Filozofia Nauki
Rok XXIII, 2015, Nr 4(92)

Cezary Cieśliński

Problemy minimalizmu*

Zgodnie z doktryną minimalizmu (Horwich 1998) wszystkie fakty dotyczące prawdy można wyjaśnić za pomocą „teorii minimalnej" (MT). Aksjomaty tej teorii to tzw. równoważności Tarskiego, czyli wyrażenia o postaci:

(T) $\langle p \rangle$ jest prawdą wtedy i tylko wtedy, gdy p,

gdzie wyrażenie „$\langle p \rangle$" odczytujemy jako „sąd, że p". Horwich twierdzi, że teoria minimalna wyczerpująco charakteryzuje treść pojęcia prawdy, a jego rozumienie polega na dyspozycji do uznawania wszystkich nieparadoksalnych podstawień schematu (T) (Horwich 1998: 35). W rezultacie pojęcie prawdy staje się proste i nieproblematyczne, bo czyż równoważności Tarskiego nie są oczywistością i banałem? Z tego właśnie względu minimaliści twierdzą, że prawda nie ma żadnej głębszej natury, której odkrycie miałoby być zadaniem filozofów (Horwich 1998: 2).

Jedną z głównych trudności Horwichowskiego minimalizmu jest tzw. problem generalizacji: w jaki sposób minimalista może wyjaśnić ogólne zasady sformułowane przy użyciu pojęcia prawdy?[1] Dla przykładu rozważmy następujące uogólnienia:

* Artykuł powstał w ramach realizacji grantu Narodowego Centrum Nauki nr 2014/13/B/HS1/02892.

[1] W dyskusjach dotyczących minimalizmu zarzut ten pojawia się po raz pierwszy w artykule Gupty (1993). Już znacznie wcześniej zauważano jednak dedukcyjną słabość pewnych teorii prawdy aksjomatyzowanych za pomocą podstawień schematu „$T(\varphi) \equiv \varphi$". Pierwsze znane mi uwagi krytyczne pod adresem takich teorii aksjomatycznych pochodzą z klasycznej pracy Tarskiego (1933). Po sformułowaniu Twierdzenia III (o niesprzeczności tego rodzaju teorii prawdy) Tarski pisze: „Wartość uzyskanego wyniku osłabia znacznie ta okoliczność, że aksjomaty, wskazane w tw. III, mają bardzo małą siłę dedukcyjną: ugruntowana na nich teoria prawdy byłaby systemem wysoce niezupełnym, pozbawionym najważniejszych i najpłodniejszych praw natury ogólnej" (1933: 105).

(1) Każdy sąd o formie „$p \rightarrow p$” jest prawdziwy.

(2) Dla dowolnego sądu φ, negacja φ jest prawdziwa wtedy i tylko wtedy, gdy φ nie jest prawdą.

(3) Dla dowolnej formuły φ, φ jest prawdziwa o pewnym przedmiocie wtedy i tylko wtedy, gdy jest prawdą, że dla pewnego x, $\varphi(x)$.

Czy teoria minimalna Horwicha, przyjmująca jako aksjomaty wyłącznie podstawienia (T), dowodzi tego rodzaju uogólnień? Nie sposób udzielić tu definitywnej odpowiedzi, ponieważ Horwich nigdy nie podał precyzyjnej definicji klasy aksjomatów MT[2]. Z jednej strony, wiemy z pewnością, że pewne znane teorie aksjomatyzowane za pomocą T-równoważności nie dowodzą *żadnych* interesujących teorioprawdziwościowych uogólnień[3]. Z drugiej strony, jak zauważył Vann McGee, każde zdanie języka z predykatem prawdziwości jest dowodliwie równoważne (na gruncie odpowiednio silnej teorii) pewnej T-równoważności. Obowiązuje bowiem następujące twierdzenie (McGee 1992: 237-238):

> **Twierdzenie**. Niech S będzie odpowiednio silną teorią arytmetyczną w języku arytmetyki pierwszego rzędu[4]. Niech ST będzie rozszerzeniem S o aksjomaty logiczne w języku L_T, który powstaje z języka arytmetyki przez wprowadzenie nowego jednoargumentowego predykatu „T”. Wtedy:
>
> $$\forall \varphi \in L_T \exists \psi \in L_T \, ST \vdash \varphi \equiv [T(\psi) \equiv \psi].$$

W rezultacie jest możliwe, że MT — z odpowiednio dobraną aksjomatyką — *będzie* dowodzić uogólnień (1)-(3). Słaba to jednak pociecha: wybór T-równoważności w roli aksjomatów motywowany był przecież początkowo ich prostotą i oczywistością! Twierdzenie McGee pokazuje jednak, że nie wszystkie T-równoważności są oczywiste i proste. Podkreślmy raz jeszcze, że *każde zdanie* jest dowodliwie równoważne pewnej T-równoważności. Dotyczy to więc również zdań fałszywych, zdań sprzecznych, a także zdań prawdziwych wyrażających nieznane nam fakty arytmetyczne. Dlaczego jednak T-równoważności dowodliwie równoważne wybranym zdaniom ogólnym (takim jak (1)-(3)) miałyby akurat należeć do kategorii „prostych i oczywistych”? O ile mi wiadomo, nikt do tej pory nie podał przekonującej odpowiedzi na to pytanie.

---

[2] Jak stwierdził, aksjomatyka ma obejmować „nieparadoksalne” T-równoważności. Jest to jednak tylko intuicja i trudno powiedzieć, co dokładnie miałoby to znaczyć. Wiemy w szczególności, że nie wystarczy tu warunek maksymalnej niesprzeczności zbioru T-równoważności, ponieważ dla przeliczalnych języków takich maksymalnych niesprzecznych zbiorów będzie continuum wiele. Zob. McGee 1992, Cieśliński 2007.

