowały szczegóły percepcyjne z poszczególnych modalności zmysłowych, zakodowane w sposób rozproszony po różnych fragmentach kory. Proces przypominania będzie więc wiązał się z koniecznością ponownej integracji takich rozproszonych reprezentacji w reprezentację całego zdarzenia. Co jednak szczególnie istotne dla bieżących celów, kiedy policjant zapyta nas o szczegół wypadku, który nie został w ogóle zakodowany, jest spore prawdopodobieństwo, że odpowiemy na pytanie, *uzupełniając* brakujący element. Jednak sposób, w jaki wypełnimy lukę, nie będzie przypadkowy, lecz spójny z ogólnym modelem statystyki środowiska (De Brigard, 2014). Jeśli znaki „ustąp pierwszeństwa" występują na skrzyżowaniach częściej niż znaki „STOP", słupy wysokiego napięcia czy gigantyczne muchomory, to właśnie takim znakiem uzupełnimy nasze wspomnienie. Próba stworzenia spójnego wspomnienia na bazie skąpego szkieletu zakodowanych fragmentów sprawi, że wspomnienie to będzie zawierać błędy.

Przykład ten można uogólnić (De Brigard, 2014). Ze względu na ograniczenia pojemności pamięci roboczej, tylko niektóre elementy zdarzeń są kodowane w pamięci długotrwałej. Tworzenie wspomnienia epizodycznego oznacza zatem mentalną rekonstrukcję zdarzenia na podstawie dość szczątkowego materiału źródłowego. Co kluczowe, luki w takiej szczątkowej reprezentacji są wypełniane zgodnie z zasadą zachowywania koherencji. „Puste" fragmenty są uzupełniane poprzez nadawanie wewnętrznej spójności wspomnieniu, w taki sposób, by elementy pasowały do siebie asocjacyjnie czy probabilistycznie (to znaczy były dopasowane do siebie zgodnie ze schematem jakiegoś zdarzenia, miejsca czy okoliczności). Proces ten jednocześnie sprawia, że pojedyncze wspomnienia stają się spójne z wiedzą na temat statystycznej struktury środowiska. Skoro jednak czasem zdarzenia, w których rzeczywiście uczestniczyliśmy, odbiegały od tego, co statystycznie typowe, nadawanie koherencji wspomnieniom musi prowadzić do błędów. W skrajnych przypadkach osoby będą tworzyły reprezentację pamięciową jakiegoś wydarzenia, które w ogóle nie miało miejsca, tylko dlatego, że jego wystąpienie jest stosunkowo wiarygodne w świetle wiedzy ogólnej (na przykład zgubienie się centrum handlowym w dzieciństwie).

Idea ta jest spójna z klasycznymi badaniami na temat popełnianych przez ludzi typów błędów pamięciowych (na przykład fałszywe wspomnienia słów, o ile te „pasują" do sekwencji, to znaczy istnieją asocjacyjne czy znaczeniowe związki między słowami rzeczywiście usłyszanymi a słowami fałszywie zapamiętanymi). Jest ona podstawą bayesowskich [7] modeli obliczeniowych, wyjaśniających błędy pamięciowe jako efekt dopasowania wspomnienia do wiedzy o statystycznym współwystępowaniu określonych obiektów czy zdarzeń w środowisku (Anderson i Schooler, 1991; Hemmer i Steyvers, 2009). Ma ona też wsparcie w wynikach empirycznych, pokazujących, że efektywność uczenia asocjacyjnego wiąże się z *większą* liczbą popełnianych błędów pamięciowych (Carpenter i Schacter, 2016). Wszystkie te wątki zbiegają się w propozycji teoretycznej, zgodnie z którą pamięć epizodyczna jest tylko aspektem działania szerszego mechanizmu, odpowiadającego również za myślenie kontrfaktyczne czy prospekcję. Do tej ostatniej tezy nawiążę jeszcze w następnej sekcji. Na obecne potrzeby wystarczy podkreślić, że maksymalizacja koherencji wydaje się odgrywać rolę w etiologii błędnych wspomnień epizodycznych.

