WYKŁAD MONOGRAFICZNY: ZAGADKI

ZAGADKI LOGICZNE 2:
SZCZĘŚCIARZE EPISTEMICZNI

JERZY POGONOWSKI

Zakład Logiki Stosowanej UAM
www.logic.amu.edu.pl
pogon@amu.edu.pl

1 Wstęp

Rozważymy teraz nieco bardziej złożone zagadki logiczne, uwzględniające nie tylko mówienie fałszu lub prawdy, ale także posiadanie przekonań. Tak jak poprzednio, naszymi bohaterami będą osoby mówiące zawsze bądź prawdę bądź fałsz. Teraz jednak dopuścimy, że osoby te mają ściśle określone przekonania i wypowiadając się, mogą być szczere (wierzyć w to co mówią) bądź nieszczere (nie wierzyć w to co mówią).

1.1 Język opisu przekonań

Poniższe zagadki pochodzą naszego tłumaczenia książki Raymonda Smullyana *Forever Undecided. A Puzzle Guide to Gödel*, które ukazało się w 2007 roku nakładem *Książki i Wiedzy*, pod tytułem *Na Zawsze Nierozstrzygnięte. Zagadkowy Przewodnik Po Twierdzeniach Gödla.* Obok zagadek o Rycerzach (mówiących zawsze prawdę) oraz Łotrach (mówiących zawsze fałsz), książka zawiera zagadki logiczne, w których w formie popularnej przedstawia się *logikę epistemiczną* oraz *logikę dowodliwości*.

Ustalimy najpierw środki językowe, za pomocą których będziemy mówili o żywieniu przekonań. Do języka klasycznego rachunku zdań dodajemy jeden funktor zdaniotwórczy o jednym argumencie zdaniowym: $\boldsymbol{B}$. Jeśli α jest formułą naszego języka, to jest nią także $\boldsymbol{B}\alpha$. Wyrażenie to czytamy: „(rozważany podmiot) wierzy, że α". Równoważnie, czytać je możemy: „α należy do zbioru przekonań (rozważanego podmiotu)".

Rozważać można też funktor $\boldsymbol{K}$, przy następującej interpretacji: zdanie $\boldsymbol{K}p$ czytamy *(rozważany podmiot) wie, że* p (gdzie p jest dowolnym zdaniem języka

logiki epistemicznej). Zwykle zakłada się, że $\boldsymbol{K}p \equiv (p \wedge \boldsymbol{B}p)$. Poniżej będziemy zajmowali się głównie funktorem $\boldsymbol{B}$.

Należy zwracać szczególną uwagę na odróżnienie języka przedmiotowego dotyczącego przekonań oraz jego metajęzyka. Niech S będzie zespołem wszystkich przekonań rozważanego podmiotu. Gdy jakaś formuła α języka przedmiotowego należy do S, to mówimy, że α należy do przekonań tego podmiotu. Uzus językowy każe w takim przypadku mówić, że podmiot ów *wierzy* w α. Jest to jednak użycie *metajęzykowe* tego słowa.

Systemy epistemiczne są interesujące same przez się – w opisie systemów przekonań, w szczególności: racjonalnych świadomych przekonań. Mają one także interesującą i ważną interpretację metalogiczną: $\boldsymbol{B}p$ można interpretować jako *zdanie p jest dowodliwe w arytmetyce PA*.

Uwaga. Angielski termin *reasoner* stosowany przez Smullyana oddajemy przez polski neologizm *myślak*.

Twój zespół przekonań to twoja osobista sprawa, możesz wierzyć w co ci się żywnie podoba. Ta swoboda przekonań wiąże się oczywiście z ryzykiem – jeśli twoje przekonania są nietrafne, to pewne twoje działania na nich oparte mogą szybko doprowadzić do usunięcia cię z areny zdarzeń. Znikniesz w niebycie o wiele prędzej od tych, których przekonania nie są tak lunatyczne jak twoje.

Nie jesteśmy natomiast całkowicie wolni w momencie, gdy zaczynamy coś mniemać na temat samych żywionych przez nas przekonań. Tu liczyć się trzeba z ingerencją logiki. Przy stosownie dobranych warunkach charakteryzujących twoje przekonania, pewnych przekonań o własnych przekonaniach mieć nie możesz, a inne znów mieć musisz. Zależy to, jak pokażemy niżej, od tego, na ile jesteś *samoświadoma*.