[3] Za przykład może posłużyć teoria TB, którą zdefiniuję w dalszej części artykułu. Słabość TB pod tym względem jest dobrze znana — zob. Cieśliński 2009: 93.

[4] W zupełności wystarczy, by S zawierała arytmetykę Peana.

W tej sytuacji Horwich jest nam winien wyjaśnienie, z jakiego właściwie powodu przyjmujemy zasady typu (1)-(3). Jeśli teoria minimalna nie dowodzi tych ogólnych zasad — co, jak widzieliśmy, nie jest bynajmniej wykluczone — to jak wytłumaczyć fakt, że je uznajemy? Na tym właśnie polega problem generalizacji.

W ostatnich latach Horwich podjął dwie próby poradzenia sobie z tą trudnością. Omówię je w dalszej części artykułu.

* * *

Pierwsza próba. Praca (Horwich 1998) zawiera propozycję wzmocnienia teorii minimalnej w taki sposób, by *dowodziła* ona pożądanych uogólnień. Wzmocnienie to nie polega na wzbogaceniu MT o nowe aksjomaty, lecz na zmodyfikowaniu technik dowodzenia dostępnych w MT. Horwich pisze:

> Wydaje się wiarygodne […] że istnieje reguła wnioskowania, która zachowuje prawdziwość i prowadzi nas od zbioru przesłanek przypisujących każdemu sądowi pewną własność F do wniosku, zgodnie z którym wszystkie sądy mają własność F (Horwich 1998: 137).

Wygląda na to, że autor proponuje tu wprowadzenie do teorii MT jako części jej dedukcyjnego aparatu zasady przypominającej tzw. ω-regułę. W standardowym sformułowaniu, jeśli potrafimy udowodnić $\varphi(n)$ o każdej liczbie naturalnej n z osobna, to na mocy ω-reguły wolno nam uznać zdanie ogólne $\forall x\, \varphi(x)$. Mamy tu do czynienia z podobnym pomysłem: jeśli dla każdego sądu α teoria MT dowodzi $F(\alpha)$, to pracując w MT, mamy prawo wywnioskować, że $\forall \alpha\, F(\alpha)$.

Nie zamierzam tu opisywać dokładniej tej strategii, ograniczę się tylko do dwóch uwag. Po pierwsze, Horwich ma rację: rzeczywiście, takie posunięcie pozwala uzyskać bardzo silną teorię, a pożądane uogólnienia stają się wówczas twierdzeniami MT. Już dodanie ω-reguły do samej tylko arytmetyki Peana prowadzi do powstania silnej teorii PA^{ω}, dowodzącej wszystkich arytmetycznych zdań prawdziwych w standardowym modelu arytmetyki. Warto podkreślić, że nie obala to klasycznego twierdzenia Tarskiego o niedefiniowalności prawdy: prawda wciąż pozostaje arytmetycznie niedefiniowalna, otrzymujemy co najwyżej wniosek, że dowodliwość w PA^{ω} nie jest arytmetycznie definiowalna[5].

Po drugie, pomysł wykorzystania ω-reguły do poradzenia sobie z problemem generalizacji został w literaturze poddany surowej krytyce (Raatikeinen 2005). W obli-

[5] Na marginesie zauważmy, że zwolennik Horwichowskiego minimalizmu nie ma najmniejszego kłopotu ze sformułowaniem i udowodnieniem (w zupełnie zwykły sposób) twierdzenia Tarskiego. W szczególności, całkiem nieproblematyczna jest dla niego syntaktyczna wersja tego twierdzenia: „nie istnieje niesprzeczna teoria *S* zawierająca arytmetykę Peana, która dla pewnej arytmetycznej formuły $T(x)$ dowodziłaby wszystkich równoważności o postaci »$T(\psi) \equiv \psi$«". Również wersja semantyczna („Nie ma modelu *M* arytmetyki Peana oraz arytmetycznej formuły $T(x)$ takich, że dla dowolnego arytmetycznego zdania ψ, $M \vDash T(\psi) \equiv \psi$") nie jest kłopotliwa dla minimalisty, którego filozoficzne tezy dotyczą pojęcia *prawdy*, a nie „prawdy w modelu". Minimalista ma pełne prawo uznać to ostatnie pojęcie za użyteczne techniczne narzędzie. Zob. Cieśliński 2015: 72-73.

czu poważnych zarzutów propozycji tej nie podtrzymuje obecnie nawet jej autor i ma ona, jak sądzę, jedynie charakter historycznej ciekawostki. Znacznie bardziej interesująca wydaje się druga strategia, również zaproponowana przez Horwicha. Jej analiza to główne zadanie tego artykułu.

Druga próba. Inaczej niż poprzednio, kolejna propozycja pozostawia nienaruszoną teoriodowodową maszynerię MT (jest to po prostu aparatura logiki klasycznej). Tym razem proponuje się użycie MT razem z pewną dodatkową przesłanką. Horwich bardzo słusznie podkreśla, że minimalista może wykorzystywać w wyjaśnieniach nie tylko aksjomaty MT, lecz także dodatkowe założenia zaczerpnięte z innych teorii. Dla przykładu zastanówmy się, dlaczego uznajemy, że „Konie są ssakami" jest prawdą. Odpowiedź uzyskamy, posługując się teorią MT i korzystając z dodatkowej informacji biologicznej, że konie są ssakami. Nasze zrozumienie pojęcia prawdy — czyli gotowość do akceptacji T-równoważności tworzących aksjomaty MT — wciąż jest istotna, ale nie zachodzi w izolacji: ważna jest tu również znajomość dodatkowych faktów, na przykład biologicznych. Ostatecznie więc odpowiedź brzmi: uznajemy prawdziwość sądu, że konie są ssakami, ponieważ uważamy, że konie są ssakami (biologiczny fakt) oraz przyjmujemy odpowiedni aksjomat MT.