Przykład 4: Dysonans poznawczy a zachowywanie koherencji

Ostatni przykład rozbratu pomiędzy maksymalizacją koherencji a poprawnością reprezentacji pochodzi z psychologii społecznej. Dysonans poznawczy to awersyjny stan wynikający z poczucia niespójności między własnymi przekonaniami lub między przekonaniami a działaniami. Zjawisko to ma w szczególności obejmować sytuacje, w których osoba natrafia na świadectwa (lub sama podejmuje działania świadczące) przeciwko przekonaniom dotyczącym samooceny (własnej wartości, moralności, racjonalności itd.), wartości grup, z którymi ta osoba się utożsamia, czy też przeciwko fundamentalnym poglądom religijnym czy politycznym tej osoby (Harmon-Jones i Mills, 1999; Travis i Aronson, 2008).

[7] Modele te, choć bayesowskie, nie są oparte na przetwarzaniu predykcyjnym.

Ludzie wykazują silną tendencję do łagodzenia lub usuwania dysonansu poznawczego. Zabiegi te są w zasadzie z definicji próbami przywrócenia koherencji systemowi przekonań. Strategie realizacji tego celu mają jednak bardzo często charakter epistemicznie nieracjonalny[8]. Rewizja przekonań służąca usunięciu dysonansu poznawczego jest zazwyczaj stronnicza, nakierowana na to, aby uchronić wybrane przekonania danej osoby przed ponurą prawdą. Choć proces ten ma służyć przywróceniu czy zachowaniu spójności, to przebiega on w taki sposób, że *a priori* przesądzona jest prawdziwość pewnych przekonań bazowych – dotyczących własnej wartości, wartości własnej tożsamości grupowej, zasadniczych przekonań metafizycznych itd. W jednym z klasycznych przykładów przywódca apokaliptycznego ruchu religijnego, skonfrontowany z faktem, że proroctwo nie spełniło się, uzna po prostu, że to modlitwy odroczyły koniec świata (por. cały szereg podobnych, barwnych przykładów, omówionych w: Travis i Aronson, 2008). Przekonanie to tworzy pas ochronny dla „twardego jądra" teorii, opartej na bezwzględnym przywiązaniu do przekonań religijnych. Jest to przykład, w którym maksymalizacja koherencji pozwala zachować jedno fałszywe przekonanie (o prawdziwości systemu religijnego) poprzez wytworzenie innego fałszywego przekonania (o modlitwach odraczających koniec świata).

Oprócz tworzenia hipotez ochronnych *ad hoc*, ludzie dysponują całym spektrum strategii łagodzenia lub usuwania dysonansu poznawczego, które wiążą się z formowaniem i ochroną fałszywych przekonań (lub odrzucaniem przekonań prawdziwych). Na przykład ludzie dużo bardziej krytycznie i sceptycznie podchodzą do świadectw zagrażających ich poczuciu wartości czy bezpieczeństwa niż do takich, które zgadzają się z ich założeniami (Ditto, Munro, Apanovitch i in., 2003). Aby przywrócić zachwiane poczucie kontroli nad światem, ludzie potrafią dostrzegać sensowne wzorce tam, gdzie ich nie ma (Whitson i Galinsky, 2008). Podjęte decyzje czy działania świadczące przeciwko własnej wartości są rutynowo uzasadniane przez (fałszywe) racjonalizacje (por. Harmon-Jones i Mills, 1999). Na przykład osoby białe wykazują tendencję do nieudzielania pomocy osobom czarnoskórym (ale nie innym białym osobom), o ile występują okoliczności, które pozwalają im zracjonalizować takie zaniechanie (Saucier, Miller i Doucet, 2005). Niektóre badania nad hipokryzją moralną pokazują też, że ludzie pozorują sprawiedliwe zachowania, aby przekonać samych siebie o własnej moralności (Batson, 2008). Wszystkie te przykłady obejmują sytuacje, w których obrona zachwianej spójności dokonuje się – w sposób systematyczny – kosztem prawdy.