Ktoś, kto zna na pamięć treść wielkiej encyklopedii jest bez wątpliwości mądralą. Jeśli wierzy we wszystko, co podaje encyklopedia, to ma mnóstwo przekonań, a ich trafność zależy od tego, na ile trafne są owe treści encyklopedyczne. Może wygrać teleturniej, błyszczeć w towarzystwie, a może nawet zdać egzamin maturalny. Tutaj interesować nas będzie jednak inny typ mądrali – nie encyklopedyczny, lecz *analityczny*, taki, który zastanawia się nad trafnością swoich mniemań o własnych przekonaniach.

1.2 Szczęściarze epistemiczni

Przypuśćmy, że jesteś racjonalną, samoświadomą Istotą. Jak to przypuszczenie przełożyć na język logiki epistemicznej? Oto propozycja. Nazwiemy *szczęściarzem epistemicznym* każdą osobę S, której system przekonań spełnia następujące warunki:

- (1a) S wierzy we wszystkie tautologie klasycznego rachunku zdań;
- (1b) system przekonań S jest domknięty na regułę *modus ponens*: jeśli S wierzy w p oraz wierzy w $p \to q$, to wierzy także w q;
- (2) dla dowolnych p oraz q, S wierzy w $(\boldsymbol{B}p \wedge \boldsymbol{B}(p \to q)) \to \boldsymbol{B}q$;
- (3) dla dowolnego p, jeśli S wierzy w p, to wierzy w $\boldsymbol{B}p$;
- (4) dla dowolnego p, S wierzy w $\boldsymbol{B}p \to \boldsymbol{B}\boldsymbol{B}p$.

Uwaga: rozważamy tylko osoby, które albo zawsze mówią prawdę, albo zawsze mówią fałsz.

Każdą osobę, która spełnia jedynie warunki (1a) i (1b) nazwiemy (bez urazy) *prostaczkiem logicznym*. Zatem, jeśli S jest prostaczkiem logicznym, to jego/jej system przekonań zawiera klasyczną logikę zdaniową, ale S może być tego nieświadom(a).

Tak więc, dana osoba jest *prostaczkiem logicznym*, gdy jej zespół przekonań zawiera wszystkie tautologie klasycznego rachunku zdań i jest domknięty na regułę *modus (ponendo) ponens*, czyli regułę odrywania:

$$\frac{\alpha \to \beta, \quad \alpha}{\beta}.$$

Jest, jak wiadomo, nieskończenie wiele tautologii klasycznego rachunku zdań. Założenie, że jakaś osoba wierzy w nie wszystkie to oczywiście swoista idealizacja. Każdy z nas ma jedynie skończenie wiele konkretnych przekonań – nasz żywot trwa najwyżej kilkadziesiąt lat i w tym czasie nie sposób zebrać nieskończonej liczby przekonań. Mamy jednak teoretyczną możliwość ustalenia – dla dowolnej formuły języka klasycznego rachunku zdań – czy formuła ta jest, czy też nie jest tautologią tego rachunku. Tu kryje się następne założenie idealizacyjne: ponieważ formuły języka klasycznego rachunku zdań mogą być dowolnie długie, a nasz żywot krótki, więc *praktycznie* zastosować możemy algorytm rozstrzygający problem tautologiczności jedynie do formuł pewnej długości. *Wiemy* jednak, że algorytm daje odpowiedź dla dowolnie długiej formuły i to nam wystarcza.

Na marginesie przypomnijmy argument Smullyana za tym, że każdy z nas jest albo zarozumiały, albo sprzeczny w swoich przekonaniach. Mamy oto skończony zbiór przekonań. Jeśli sądzimy, że wszystkie z nich są trafne, to jesteśmy zarozumiali. Jeśli natomiast skromnie dopuszczamy, że co do pewnych naszych przekonań możemy się mylić, to automatycznie uznajemy, że cały nasz system przekonań jest wewnętrznie sprzeczny.