W celu wyjaśnienia, dlaczego uznajemy uogólnienia typu (1)-(3), Horwich próbuje posłużyć się analogiczną strategią[6]. Proponuje skorzystać z następującej dodatkowej przesłanki (2010: 45):

(A) Ilekroć dla każdego sądu o strukturze F ktoś ma dyspozycję do uznania, że jest on G (a przy tym czyni to za każdym razem z tego samego powodu), tylekroć będzie miał dyspozycję do uznania, że każdy F-sąd jest G.

Dzięki przesłance (A) Horwich jest w stanie wyjaśnić, dlaczego mamy skłonność uznawać generalizacje o postaci „Każdy F-sąd jest prawdziwy" (np. „Każdy sąd o formie »$p \rightarrow p$« jest prawdziwy"). Wyjaśnienie przebiega następująco:

Wyjaśnienie 1

P_1: Dla każdego F-sądu γ, mamy dyspozycję do uznania prawdziwości γ (a przy tym czynimy to za każdym razem z tego samego powodu).

P_2: Jeśli P_1, to na mocy (A) będziemy skłonni uznać, że każdy F-sąd jest prawdziwy.

Wniosek: będziemy skłonni uznać, że każdy F-sąd jest prawdziwy.

Skoro rozumowanie to jest logicznie poprawne, przyjrzyjmy się jego przesłankom. P_1 to założenie dotyczące faktów — przyjmujemy, że rzeczywiście opisuje nas ono

[6] Praca (Horwich 2001) zawiera pierwsze sformułowanie tej idei. Zob. również Tennant 2002, gdzie zaproponowane jest nieco inne, lecz pod wieloma względami podobne podejście. Zob. też Cieśliński 2010, Ketland 2010, Tennant 2010.

jako użytkowników MT. Inaczej mówiąc, zakładamy tu, że F-sądy mają podaną własność — to przecież właśnie dla takiego przypadku chcemy uzyskać końcowy wniosek. W takim razie do rozważenia pozostaje przesłanka P_2, którą możemy uznać za podstawienie ogólnej zasady (A). Czy P_2 jest prawdziwa?

Bradley Armour-Garb zgłosił tu zastrzeżenia. Zauważył, że:

> nie będziemy mieć dyspozycji do uznania (sądu), że wszystkie F-sądy są G, na podstawie faktu, że dla dowolnego F-sądu jesteśmy skłonni uznać, że jest on G […] chyba że jesteśmy *świadomi* faktu, że dla dowolnego F-sądu jesteśmy skłonni uznać, że jest on G (Armour-Garb 2010: 699).

Zarzut ten wydaje się słuszny. Rzeczywiście, nie widać powodu, by dyspozycji do uznania prawdziwości wszystkich sądów danej postaci (każdego z osobna!) towarzyszyła dyspozycja do uznania prawdziwości zdania ogólnego: mogłoby się przecież zdarzyć, że dana osoba nie zdaje sobie sprawy z posiadania pierwszej z wymienionych dyspozycji.

W kolejnym kroku Armour-Garb zauważa, że Horwich mógłby wziąć pod uwagę wspomniane zastrzeżenie i zmodyfikować swoje wyjaśnienie w następujący sposób:

Wyjaśnienie 2

R_1:	Dla każdego F-sądu γ, mamy dyspozycję do uznania prawdziwości γ [R_1 nie różni się od P_1].
R_2:	Jesteśmy świadomi tego, że R_1.
R_3:	Jeśli R_1 i R_2, to będziemy skłonni uznać, że każdy F-sąd jest prawdziwy.
Wniosek:	będziemy skłonni uznać, że każdy F-sąd jest prawdziwy.

Armour-Garb uważa jednak, że R_2 jest problematyczne. Cóż to bowiem znaczy „być świadomym" takiego faktu?

> Oto wiarygodna odpowiedź: być świadomym faktu, że dla każdego F-sądu mamy dyspozycję do zaakceptowania, że jest on prawdziwy, to tyle co być świadomym faktu, że mamy dyspozycję do uznania, że każdy F-sąd jest prawdziwy (Armour-Garb 2010: 700).

Na tej podstawie autor stwierdza, że *Wyjaśnienie 2* jest nie do przyjęcia: w przesłance R_2 stwierdzamy bowiem w istocie treść wniosku i w ten sposób zakładamy to, co powinno zostać dowiedzione.

Krytyka Armour-Garba jest moim zdaniem chybiona, a jego interpretacja przesłanki R_2 zupełnie nieprzekonująca. W dalszej części artykułu przedstawię alternatywne (i według mnie lepsze) ujęcie. Naszkicuję również program badawczy, którego realizacja wydaje się warunkiem powodzenia strategii Horwicha.

Rozważę teraz tę strategię na modelowym ograniczonym przykładzie teorii prawdy dla języka arytmetyki. Takie ograniczenie ma zarówno zalety, jak i wady. Niewątpliwą zaletą jest konkretność ujęcia: jak już wspominałem, nie wiadomo dokładnie, *które* T-równoważności tworzą kolekcję aksjomatów MT. Tu będziemy zaś pracować z precyzyjnie scharakteryzowanymi teoriami. Natomiast po stronie mankamentów należy odnotować fakt, że nie wiadomo z góry, czy i w jaki sposób wyni-

ki dotyczące konkretnego modelowego przykładu dają się uogólnić, tak by stosowały się również do innych teorii prawdy.

* * *

Niech L_{PA} będzie językiem arytmetyki dodawania i mnożenia, a L_T rozszerzeniem L_{PA} o jednoargumentowy predykat T. Dalej określimy teorię TB[7], będącą w następującym sensie odpowiednikiem MT dla klasy zdań arytmetycznych: aksjomaty TB to wyłącznie równoważności Tarskiego dla zdań arytmetycznych[8].