4. W obronie epistemologicznej roli koherencji

W tej sekcji pokażę, że powyższe przypadki nie są dobrymi kontrprzykładami dla zasady KDP. Choć może to brzmieć paradoksalnie, nie zamierzam zaprzeczyć tezie, że w omówionych przypadkach dbanie o koherencję systematycznie prowadzi systemy poznawcze do tworzenia błędnych reprezentacji. Chcę jednak wskazać, że żaden z tych przypadków nie podpada pod *rodzaj mechanizmu poznawczego*, który stanowiłby dobry kontrprzykład dla zasady KDP.

[8] Pod tym względem przypadek dysonansu poznawczego różni się od trzech omawianych wcześniej w tej sekcji przykładów. Te ostatnie wydają się cechować pewną racjonalnością (np. wnioskowanie Bayesowskie jest wręcz paradygmatycznie racjonalnym procesem).

Rozważmy następującą możliwość, zaczerpniętą z pracy Ryana McKaya i Daniela Dennetta (2009). Mamy komputer, na którym zainstalowany jest program pozwalający rozwiązywać problemy z zakresu fizyki. Funkcją tego programu jest wygenerowanie poprawnego rozwiązania (na przykład poprawnej reprezentacji trajektorii ruchu jakiegoś obiektu w polu grawitacyjnym Ziemi) przy otrzymaniu określonych danych wejściowych. Załóżmy, że program wykorzystuje prawa fizyki newtonowskiej i jest zaprojektowany do rozwiązywania problemów niewymagających wykorzystania fizyki relatywistycznej. Możemy pomyśleć o trzech potencjalnych scenariuszach:

(1) Mechanizm działa poprawnie (zgodnie ze swoim projektem czy funkcją), jego funkcją jest tworzenie poprawnych reprezentacji i taką poprawną reprezentację wytwarza. W naszym przykładzie komputer z zainstalowanym programem działa poprawnie, otrzymuje zadanie rozwiązywalne na gruncie fizyki newtonowskiej i generuje poprawny wynik.

(2) Funkcją mechanizmu jest tworzenie poprawnych reprezentacji, jednak działa on niepoprawnie (w wyniku uszkodzenia bądź losowego zaburzenia) i generuje błędną reprezentację. W naszym przykładzie komputer otrzymuje zadanie rozwiązywalne na gruncie fizyki newtonowskiej, jednak podczas działania programu dochodzi do losowego błędu i wynik wyjściowy jest niepoprawny.

(3) Mechanizm działa poprawnie (zgodnie ze swoim projektem czy funkcją), jego funkcją jest tworzenie poprawnych reprezentacji, jednak generuje on błędną reprezentację jako oczekiwany skutek uboczny. W naszym przypadku program otrzymuje zadanie nierozwiązywalne na gruncie fizyki newtonowskiej (na przykład określenie masy obiektu poruszającego z prędkością bliską prędkości światła) i pomimo że w procesie obliczeniowym nie występują żadne losowe błędy, wynik obliczeń jest błędny.

Zwróćmy szczególną uwagę na trzecią ewentualność. Dennett i McKay (2009) określają tego rodzaju błąd jako „wybaczalny" z perspektywy projektu czy funkcji mechanizmu. Błąd ten jest naturalnym, oczekiwanym produktem ubocznym faktu, że nasz mechanizm został zaprojektowany do radzenia sobie z określoną klasą problemów; wykorzystano w nim rozwiązania nieoptymalne, lecz wystarczająco dobre dla tego celu (na przykład implementacja fizyki newtonowskiej była łatwiejsza i bardziej oszczędna obliczeniowo niż implementacja fizyki einsteinowskiej). Kiedy poprawnie działający mechanizm wytwarza błędne rozwiązania dla problemów wykraczających poza klasę problemów, do rozwiązywania których jest on zaprojektowany, to popełniane błędy są niewinne. Takie błędy to nie dysfunkcje.