Bycie prostaczkiem logicznym to nie hańba – w istocie, każdy prostaczek logiczny jest całkowicie poprawny w swoich rozumowaniach, a dokładniej: nie popełnia żadnych błędów formalnych. Może popełniać błędy materialne, o ile do zespołu jego najbardziej elementarnych przekonań (reprezentowanych przez zmienne zdaniowe) należą zdania fałszywe.

1.3 Poziomy samoświadomości

Powiemy, że osoba S jest:

- *normalna*, gdy jeśli wierzy w p, to wierzy też w $\boldsymbol{B}p$;
- *regularna*, gdy jeśli wierzy w $p \to q$, to wierzy też w $\boldsymbol{B}p \to \boldsymbol{B}q$;
- *sprzeczna*, gdy do jej systemu przekonań należy jakaś para zdań wzajem sprzecznych, lub – co na jedno wychodzi – *fałsz logiczny*, który oznaczamy przez $\bot$.

Uwaga. Może bardziej właściwe byłoby mówienie o własnościach *systemów przekonań*, a nie *osób*.

Można udowodnić, że: $(*)$ dowolny szczęściarz epistemiczny S *wie*, że jeśli uwierzy w jakieś zdanie p oraz w jego negację $\neg p$, to stanie się sprzeczny.

O szczęściarzach epistemicznych można udowodnić wiele innych ciekawych rzeczy. Nie wszystkie z nich będą nam dalej potrzebne. Dodajmy może jedynie, że:

- każdy szczęściarz epistemiczny jest normalny, a nawet *wie*, że jest normalny;
- każdy szczęściarz epistemiczny jest regularny i o tym także *wie*;
- wreszcie, każdy szczęściarz epistemiczny jest przekonany o tym, że jest szczęściarzem epistemicznym; a zatem to jego przekonanie jest *trafne* i, w konsekwencji, każdy szczęściarz epistemiczny *wie*, że jest szczęściarzem epistemicznym.

Można rozważać pięć typów myślaków, o wstępujących poziomach samoświadomości:

- Typ 1: prostaczek logiczny.
- Typ 1^*: prostaczek logiczny, który, jeśli uwierzył w $p \to q$, to uwierzy, że jeśli uwierzył w p, to uwierzy w q.

- Typ 2: prostaczek logiczny, który wierzy we wszystkie zdania postaci $(\boldsymbol{B}p \wedge \boldsymbol{B}(p \to q)) \to \boldsymbol{B}q$.
- Typ 3: myślak typu 2, który, jeśli wierzy w p, to wierzy w $\boldsymbol{B}p$.
- Typ 4: szczęściarz epistemiczny, tj. normalny i regularny prostaczek logiczny, który wierzy we wszystkie zdania postaci $\boldsymbol{B}p \to \boldsymbol{B}\boldsymbol{B}p$, czyli wierzy, że jest normalny.

Uwaga. Terminy: *prostaczek logiczny* oraz *szczęściarz epistemiczny* nie występują w *Forever Undecided*; wprowadzamy je na użytek tej prezentacji.

Z podanych definicji wynika, że:

- Każdy prostaczek logiczny jest myślakiem typu 1^*.
- Każdy myślak typu 1^* jest regularnym prostaczkiem logicznym (i *vice versa*).
- Każdy myślak typu 2 wie, że jest typu 1^*.
- Myślaki typu 3 to dokładnie normalne myślaki typu 2.
- Dla $1 \leqslant n < 4$: każdy myślak typu n jest też myślakiem typu $n + 1$.
- $1 < n \leqslant 4$: każdy myślak typu n wierzy, że jest myślakiem typu $n - 1$.

Uwaga. Ponieważ każdy szczęściarz epistemiczny *wie*, że jest szczęściarzem epistemicznym, więc stanowi on zwieńczenie hierarchii samoświadomych myślaków. Inaczej mówiąc, gdybyśmy chcieli zdefiniować myślaka typu 5 jako takiego, który jest typu 4 i wierzy, iż jest typu 4, to otrzymalibyśmy jedynie myślaka typu 4.

2 II Twierdzenie Gödla

Za chwilę dowiesz się czegoś naprawdę frapującego o swoim systemie przekonań. Udowodnimy mianowicie:

Twierdzenie 1.