Definicja. $\mathrm{TB} = \mathrm{PA} \cup \{T(\varphi) \equiv \varphi\colon \varphi \in L_{PA}\}$

TB jest teoriodowodowo słaba. Po pierwsze, nie dowodzi żadnych twierdzeń arytmetycznych, których nie dowodziłaby już sama arytmetyka Peana[9]. Po drugie, nie dowodzi również wielu podstawowych zasad ogólnych odwołujących się do pojęcia prawdy. Dla ilustracji:

Fakt

(a) $\mathrm{TB} \nvdash \forall \varphi \in L_{PA}\ T(\varphi \rightarrow \varphi)$

(b) $\mathrm{TB} \nvdash \forall \varphi \in L_{PA}\ [T(\neg\varphi) \equiv \neg T(\varphi)]$

(c) $\mathrm{TB} \nvdash \forall \varphi \in L_{PA}\ [\exists x\ T(\varphi(x)) \equiv T(\exists x\ \varphi(x))]$

Załóżmy, że przyjmujemy TB jako teorię prawdy arytmetycznej. Ze względu na teorioprawdziwościową słabość TB staniemy wówczas przed problemem generalizacji. Dla przykładu: jesteśmy skłonni uznać, że dla wszystkich arytmetycznych φ, zdanie o postaci „$\varphi \rightarrow \varphi$” jest prawdziwe. W jaki sposób mamy wyjaśnić ten fakt za pomocą TB, skoro TB tego nie dowodzi?

Przedstawię wytłumaczenie zgodne ze strategią Horwicha, ale uwzględnię również uwagę Armour-Garba. Wyjaśnienie przeprowadzamy w metateorii, co do której przyjmujemy następujące założenia:

I. Język metateorii pozwala mówić o dyspozycjach do uznania zdań. Język ten zawiera również predykat „jesteśmy świadomi, że…”, orzekany o zdaniach języka metateorii.

II. Metateoria obejmuje informację, że TB jest akceptowaną przez nas teorią.

III. Metateoria zawiera arytmetykę Peana.

---

[7] Skrót pochodzi od określenia „Tarski biconditionals”.

[8] W literaturze używa się czasem dwóch oznaczeń: TB oraz TB$^-$, gdzie TB (w odróżnieniu od TB$^-$) ma dodatkowo zawierać wszystkie aksjomaty indukcji w języku z predykatem prawdziwości. W tym artykule pomijam tę subtelność.

[9] Inaczej mówiąc, TB jest konserwatywnym rozszerzeniem PA. Zob. Cieśliński 2009: 99.

IV. Oprócz zasad logiki klasycznej metateoria zawiera dwie dodatkowe reguły wnioskowania:

- regułę NEC: jeśli w metateorii udowodnimy φ, to wolno nam dopisać do dowodu „jesteśmy świadomi, że φ";
- regułę Horwicha: z informacji o postaci „jesteśmy świadomi, że dla każdego x mamy dyspozycję do uznania $A(x)$" wolno nam wywnioskować „mamy dyspozycję do uznania »dla każdego x, $A(x)$«".

Założenie IV kryje w sobie odpowiedź na pytanie Armour-Garba: w jakich okolicznościach wolno nam stwierdzić, że jesteśmy świadomi, iż φ? Otóż rozsądnym warunkiem wystarczającym wydaje się nasza umiejętność udowodnienia φ. Mając dane φ jako twierdzenie naszej metateorii, wolno nam wywnioskować: jesteśmy świadomi, że φ. Taki jest sens reguły NEC.

Dlaczego jesteśmy skłonni uznać, że każde zdanie arytmetyczne o postaci „$(\varphi \rightarrow \varphi)$" jest prawdziwe? Oto proponowane wyjaśnienie:

Wyjaśnienie 3

(1) Dla każdego zdania $\varphi \in L_T$, jeśli $TB \vdash \varphi$, to mamy dyspozycję do uznania φ.
(2) Dla każdego zdania $\varphi \in L_{PA}$, $TB \vdash T(\varphi \rightarrow \varphi)$.
(3) Dla każdego zdania $\varphi \in L_{PA}$ mamy dyspozycję do uznania „$T(\varphi \rightarrow \varphi)$".
(4) Jesteśmy świadomi tego, że: dla każdego zdania $\varphi \in L_{PA}$ mamy dyspozycję do uznania „$T(\varphi \rightarrow \varphi)$".

Wniosek: mamy dyspozycję do uznania „Dla każdego zdania $\varphi \in L_{PA}$, $T(\varphi \rightarrow \varphi)$".

Krok (1) jest uprawniony na mocy założenia II — przyjmujemy, że ta informacja należy do naszej metateorii. Skoro metateoria zawiera arytmetykę Peana, to zachodzi również (2)[10]. Krok (3) wynika z (1) i (2). Krok (4) to rezultat zastosowania NEC do (3). Wyciągamy wówczas wniosek na mocy reguły Horwicha zastosowanej do (4).

W analogiczny sposób potrafimy uporać się z wieloma innymi uogólnieniami. Dla ilustracji przedstawię wyjaśnienie dla przypadku (b) z podanego wcześniej *Faktu* charakteryzującego teoriodowodową słabość TB[11]. Dlaczego jesteśmy skłonni uznać, że dla każdej formuły $\varphi \in L_{PA}$ zachodzi: $T(\neg\varphi) \equiv \neg T(\varphi)$? Oto proponowane wyjaśnienie.

Wyjaśnienie 4

(1) Dla każdego zdania $\varphi \in L_T$, jeśli $TB \vdash \varphi$, to mamy dyspozycję do uznania φ.
(2) $\forall \varphi \in L_{PA}\ TB \vdash T(\neg\varphi) \equiv \neg T(\varphi)$.

[10] Chodzi o to, że potrafimy udowodnić (2) środkami arytmetyki Peana.
[11] W sprawie najbardziej złożonego przypadku (c) — zob. Appendix.

(3) Dla każdego zdania $\varphi \in L_{PA}$, mamy dyspozycję do uznania „$T(\neg\varphi) \equiv \neg T(\varphi)$”.