Rozważmy jeszcze typ mechanizmu ilustrowany przykładem budzika zaprojektowanego w taki sposób, aby w środku nocy przestawiał się on na wskazywanie godziny o 10 minut późniejszej niż rzeczywista (McKay i Dennett, 2009). Budzik taki miałby motywować swojego posiadacza (przynajmniej do czasu, kiedy nie odkryje on zasady działania) do wstawania nieco wcześniej, dając mu więcej czasu na poranny „rozruch". Mechanizm taki ilustrowałby następującą możliwość:

(4) Mechanizm działa poprawnie, jego funkcją jest wytwarzanie reprezentacji, jednak poprawność jego funkcjonowania nie zależy od tego, czy tworzone reprezentacje są poprawne. Funkcją mechanizmu jest tworzenie reprezentacji o określonej treści, niezależnie od tego, czy są one poprawne.

Możemy teraz powiązać tę dyskusję z pojęciem koherencji i zasadą KDP. Weźmy pod uwagę pięć możliwości:

(1-K) Mechanizm ma funkcję wytwarzania poprawnych reprezentacji (lub realizacja jego funkcji systematycznie zależy od generowania poprawnych reprezentacji[9]) i do realizacji tej funkcji przyczynia się maksymalizacja koherencji.

(2-K) Mechanizm ma funkcję wytwarzania poprawnych reprezentacji (lub realizacja jego funkcji systematycznie zależy od generowania poprawnych reprezentacji), do realizacji tej funkcji normalnie przyczynia się maksymalizacja koherencji, jednak działanie mechanizmu jest zaburzone i działa on niepoprawnie, przez co utrzymywanie koherencji *de facto* przyczynia się do wytwarzania reprezentacji błędnych[10].

(3-K) Mechanizm ma funkcję wytwarzania poprawnych reprezentacji (lub realizacja jego funkcji systematycznie zależy od generowania poprawnych reprezentacji), do realizacji tej funkcji normalnie przyczynia się maksymalizacja koherencji. Czasem jednak maksymalizacja spójności prowadzi do wytwarzania błędnych reprezentacji. Popełniane w ten sposób błędy są wybaczalne w sensie McKaya i Dennetta.

(4-K) Mechanizm ma funkcję wytwarzania poprawnych reprezentacji (lub realizacja jego funkcji systematycznie zależy od generowania poprawnych reprezentacji), jednak proces maksymalizacji koherencji prowadzi mechanizm do wytwarzania błędnych reprezentacji, czyli do działania niezgodnie z jego funkcją. Błędy te nie są wybaczalne w sensie McKaya i Dennetta.

(5-K) Mechanizm ma funkcję wytwarzania reprezentacji o określonej treści (lub realizacja jego funkcji systematycznie zależy od generowania reprezentacji o określonej treści), jednak realizacja tej funkcji nie zależy w żaden sposób od tego, czy reprezentacje są poprawne. Proces maksymalizacji koherencji przyczynia się do nabywania i utrzymywania reprezentacji o określonej treści, niezależnie od tego, czy są one poprawne.

Zwróćmy uwagę, że tylko mechanizm opisany w punkcie (4-K) stanowiłby jasny kontrprzykład dla zasady KDP. Jedynie w tym przypadku mechanizm zajmuje się (zgodnie ze swoją funkcją czy projektem) wytwarzaniem poprawnych reprezentacji, a zarazem proces nadawania spójności tym reprezentacjom powoduje błędy. Inaczej mówiąc, kategoria (4-K) obejmuje sytuacje, w których poszukiwanie spójności koliduje z poszukiwaniem prawdy.

[9] Na przykład, chociaż funkcją mechanizmu wytwarzania map przestrzennych jest umożliwianie nawigacji przestrzennej, poprawna realizacja tej funkcji systematycznie zależy od poprawności map (por. Gładziejewski i Miłkowski, 2017).

[10] Przykładem należącym do tej kategorii mogą być wspomniane w poprzedniej sekcji spójne systemy urojeniowe, towarzyszące niektórym zaburzeniom psychicznym.