Przypuśćmy, że normalny prostaczek logiczny S wierzy w zdanie postaci $p \equiv \neg\boldsymbol{B}p$. Wtedy:

- (a) Jeśli S kiedykolwiek uwierzy w p, to stanie się sprzeczny.
- (b) Jeśli S jest szczęściarzem epistemicznym, to wie, iż jeśli kiedykolwiek uwierzy w p, to stanie się sprzeczny – tj. uwierzy w $\boldsymbol{B}p \to \boldsymbol{B}\ \bot$.

- (c) Jeśli S jest szczęściarzem epistemicznym i wierzy, że nie może być sprzeczny, to stanie się sprzeczny.

Dowód Twierdzenia 1.

(a) Przypuśćmy, że S wierzy w p. Będąc normalnym, uwierzy w $\boldsymbol{B}p$. Nadto, ponieważ wierzy w p oraz wierzy w $p \equiv \neg\boldsymbol{B}p$, więc musi uwierzyć w $\neg\boldsymbol{B}p$ (bo jest prostaczkiem logicznym). A więc uwierzy jednocześnie w $\boldsymbol{B}p$ oraz w $\neg\boldsymbol{B}p$, a stąd stanie się sprzeczny.

(b) Przypuśćmy, że S jest szczęściarzem epistemicznym. Ponieważ jest wtedy prostaczkiem logicznym i wierzy w $p \equiv \neg\boldsymbol{B}p$, więc musi także wierzyć w $p \rightarrow \neg\boldsymbol{B}p$. Nadto, S jest regularny, a stąd uwierzy w $\boldsymbol{B}p \rightarrow \boldsymbol{B}\neg\boldsymbol{B}p$. Wierzy też w $\boldsymbol{B}p \rightarrow \boldsymbol{B}\boldsymbol{B}p$ (ponieważ wie, że jest normalny). Zatem S uwierzy w $\boldsymbol{B}p \rightarrow (\boldsymbol{B}\boldsymbol{B}p \wedge \boldsymbol{B}\neg\boldsymbol{B}p)$, które jest logiczną konsekwencją ostatnich dwóch zdań. Wierzy również w $(\boldsymbol{B}\boldsymbol{B}p \wedge \boldsymbol{B}\neg\boldsymbol{B}p) \rightarrow \boldsymbol{B}\ \bot$ (na mocy $(*)$, ponieważ dla dowolnego zdania X, S wierzy w $(\boldsymbol{B}X \wedge \boldsymbol{B}\neg X) \rightarrow \boldsymbol{B}\ \bot$, a więc wierzy w jego szczególny przypadek, gdzie X jest zdaniem $\boldsymbol{B}p$). Gdy S już uwierzy jednocześnie w $\boldsymbol{B}p \rightarrow (\boldsymbol{B}\boldsymbol{B}p \wedge \boldsymbol{B}\neg\boldsymbol{B}p)$ oraz w $(\boldsymbol{B}\boldsymbol{B}p \wedge \boldsymbol{B}\neg\boldsymbol{B}p) \rightarrow \boldsymbol{B}\ \bot$, będzie musiał uwierzyć w $\boldsymbol{B}p \rightarrow \boldsymbol{B}\ \bot$ (ponieważ jest prostaczkiem logicznym).

(c) Ponieważ S wierzy w $\boldsymbol{B}p \rightarrow \boldsymbol{B}\ \bot$ (jak właśnie udowodniliśmy), więc wierzy także w $\neg\boldsymbol{B}\ \bot\rightarrow \neg\boldsymbol{B}p$. Załóżmy teraz, że S wierzy w $\neg\boldsymbol{B}\ \bot$ (wierzy, że nie może być sprzeczny). Ponieważ wierzy też w $\neg\boldsymbol{B}\ \bot\rightarrow \neg\boldsymbol{B}p$ (jak właśnie widzieliśmy), więc uwierzy w $\neg\boldsymbol{B}p$. A ponieważ wierzy również w $p \equiv \neg\boldsymbol{B}p$, więc uwierzy w p, a stąd stanie się sprzeczny, na mocy (a).

Udowodniliśmy przed chwilą nie byle co, bo modalną (epistemiczną) wersję *II Twierdzenia Gödla* (o niedowodliwości niesprzeczności arytmetyki w samej arytmetyce). Oczywiście był to dowód w postaci wielce uproszczonej – precyzyjny dowód wymagałby, powiedzmy, jednosemestralnego wykładu wstępnego.