(4) Jesteśmy świadomi tego, że: dla każdego zdania $\varphi \in L_{PA}$, mamy dyspozycję do uznania „$T(\neg\varphi) \equiv \neg T(\varphi)$”.

Wniosek: mamy dyspozycję do uznania „Dla każdego zdania $\varphi \in \mathrm{L_{PA}}$, $T(\neg\varphi) \equiv \neg T(\varphi)$”.

Pomijam uzasadnienie poszczególnych kroków, ponieważ jest ono bardzo podobne do komentarza zamieszczonego pod *Wyjaśnieniem 3*.

Jak oceniać tego rodzaju wyjaśnienia? Czy proponowana strategia rzeczywiście pozwoli poradzić sobie z problemem generalizacji? Sformułuję dwie uwagi. Pierwsza z nich wskaże pewną szczególną cechę Horwichowskich wyjaśnień, którą uważam za pozytywną — jest to moim zdaniem ruch w dobrym kierunku. Druga będzie miała charakter krytyczny: sądzę, że podane wytłumaczenia wciąż wymagają przeformułowania.

Mają one pewną szczególną cechę, która odróżnia je zarówno od wyjaśnień wykorzystujących ω-regułę, jak i od sposobu, w jaki tłumaczyliśmy prawdziwość zdania „Konie są ssakami”. Za każdym razem nasze wyjaśnienie przyjmuje postać odpowiedzi na pytanie „Dlaczego jesteśmy skłonni uznać zdanie *A*?” (gdzie *A* to np. „Wszystkie podstawienia schematu $\varphi \rightarrow \varphi$ są prawdziwe” albo „Zdanie »Konie są ssakami« jest prawdziwe”). Jednakże w dwóch ostatnich przypadkach wyjaśnienie przebiega według następującego schematu:

— uznajemy teorię T,
— teoria T dowodzi *A*,
— z tego względu jesteśmy skłonni uznać *A*.

W wypadku strategii z ω-regułą teoria T to nic innego jak Horwichowska MT wzbogacona o ω-regułę: przyjmujemy prawdziwość wszystkich podstawień „$\varphi \rightarrow \varphi$”, ponieważ MT z ω-regułą dowodzi odpowiedniego zdania ogólnego. Z kolei prawdziwość zdania „Konie są ssakami” to konsekwencja MT wzbogaconej o dodatkową (biologiczną) informację, że konie są ssakami — akceptowana przez nas teoria T to MT $\cup$ {„Konie są ssakami”}. Tymczasem w wyjaśnieniach (1)-(4) odeszliśmy od tego wzorca: zdanie *A* nie jest po drodze udowodnione w *żadnej* uznawanej przez nas teorii; nie na tym polega to wyjaśnienie! Dowodzi się natomiast bezpośrednio (w odpowiedniej metateorii), że jesteśmy skłonni uznać *A*. Ten brak wyprowadzenia *A* z przyjmowanej przez nas teorii uważam za istotną, nową własność proponowanych tu wyjaśnień. Cecha ta sama w sobie nie jest kłopotliwa. Inne aspekty pozostają jednak problematyczne.

Chciałbym teraz wskazać dwie z tych trudności; napiszę również parę słów o możliwości ich przezwyciężenia.

Problem 1. Podane wyjaśnienia mają charakter psychologiczny. Tłumaczy się tu, dlaczego *jesteśmy skłonni* uznać pewne zdania (fakt dotyczący naszych umysłowych predyspozycji) za pomocą przesłanek i reguł charakteryzujących nasze *dyspozycje* (czyli elementy naszego psychicznego wyposażenia). Rzecz jasna, nie jest to problematyczne samo w sobie: nie ma nic złego w psychologicznych wyjaśnieniach jako takich. Kłopot w tym, że przy tego rodzaju tłumaczeniu umyka nam całkowicie element normatywny. Pojawia się następujące dodatkowe zagadnienie: czy ktoś, kto uznaje TB (albo Horwichowską MT), jest w jakikolwiek sposób *zobowiązany* do uznania generalizacji, w których występuje predykat prawdziwości? Załóżmy na przykład, że rzeczywiście mamy cechy przypisane nam w *Wyjaśnieniu 4*. W związku z tym, uznając TB (przesłanka (1)), mamy również skłonność uznawać zdanie ogólne: $\forall\varphi \in L_{PA}\ T(\neg\varphi) \equiv \neg T(\varphi)$. Czy istnieje jednak powód, dla którego *powinniśmy* przyjąć to uogólnienie? Zauważmy dla kontrastu, że wyjaśnienia, w których wyprowadza się rozważane uogólnienie z aksjomatów uznawanej przez nas teorii, wolne są od tego problemu: można by twierdzić, że uznanie wyjściowej teorii T pociąga za sobą zobowiązanie do akceptacji twierdzeń (nie tylko aksjomatów) T. *Wyjaśnienie 4* nie przedstawia jednak wspomnianej generalizacji jako twierdzenia uznawanej przez nas teorii T. Czy zatem powinniśmy uznać rozważane zdanie ogólne, a jeżeli tak, to dlaczego?

Problem 2. Treść przesłanki (1) budzi wątpliwości. Wydaje się, że jakieś założenie w rodzaju „TB (albo MT) jest uznawaną przez nas teorią" jest nieodzowne w Horwichowskich wyjaśnieniach. Co to jednak znaczy „uznawać teorię"?

Rzecz w tym, że przesłanka (1) nie oddaje tej myśli. Niech φ_1 i φ_2 będą zdaniami języka L_T, o których w danym momencie *nie wiem*, czy wynikają z TB. Załóżmy przy tym, że w rzeczywistości $\mathrm{TB} \vdash \varphi_1$, ale $\mathrm{TB} \nvdash \varphi_2$. Jeśli jednak nie jestem świadom tych faktów, to nie wpływają one w żaden sposób na moje dyspozycje — poprzednik przesłanki (1) wydaje się więc pustym warunkiem. Z tego względu przesłanka (1) powinna zostać zmodyfikowana, być może do postaci:

(1') Dla każdego zdania $\varphi \in L_T$, jeśli jesteśmy świadomi, że $\mathrm{TB} \vdash \varphi$, to mamy dyspozycję do uznania φ[12].