Kluczowym punktem mojej obrony zasady KDP jest obserwacja, że żaden potencjalny kontrprzykład omówiony w sekcji 3. nie należy do kategorii (4-K). Zatem żaden z omówionych tam przypadków nie zagraża zasadzie KDP.

Omówię teraz krótko przykłady omówione w sekcji 3. w kontekście pięciu wskazanych wyżej możliwości. Spróbuję pokazać, że wszystkie należą do kategorii (3-K) lub (5-K). Są to zatem przypadki, gdzie wynikające z nadawania spójności reprezentacjom błędy są wybaczalne z perspektywy funkcji czy projektu mechanizmu lub przypadki, gdzie błędy wynikają z faktu, że dany mechanizm w ogóle nie zajmuje się wytwarzaniem poprawnych reprezentacji. Ta druga ewentualność nie zagraża zasadzie KDP, ponieważ zasada ta nie przeczy twierdzeniu, że koherencja przynajmniej czasem może być (skutecznie) wykorzystywana w *innych celach* niż tworzenie poprawnych reprezentacji[11].

Błędne wnioskowania percepcyjne należą do kategorii (3-K). Do natury wnioskowania należy możliwość dojścia do fałszywego wniosku jeśli wychodzimy z fałszywych przesłanek. Omówione przykłady iluzji wzrokowych obejmują przypadki, gdzie system poznawczy dostaje na wejściu dane odbiegające od zwykłych statystycznych wzorców, na których oparta jest jego wiedza uprzednia. Jeśli percepcja jest nieuświadamianym wnioskowaniem probabilistycznym, to naturalną, oczekiwaną konsekwencją tego faktu będzie generowanie błędnych reprezentacji percepcyjnych pod wpływem mylących danych wejściowych. Takie błędy są niewinną konsekwencją inferencyjnego charakteru mechanizmu percepcji.

Błędne oszacowania precyzji sygnałów proprioceptywnych w inicjacji ruchu należą do kategorii (5-K). Na gruncie koncepcji przetwarzania predykcyjnego od systematycznej błędności tych reprezentacji zależy powodzenie procesu kontroli ruchu. Inaczej mówiąc, aby zainicjować ruch w ramach wnioskowania aktywnego, system poznawczy *musi* obniżyć szacowaną precyzję bieżących sygnałów proprioceptywnych dotyczących układu przestrzennego ciała, niezależnie od rzeczywistej precyzji tych sygnałów (Wiese 2016). Jest to zatem wykorzystanie spójności w celu innym niż tworzenie poprawnych reprezentacji świata.

Błędne wspomnienia epizodyczne mieszczą się w kategorii (3-K). Współczesne koncepcje pamięci epizodycznej na ogół zakładają, że jest ona tylko aspektem szerszego mechanizmu „mentalnej podróży w czasie". Funkcją tego mechanizmu jest przewidywanie przyszłych lub kontrfaktycznych sekwencji wydarzeń (De Brigard, 2014; Conway i Loveday, 2015; Schacter, Benoit, De Brigard i Szpunar, 2015). Wspominanie epizodyczne jest wykorzystaniem tego mechanizmu w celu innym niż pierwotny, mianowicie aby skonstruować wyobrażenie o przeszłym wydarzeniu. Mechanizm pamięci epizodycznej tworzy scenariusze przeszłości, które są najbardziej prawdopodobne w świetle posiadanej wiedzy, a zatem niekoniecznie takie, które się wydarzyły. Błędne wspomnienia są produktem ubocznym tego procesu. Są one zatem wybaczalnym skutkiem działania mechanizmu, którego zadanie polega na wyobrażaniu przyszłości i wydarzeń kontrfaktycznych.

[11] Analogicznie, choć skarpetka to dobre narzędzie ogrzewania stopy, może ona zostać z powodzeniem nałożona na dłoń i użyta jako pacynka.