W prezentacji korzystaliśmy z rozdziału 12 tłumaczenia książki Raymonda Smullyana *Forever Undecided*. Poddajemy ocenie audytorium, czy ten sposób popularyzacji wiedzy (meta)logicznej można uznać za dydaktycznie przydatny.

Przykład teologiczny. Przypuśćmy, że jesteś studentką teologii i że Twój Ulubiony Profesor teologii mówi do Ciebie:

Bóg istnieje wtedy i tylko wtedy, gdy nigdy nie uwierzysz, że Bóg istnieje.

Jeśli wierzysz profesorowi, to wierzysz w zdanie $g \equiv \neg\boldsymbol{B}g$, gdzie g jest zdaniem stwierdzającym, że Bóg istnieje. Wtedy, zgodnie z Twierdzeniem 1, nie możesz wierzyć w swoją własną niesprzeczność bez popadnięcia w sprzeczność. Oczywiście, możesz wierzyć we własną niesprzeczność, bez popadnięcia przy tym w sprzeczność – wystarczy, że przestaniesz ufać Twojemu Ulubionemu Profesorowi. Coś za coś. Przy modalnej interpretacji *dowodliwości* nie mamy jednak takiej

możliwości ucieczki, jak w powyższym przykładzie. Wiadomo, że formuła $god(\overline{n})$, stwierdzająca swoją własną niedowodliwość w PA (gdzie $\overline{n}$ jest stosownym numerem gödlowskim), jest prawdziwa, lecz dowodu w PA nie posiada. Można pokazać, że twierdzeniem stosownego systemu modalnego (w którym reprezentujemy dowodliwość w PA) jest:

$$god(\overline{n}) \equiv \neg \boldsymbol{B} god(\overline{n}).$$

3 Zastosowanie Twierdzenia Löba

Ludziska tracą mnóstwo czasu i energii na przekonywanie samych siebie, że przydarzyć się może coś, czego bardzo pragną, i to jedynie dlatego, że właśnie tego pragną. Typowe tego typu sytuacje to klepanie modlitw lub zaklęć. Podobno przeprowadzano badania empiryczne, mające dostarczyć przekonujących dowodów na rzecz skuteczności bądź nieskuteczności modlitw. Każdy zresztą może takie badania przeprowadzić – wystarczy przekonać dwie osoby, aby modliły się o coś wzajem przeciwnego i czekać na wyniki. Może trudno w to uwierzyć, ale w niektórych krajach zdarza się, że ślubowania poselskie wzmacniane są formułką przywołującą moce nadprzyrodzone, w tekst konstytucji wpisane są deklaracje wiary, a czasem nawet określona religia nauczana bywa w szkołach.

Zdarzają się jednak sytuacje, w których wiara w zajście jakiegoś zdarzenia – przy założeniu pewnych dodatkowych warunków – istotnie implikuje, że zdarzenie to zajść musi.

Pokażemy teraz, co wystarcza, aby każda z obecnych tu Uroczych Pań została – powiedzmy – *Miss World 2013*. Będzie to przykład *samospełniającego się przekonania*. Przypuśćmy, że:

- jesteś szczęściarą epistemiczną;
- osoby, które rozważamy albo zawsze mówią fałsz, albo zawsze mówią prawdę (i Ty wiesz, że tak jest);
- wierzysz swojemu chłopakowi, który prawdziwie (!) mówi:

  (†) *Jeśli uwierzysz, że zostaniesz Miss World 2013, to zostaniesz Miss World 2013.*

- wierzysz też mnie (JP), który mówi:

  (‡) *Jeśli wierzysz, że ja zawsze mówię prawdę, to zostaniesz Miss World 2013.*

Twierdzenie 2. Przy powyższych założeniach *zostaniesz* Miss World 2013. Cieszysz się?

Dla skrótu, przyjmijmy oznaczenia:

- k zastępuje zdanie stwierdzające, iż ja (JP) zawsze mówię prawdę;
- α zastępuje zdanie stwierdzające, że zostaniesz Miss World 2013.

Dowód składa się z dwóch części.

1. W pierwszej pokazujemy, że nasze założenia implikują $\boldsymbol{B}\alpha$. Jest to dowód założeniowy, dostępny dla każdej szczęściary epistemicznej.