[12] Jeśli zamierzonym sensem przesłanki (1) jest „uznajemy daną teorię", to nawet taka modyfikacja nie jest zadowalająca. Może się przecież zdarzyć, że uznaję kilka znanych mi konsekwencji teorii T (innych po prostu nie znam), lecz mimo to nie mam do teorii T zaufania: tak się po prostu składa, że akurat kilka znanych mi konsekwencji uważam za prawdziwe. Podkreślmy, że wzmocnienie (1') przez użycie trybu przypuszczającego nie będzie skuteczne. Rozważmy: (1") Dla każdego zdania $\varphi \in L_T$, gdybyśmy byli świadomi, że $\mathrm{TB} \vdash \varphi$, to mielibyśmy dyspozycję do uznania φ. Oto problem z (1"): uznaję arytmetykę Peana, lecz gdybym zdawał sobie sprawę, że $\mathrm{PA} \vdash 0 = 1$, to nie miałbym dyspozycji do uznania „0 = 1". Wręcz przeciwnie, w tej sytuacji odrzuciłbym PA.

Po uwzględnieniu tej zmiany należałoby zmodyfikować także pozostałe fragmenty podanych wyjaśnień. W obliczu *Problemu 1* zamierzam jednak zaproponować drastyczniejsze rozwiązanie: należy zrezygnować z psychologizmu, zastąpić *Wyjaśnienia (1)-(4)* wariantami nieodwołującymi się do żadnych pojęć psychologicznych.

Pomysł polega na pozbyciu się zarówno „dyspozycji", jak i „bycia świadomym, że…" na rzecz jednego epistemicznego predykatu *wiarygodności*. Predykat ten byłby orzekany o zdaniach, a intuicyjna interpretacja wypowiedzi „φ jest wiarygodne" (w skrócie „$W(\varphi)$") brzmiałaby: „Istnieje silna racja przemawiająca za uznaniem φ". Predykat ten byłby scharakteryzowany aksjomatycznie. Do teorii *Th*, w której budowalibyśmy nasze wyjaśnienia, należałyby w szczególności następujące aksjomaty[13]:

(Ax1) $\forall \varphi \in L_{T,W}$ [jeśli TBW $\vdash \varphi$, to $W(\varphi)$]

(Ax2) $\forall \varphi,\psi \in L_{T,W}$ [jeśli $W(\varphi)$ i $W(\varphi \rightarrow \psi)$, to $W(\psi)$]

Przez „$L_{T,W}$" rozumiem tu rozszerzenie języka L_T o nowy jednoargumentowy predykat W; z kolei TBW to teoria w języku $L_{T,W}$ będąca rozszerzeniem TB o aksjomaty logiczne dla formuł z predykatem W. (Ax1) wyraża nasze zaufanie do TB oraz do logiki w rozszerzonym języku. Stoi za tym intuicja, zgodnie z którą jeśli zdanie φ ma dowód zbudowany przy użyciu środków TB oraz logiki rozszerzonego języka, to φ jest wiarygodne: istnieje silna racja przemawiająca za uznaniem φ (mianowicie dowód w TBW). (Ax2) również uważam za niekontrowersyjny, a myśl w nim wyrażona wydaje się oczywista: jeśli istnieją silne racje przemawiające za uznaniem implikacji oraz jej poprzednika, to istnieje silna racja przemawiająca za uznaniem następnika.

W systemie *Th* poza standardowymi regułami logicznymi mielibyśmy dwie reguły specjalne: NEC oraz regułę Horwicha.

$$\text{NEC} \quad \frac{\vdash \varphi}{\vdash W(\varphi)} \qquad \text{RH} \quad \frac{\vdash \forall x W(\varphi(x))}{\vdash W(\forall x\, \varphi(x))}$$

Intuicja stojąca za regułą NEC jest następująca: jeśli udowodniliśmy φ w naszej teorii, to istnieje silna racja przemawiająca za uznaniem φ (mianowicie nasz dowód φ), czyli φ jest wiarygodna.

Druga reguła to RH, czyli reguła Horwicha. Zgodnie z RH, jeśli udowodniliśmy, że każde numeryczne podstawienie φ jest wiarygodne, to wolno nam uznać za wiarygodne odpowiednie zdanie ogólne. Dla ilustracji załóżmy, że udowodniliśmy, iż każde numeryczne podstawienie formuły „x nie jest dowodem sprzeczności w PA" jest wiarygodne. Zatem (na mocy NEC) dysponujemy silną racją, by uznać, że każde numeryczne podstawienie „x nie jest dowodem sprzeczności w PA" jest wiarygodne. Inaczej mówiąc, mamy: $W(\forall x\ W(x$ nie jest dowodem sprzeczności w PA)). Zasadniczy jest przy tym fakt, że ogólny kwantyfikator pozostaje w zasięgu zewnętrznego

[13] Przyjmuję dodatkowo, że *Th* jest rozszerzeniem arytmetyki Peana.

predykatu wiarygodności. Zgodnie z intuicją kryjącą się za RH w opisanej sytuacji mamy silną rację, by uznać, że dla każdego *x*, *x* nie jest dowodem sprzeczności w PA — czyli że niesprzeczność PA jest wiarygodna[14]. Tą silną racją jest zaś właśnie nasz argument: pokazaliśmy przecież, że jest wiarygodne, iż *każdy dowód* jest wiarygodną (cząstkową) ilustracją faktu, że w PA nie ma sprzeczności.