Kwestia kategoryzacji fałszywych przekonań powstających lub utrzymywanych w wyniku rozwiązywania dysonansu poznawczego jest najbardziej problematyczna. Chcę jednak roboczo zaproponować, że mamy tu do czynienia z przykładem należącym do (5-K). To znaczy sugeruję, że błędy te powstają w wyniku działania mechanizmu, którego funkcją jest wytwarzanie określonych reprezentacji (przekonań), niezależnie od ich poprawności (prawdziwości). Nie jest obecnie zbyt kontrowersyjne twierdzenie, że ludzkie sposoby rozumowania i aktualizacji przekonań przynajmniej częściowo były kształtowane przez presje adaptacyjne związane z życiem społecznym. Oznacza to, że ścieżki ludzkiego myślenia są często wyznaczane przez to, co jest społecznie „strategiczne", a niekoniecznie prawdziwe lub uzasadnione epistemicznie (por. Williams, w druku). W szczególności wydaje się, że ludzie posiadają mechanizmy pozwalające im podtrzymywać wyolbrzymiony obraz swojej własnej wartości (von Hippel i Trivers, 2011) oraz utrzymywać konsekwentne i przewidywalne wzorce zachowania w ramach interakcji społecznych (por. Zawidzki, 2013). Mechanizmy te „nie troszczą" się o to, czy powstające systemy przekonań są prawdziwe. Proponuję przyjąć, że sposoby rozwiązywania dysonansu poznawczego są efektem działania takich mechanizmów. To znaczy są one próbą ochrony przekonań na temat własnej wartości (inteligencji, atrakcyjności itd.) przed dowodami świadczącymi przeciwko tym przekonaniom. Zarazem mają one „chronić" przed zbyt częstą zmianą zdań, tak by zachowanie było stabilne i konsekwentne, a przez to przewidywalne dla innych członków grupy. Kiedy rozwiązujemy dysonans poznawczy, za wszelką cenę broniąc spójności naszych przekonań, to funkcją tego procesu nie jest formowanie przekonań prawdziwych.

5. Podsumowanie

W artykule tym podjąłem próbę uogólnienia i rozszerzenia koherencyjnej koncepcji rozpoznawania błędu reprezentacyjnego Krystyny Bieleckiej. Pokazałem, w jakim sensie rola odgrywana przez koherencję w procesie rewizji reprezentacji może wykraczać poza „odfiltrowywanie" reprezentacji fałszywych. Wskazałem też kilka przekornych przykładów zjawisk poznawczych, w których poszukiwanie spójności systematycznie prowadzi do tworzenia błędnych reprezentacji. Argumentowałem jednak, że nie są to dobre kontrprzykłady dla idei, iż koherencja jest potrzebna systemom poznawczym jako drogowskaz prawdy. Jeśli naszym celem są poprawne reprezentacje, to troszczenie się o spójność jest drogą do celu; a przynajmniej jedną z dróg.

Finansowanie

Pracę nad tym artykułem sfinansowano z grantu NCN Opus-17 nr 2019/33/B/HS1/00677.

Podziękowania

Dziękuję Krystynie Bieleckiej za dyskusję dotyczącą wczesnej wersji tego komentarza, która odbyła się w IFiS PAN w maju 2019, w ramach sympozjum poświęconego książce *Błądzę więc myślę*.

Bibliografia

Adams, R. A., Shipp, S., Friston, K. J. (2013). Predictions not commands: Active inference in the motor system. *Brain Structure and Function*, 218, 611–643.

Anderson, J. R., Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science, 2,* 396–408.

Batson, C.D. (2008). Moral masquerades: Experimental exploration of the nature of moral motivation. *Phenomenology and the Cognitive Sciences*, 7, 51–66.

Bickhard, M. H. (1999). Interaction and representation. *Theory and Psychology, 9*, 435–458.

Bickhard, M. H. (2004). The dynamic emergence of representation. W: H. Clapin, P. Staines, P. Slezak (red.). *Representation in mind: New approaches to mental representation* (71–90). Oxford: Elsevier Science.

Bielecka, K. (2019). *Błądzę, więc myślę. Co to jest błędna reprezentacja?* Warszawa: UW.