Mamy udowodnić formułę:

$$(\bigstar) \qquad ((\boldsymbol{B}\alpha \to \alpha) \wedge (k \equiv (\boldsymbol{B}k \to \alpha))) \to \boldsymbol{B}\alpha.$$

Uwaga. Zdanie k stwierdza, iż JP zawsze mówi prawdę; a więc prawdą jest, że JP wypowiada $(\ddagger)$ dokładnie wtedy, gdy prawdziwe jest $k \equiv (\ddagger)$, czyli dokładnie wtedy, gdy prawdziwe jest $k \equiv (\boldsymbol{B}k \to \alpha)$.

1.	$(\boldsymbol{B}\alpha \to \alpha) \wedge (k \equiv (\boldsymbol{B}k \to \alpha))$	założenie
2.	$\boldsymbol{B}\alpha \to \alpha$	OK: 1
3.	$k \equiv (\boldsymbol{B}k \to \alpha)$	OK: 1
4.	$k \to (\boldsymbol{B}k \to \alpha)$	OR: 3
5.	$(\boldsymbol{B}k \to \alpha) \to k$	OR: 3
6.1.	k	założenie dodatkowe
6.2.	$\boldsymbol{B}k \to \alpha$	MP: 4, 6.1.
6.3.	$\boldsymbol{B}k$	6.1. i warunek **(3)**
6.4.	α	MP: 6.2., 6.3.
7.	$k \to \alpha$	6.1.→6.4.
8.	$\boldsymbol{B}(k \to \alpha)$	7 i warunek **(3)**
9.	$\boldsymbol{B}k \to \boldsymbol{B}\alpha$	8 i warunki **(1a)** i **(2)**
10.	$\boldsymbol{B}k \to \alpha$	2, 9 i warunki **(1b)**, **(1a)** (prawo sylog. hipotet.)
11.	k	MP: 5, 10
12.	$\boldsymbol{B}k$	11 i warunek **(3)**
13.	α	MP: 10, 12
14.	$\boldsymbol{B}\alpha$	13 i warunek **(3)**.

2. Ponieważ proroctwo $(\dagger)$ Twojego chłopaka (tj. zdanie $\boldsymbol{B}\alpha \to \alpha$) jest z założenia prawdziwe, a powyższy dowód formuły $(\bigstar)$ pokazuje, iż nasze założenia implikują $\boldsymbol{B}\alpha$, więc na mocy reguły odrywania otrzymujemy α, czyli tezę.

Zostaniesz Miss World 2013!!! Cieszysz się???

Uwaga. Powyższy dowód był przykładem *dowodu wprost*. Aby pokazać, że zostaniesz Miss World 2013 nie musieliśmy odwoływać się do *absurdu*. Cieszysz się?
Ciekawostka prowincjonalna. 16 maja 2005 roku odbyły się demokratyczne wybory Dyrektora Instytutu Językoznawstwa UAM. Dwa tygodnie wcześniej, na Seminarium Zakładu Logiki Stosowanej UAM, odczyt *Kto będzie Dyrektorem Instytutu Językoznawstwa UAM?* wygłosiła Pani Dr *Alice Ann Hunter* (*Department of Independent Logic*, *King David University*, Negev Desert). Korzystając z twierdzeń logiki epistemicznej (z Twierdzenia Löba), Dr *Hunter* trafnie przewidziała wynik wyborów. Jak się domyślasz, dowód był podobny do podanego wyżej dowodu, że zostaniesz Miss World 2013. Tekst odczytu dostępny na stronie:

www.logic.amu.edu.pl

* * *

Dalsze zagadki o szczęściarzach epistemicznych, związane m.in. z: I Twierdzeniem Gödla (o niezupełności arytmetyki), twierdzeniem Tarskiego (o niedefiniowalności pojęcia prawdy arytmetycznej w języku arytmetyki) oraz innymi jeszcze twierdzeniami podano w prezentacji *Alicja, labirynty i magiczny ogród*. Zachęcamy też oczywiście do lektury całej książki Raymonda Smullyana *Na Zawsze Nierozstrzygnięte. Zagadkowy Przewodnik Po Twierdzeniach Gödla.*