Tego rodzaju teoria *Th* pozwoliłaby nam dowodzić wiarygodności uogólnień niezależnych od wyjściowej teorii prawdy TB, mimo że same te generalizacje nie byłyby jej twierdzeniami. W ten sposób dostarczałaby odpowiedzi na pytanie, dlaczego ktoś, kto uznaje TB, powinien również uznać szereg zdań ogólnych od TB niezależnych. Odpowiedź brzmi: ktoś, kto uznaje TB, uznaje też wiarygodność TB (taka właśnie intuicja kryje się za aksjomatem (Ax1)), a to przy naturalnych dodatkowych założeniach pozwala wyprowadzić wniosek o wiarygodności dodatkowych zdań ogólnych.

APPENDIX

W artykule nie zajmuję się badaniem formalnych własności teorii *Th*. Własności te będzie można precyzyjnie badać dopiero po dokonaniu ostatecznego wyboru aksjomatyki. Na podkreślenie zasługuje jednak fakt, że już podane dwa aksjomaty pozwalają udowodnić wiarygodność *wszystkich* standardowych kompozycyjnych zasad charakteryzujących pojęcie prawdy. Dowody dla spójników są proste; w wypadku negacji wystarczy np. drobna modyfikacja *Wyjaśnienia 4*. Dalej przedstawię nieco bardziej złożony przypadek zasady kwantyfikatorowej. Udowodnimy mianowicie wiarygodność zdania ogólnego „$\forall\varphi\,[\exists x\, T(\varphi(x)) \equiv T(\exists x\, \varphi(x))]$” w *Th*.

W dowodzie wykorzystamy następujące konwencje notacyjne:

E_n — klasa formuł o złożoności syntaktycznej co najwyżej *n*. Dokładny sposób rozumienia pojęcia złożoności syntaktycznej nie ma tu dla nas większego znaczenia[15]. Ważne jest natomiast, by istniała formuła arytmetyczna „$x \in E_n$” klasy Σ_1 reprezentująca E_n w PA. Zakładam, że mamy do dyspozycji taką formułę.

---

[14] Nie to znaczy, że zdanie Con_{PA} (czyli „PA jest niesprzeczna”) jest twierdzeniem *Th*. W istocie łatwo zauważyć, że tak nie będzie, a opisany system z dwoma aksjomatami i podanymi regułami wnioskowania jest konserwatywnym rozszerzeniem PA. Argument jest banalny: twierdzenia *Th* stają się prawdziwe po zinterpretowaniu predykatu *W* jako zbioru wszystkich formuł języka L_T. Dla przykładu, przy tej interpretacji (Ax1) znaczy: „Każde twierdzenie TB jest formułą języka L_T”. Oznacza to, że *Th* jest interpretowalna w PA. Twierdzeniem *Th* nie będzie zatem Con_{PA}, lecz $W(Con_{PA})$.

[15] Na przykład, E_n można zdefiniować jako klasę formuł, których drzewo syntaktyczne ma wysokość co najwyżej *n*. Jednakże te same rezultaty (Σ_1-reprezentowalność, istnienie arytmetycznego predykatu prawdziwości) potrafimy uzyskać także dla innych miar złożoności syntaktycznej.

$Tr_n(x)$ — arytmetyczny predykat prawdy dla formuł klasy $\mathrm{E_n}$. Przyjmujemy, że predykat ten został tak dobrany, by zachodziło: $\mathrm{PA} \vdash \forall x \forall \varphi \in \mathrm{E_n} [Tr_n(\varphi(x)) \equiv \varphi(x)]$.

Po drodze skorzystamy ze sformalizowanego twierdzenia o Σ_1-zupełności w następującej wersji (zob. Rautenberg 2006: 186):

> **(Sformalizowana Σ_1-zupełność)**. Niech $A(x_1 \ldots x_n)$ będzie arytmetyczną formułą klasy Σ_1. Wówczas:
>
> $$\mathrm{PA} \vdash \forall x_1 \ldots x_n [A(x_1 \ldots x_n) \rightarrow Pr_{\mathrm{PA}}(A(x_1 \ldots x_n))].$$

Przechodzimy teraz do przedstawienia dowodu kompozycyjnego warunku dla kwantyfikatora egzystencjalnego w teorii *Th*.

> **Twierdzenie**. $Th \vdash W(\forall \varphi [\exists x\, T(\varphi(x)) \equiv T(\exists x\, \varphi(x))])$
>
> *Dowód:*

(1) $\forall \varphi \forall n \forall x\ \mathrm{TB} \vdash \varphi \in \mathrm{E_n} \rightarrow T(\varphi(x)) \equiv Tr_n(\varphi(x))$

(2) $\forall \varphi \forall n \forall x\ W(\varphi \in \mathrm{E_n} \rightarrow T(\varphi(x)) \equiv Tr_n(\varphi(x)))$

(3) $W(\forall \varphi \forall n \forall x [\varphi \in \mathrm{E_n} \rightarrow T(\varphi(x)) \equiv Tr_n(\varphi(x))])$

(4) $\forall \varphi \forall n\ W(\forall x [\varphi \in \mathrm{E_n} \rightarrow T(\varphi(x)) \equiv Tr_n(\varphi(x))])$

(5) $\forall \varphi \forall n [\varphi \in \mathrm{E_n} \rightarrow W(\forall x [T(\varphi(x)) \equiv Tr_n(\varphi(x))])]$

(6) $\forall \varphi \forall n (\varphi \in \mathrm{E_n} \rightarrow \mathrm{TB} \vdash \forall x [Tr_n(\varphi(x)) \equiv \varphi(x)])$

(7) $\forall \varphi \forall n (\varphi \in \mathrm{E_n} \rightarrow W(\forall x [Tr_n(\varphi(x)) \equiv \varphi(x)]))$

(8) $\forall \varphi \forall n (\varphi \in \mathrm{E_n} \rightarrow W(\forall x [T(\varphi(x)) \equiv \varphi(x)]))$