BonJour, L. (1985). *The Structure of Empirical Knowledge*. Cambridge (MA): Harvard University Press.

Carpenter, A., Schacter D. (2016). Flexible retrieval: When true inferences produce false memories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 43, 335–349.

Chemero, A. (2014). Antyreprezentacjonizm i nastawienie dynamiczne. *Przegląd Filozoficzno-Literacki*, 39, 79–107.

Conway, M.A., Loveday, C. (2015). Remembering, imagining, false memories & personal meanings. *Consciousness and Cognition*, 33, 574–581.

Daw, N. D., O'Doherty, J.P., Dayan, P., Seymour, B., Dolan, R.J. (2006). Cortical substrates for exploratory decisions in humans. *Nature*, 441, 876–879.

De Brigard, F. (2014). Is memory for remembering? Recollection as a form episodic hypothetical thinking. *Synthese*, 191, 155–185.

Ditto, H., Munro, G.D., Apanovitch, A.M., Scepansky, J.A., Lockhard, L.K. (2003). Spontaneous skepticism: The interplay of motivation and expectation in responses to favorable and unfavorable medical diagnoses. *Personality and Social Psychology Bulletin*, 29, 1120–1132.

Friston, K. J. (2018). Active inference and cognitive consistency. *Psychological Inquiry*, 29, 67–73.

Geisler, W. S., Kesler, D. (2002). Illusion, perception and Bayes. *Nature neuroscience*, 5, 508–510.

Gładziejewki, P. (2015). *Wyjaśnianie za pomocą reprezentacji mentalnych*. Toruń: UMK (FNP).

Gładziejewski, P. (2017). Evidence of the senses: A Predictive Processing-based take on the Sellarsian dilemma. W: T. Metzinger, W. Wiese (red.). *Philosophy and Predictive Processing*. MIND Group.

Gładziejewski, P., Miłkowski, M. (2017). Structural representations: causally relevant and different from detectors. *Biology & Philosophy*, 32: 337–355.

Goldman, A. (1986). *Epistemology and Cognition*. Cambridge (MA): Harvard University Press.

Gregory, R. L. (1963). Distortion of visual space as inappropriate constancy scaling. *Nature*, 199, 678–680.

Gregory, R. L. (1980). Perceptions as hypotheses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 290, 181–97.

Gregory, R. L. (1997). Knowledge in perception and illusion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 352, 1121–7.

Haack, S. (1993). *Evidence and Inquiry: Towards Reconstruction in Epistemology.* Oxford: Blackwell.

Harman, G. (1973). *Thought.* Princeton: Princeton University Press.

Harmon-Jones, E., Mills, J. (1999). An introduction to cognitive dissonance theory and an overview of current perspectives on the theory. W: E. Harmon-Jones, J. Mills (red.). *Science conference series. Cognitive dissonance: Progress on a pivotal theory in social psychology*. Washington: American Psychological Association.

Hemmer, P., Steyvers, M. (2009). A Bayesian account of reconstructive memory. *Topics in Cognitive Science*, 1, 189–202.

Hohwy, J. (2013). *The Predictive Mind*. Oxford: Oxford University Press.

Hohwy, J., Roestorff, A, Friston, K. (2008). Predictive coding explains binocular rivarly: An epistemological review. *Cognition*, 108, 687–701.

Hutto, D., Myin, E. (2012). *Radicalizing Enactivism: Basic Minds without Content*. Cambridge (MA): The MIT Press.

Jagodzińska, M. (2008). *Psychologia pamięci. Badania, teorie, zastosowania*. Warszawa: Helion.

Kiefer, A. (2017). Literal perceptual inference. W: T. Metzinger, W. Wiese (red.). *Philosophy and Predictive Processing*. MIND Group.

Magnotti, J. F., Beauchamp, M. S. (2017). A causal inference model explains perception of the McGurk Effect and other incongruent audiovisual speech. *PLoS Computational Biology*, 13, e1005229.