(9) $\forall \varphi\ W(\forall x [T(\varphi(x)) \equiv \varphi(x)])$

(10) $\forall \varphi\ W(\exists x\, T(\varphi(x)) \equiv \exists x\, \varphi(x))$

(11) $\forall \varphi\ W(T(\exists x\, \varphi(x)) \equiv \exists x\, \varphi(x))$

(12) $\forall \varphi\ W(\exists x\, T(\varphi(x)) \equiv T(\exists x\, \varphi(x)))$

(13) $W(\forall \varphi [\exists x\, T(\varphi(x)) \equiv T(\exists x\, \varphi(x))])$

A oto komentarze do poszczególnych kroków: (1) jest dowodliwe już w samej PA, a zatem tym bardziej w *Th*; (2) wynika z (1) na mocy (Ax1); (3) uzyskujemy przez zastosowanie RH. W celu uzyskania (4) wykorzystujemy zdanie ogólne „$\forall \varphi(x)\ [W(\forall x\ \varphi(x)) \rightarrow \forall x\ W(\varphi(x))]$”, dowodliwe w *Th*. W (5) stosujemy sformalizowane twierdzenie o Σ_1-zupełności: jeśli $\varphi \in \mathrm{E_n}$, to PA dowodzi tego faktu (wyrażająca go formuła jest bowiem klasy Σ_1), a zatem na mocy (Ax1) $W(\varphi \in \mathrm{E_n})$ i nasza konkluzja wynika z (4); (6) jest dowodliwe w PA; (7) wynika z (6) na mocy

(Ax1); (8) uzyskujemy z (7) i (5); (9) w sposób oczywisty wynika z (8); (10) otrzymujemy z (9) oraz z aksjomatów *Th* — w szczególności korzystamy tu z faktu, że wszystkie tautologie logiki pierwszego rzędu języka $L_{T,W}$ są wiarygodne[16]; (11) zachodzi na mocy (Ax1), mamy bowiem: $\forall\varphi\ TB \vdash T(\exists x\ \varphi(x)) \equiv \exists x\ \varphi(x)$; (12) wynika bezpośrednio z (10) i (11) oraz z aksjomatów *Th*; wreszcie (13) to wynik zastosowania RH do (12).

Na zakończenie chciałbym podkreślić, że wyboru przedstawionej tu aksjomatyki nie uważam za ostateczny. Na rozważenie zasługuje np. następujący aksjomat niesprzeczności:

(Ax3) $\forall\varphi \in L_{T,W}\ \neg W(\varphi \wedge \neg\varphi)$.

Teoria *Th* bez (Ax3) jest konserwatywnym rozszerzeniem PA, ale dodanie go znacząco zmienia postać rzeczy — np. niesprzeczność PA staje się wówczas twierdzeniem *Th*. Na obecnym etapie intuicje kryjące się za (Ax3) nie są dla mnie jednak do końca jasne. Z jednej strony, można by twierdzić, że sprzeczność *nigdy* nie jest wiarygodna, a jeśli jakiś argument prowadzi do sprzeczności, świadczy to tylko o tym, że nie jest on silną racją do przyjęcia czegokolwiek. Istnieje jednak intuicja odwrotna. Czasami stajemy przecież w obliczu paradoksów — zdarza się, że pewne silne argumenty przemawiają za φ, a inne równie silne argumenty przemawiają za jego negacją. W naszej terminologii oba te zdania byłyby wówczas wiarygodne. To oczywiście nie znaczy, że zdanie i jego negacja są jednocześnie prawdziwe. Wiarygodność i prawdziwość to pojęcia, których nie powinniśmy utożsamiać.

BIBLIOGRAFIA

Armour-Garb B. (2010), *Horwichian Minimalism and the Generalization Problem*, „Analysis" 70(4), 693-703.

Cieśliński C. (2007), *Deflationism, Conservativeness and Maximality,* „Journal of Philosophical Logic" 36(6), 695-705.

Cieśliński C. (2009), *Deflacyjna koncepcja prawdy*, Warszawa: Wydawnictwo Naukowe Semper.

Cieśliński C. (2010), *Truth, Conservativeness, and Provability*, „Mind" 119(474), 409-422.

Cieśliński C. (2015), *The Innocence of Truth*, „Dialectica" 69(1), 61-85.

Gupta A. (1993), *Minimalism*, „Philosophical Perspectives" 7, 359-369.

Horwich P. (1998), *Truth*, Oxford: Blackwell.

Horwich P. (2001), *A Defense of Minimalism*, „Synthese" 126(1), 149-165.

Horwich P. (2010), *Truth — Meaning — Reality*, Oxford: Clarendon Press.

Ketland J. (2010), *Truth, Conservativeness, and Provability. Reply to Cieśliński*, „Mind" 119(474), 423-436.

McGee V. (1992), *Maximal Consistent Sets of Instances of Tarski's Schema (T)*, „Journal of Philosophical Logic" 21(3), 235-241.

Raatikainen P. (2005), *On Horwich's Way Out*, „Analysis" 65(287), 175-77.

[16] Na mocy (Ax1) mamy: $\forall\varphi\ W(\forall x\ [T(\varphi(x)) \equiv \varphi(x)] \rightarrow (\exists x\ T(\varphi(x)) \equiv \exists x\varphi(x)))$. Skoro (9), to uzyskujemy (10) na mocy (Ax2).

Rautenberg W. (2006), *A Concise Introduction to Mathematical Logic*, New York, NY: Springer.
Tarski A. (1933), *Pojęcie prawdy w językach nauk dedukcyjnych*, Warszawa: Towarzystwo Naukowe Warszawskie.
Tennant N. (2002), *Deflationism and the Gödel Phenomena*, „Mind" 111(443), 551-582.
Tennant N. (2010), *Deflationism and the Gödel Phenomena. Reply to Cieśliński*, „Mind" 119(474), 437-450.