McKay, R. T., Dennett, D. (2009). The evolution of misbelief. *Behavioral and Brain Sciences*, 32, 493–510.

Olsson, E. (2003). Coherentist theories of epistemic justification. W: E. Zalta (red.). *Stanford Encyclopedia of Philosophy*. Dostępne 03.12.2019 na stronie: https://plato.stanford.edu/entries/justep-coherence/.

Quine, W.v.O. (2000). Dwa dogmaty empiryzmu. W: Tegoż. *Z punktu widzenia logiki: dziewięć esejów logiczno-filozoficznych*. Przeł. Barbara Stanosz. Warszawa: Aletheia.

Rescorla, M. (2015). Bayesian perceptual psychology. W: M. Matthen (red.). *The Oxford Handbook of Philosophy of Perception*. Oxford: Oxford University Press.

Saucier, D. A., Miller, C. T., Doucet, N. (2005). Differences in helping whites and blacks: a meta-analysis. *Personality and Social Psychology Review*, 9, 2–16.

Schacter, D. (2001). *Siedem grzechów pamięci*. Warszawa: PIW.

Schacter, D. Benoit, R.G., De Brigard, F., Szpunar, K.K. (2015). Episodic future thinking and episodic counterfactual thinking: Intersections between memory and decisions. *Neurobiology of Learning and Memory*, 117, 14–21.

Thagard, P. (2000). *Coherence in Thought and Action*. Cambridge (MA): The MIT Press.

Travis, C., Aronson, E. (2008). *Błądzą wszyscy (ale nie ja). Dlaczego usprawiedliwiamy głupie poglądy, złe decyzje i szkodliwe działania.* Warszawa: Smak Słowa.

Ullman, T. Spelke, E., Battaglia, P., Tenenbaum, J. (2017). Mind games: Game engines as an architecture for intuitive physics. *Topics in Cognitive Sciences*, 21, 649–665.

von Hippel, W., Trivers, R. (2011). The evolution and psychology of self-deception. *Behavioral and Brain Sciences*, 34, 1–16.

Whitson, J.A., Galinsky, A.D. (2008). Lacking control increases illusory pattern perception. *Science*, 322, 115–117.

Wiese, W. (2016). Action is enabled by systematic misrepresentations. *Erkentnnis*, 82, 1233–1252.

Williams, D. (w druku). Socially adaptive belief. *Mind and Language*.

Zawidzki, T. (2013). *Mindshaping.* Cambridge (MA): MIT Press.

Is coherence a guide to truth? Coherence and false representations

Abstract: This paper is a commentary on Krystyna Bielecka's book "I err, therefore I think. What is misrepresentation?". I build on Bielecka's coherentist account of detectable representational error in order to draw a wider picture of the role of coherence in evaluating representations. In particular, I explore the idea that the strive to maximize coherence is a major factor guiding the acquisition and revision of internal representations, whose role extends beyond error detection. The idea critically discussed in the paper is that coherence owes this role to its being truth-conductive, i.e. to the (alleged) fact that coherence tends to increase the truth-ratio of representations. I draw from historical and recent work in cognitive science to discuss a number cases where coherence maximization systematically produces misrepresentations. I argue that despite the appearances, these cases do not constitute proper counterexamples to the idea that coherence is, all else being equal, a guide to truth.

Keywords: epistemology; coherence; episodic memory; predictive processing; mental representations; reprezentacjonizm

Paweł Gładziejewski jest adiunktem w Katedrze Kognitywistyki UMK w Toruniu. Interesuje się epistemologią percepcji w świetle teorii kognitywistycznych, rolą eksplanacyjną reprezentacji mentalnych oraz naturą pamięci epizodycznej i jej związkiem z poczuciem tożsamości osobowej.

Redakcję i publikację tekstu sfinansowano ze środków Ministerstwa Nauki i Szkolnictwa Wyższego na działalność upowszechniającą naukę (DUN), działalność wydawnicza, nr umowy: 711/P-DUN/2019, okres realizacji: 2019–2020.