Logika w zastosowaniach kognitywistycznych

Paradoks wszechwiedzy logicznej (*logical omniscience paradox*) i wybrane metody jego unikania

(notatki do wykładów)

Andrzej Wiśniewski
Andrzej.Wisniewski@amu.edu.pl

wersja beta 1.1

Niech $\otimes$ będzie metazmienną, której wartościami są operatory wiedzy lub operatory przekonań. Niech A, B, ... będą metajęzykowymi zmiennymi przebiegającymi zbiór formuł języka (zdaniowej) logiki epistemicznej. Symbol $\vdash$ czytamy "jest tezą". Rozważmy zależności o schematach:

LO1: *jeśli* $\vdash A$, *to* $\vdash \otimes A$

LO2: *jeśli* $\vdash A \rightarrow B$, *to* $\vdash \otimes A \rightarrow \otimes B$

LO3: *jeśli* $\vdash A \leftrightarrow B$, *to* $\vdash \otimes A \leftrightarrow \otimes B$

LO4: $\vdash \otimes A \rightarrow \otimes(A \vee B)$

LO5: $\vdash (\otimes A \vee \otimes B) \rightarrow \otimes(A \vee B)$

LO6: $\vdash \neg(\otimes A \wedge \otimes \neg A)$

LO7: $\vdash \otimes(A \rightarrow B) \rightarrow (\otimes A \rightarrow \otimes B)$

LO8: $\vdash \otimes(A \leftrightarrow B) \rightarrow (\otimes A \leftrightarrow \otimes B)$

LO9: $\vdash \otimes(A \wedge B) \rightarrow \otimes A \wedge \otimes B$

LO10: $\vdash \otimes A \wedge \otimes B \rightarrow \otimes(A \wedge B)$

Zależności te obowiązują we wszystkich rozważanych dotąd logikach wiedzy i/lub przekonań. Z drugiej strony, co najmniej niektóre z tych zależności można interpretować jako przypisujące podmiotowi epistemicznemu wszechwiedzę logiczną (*logical omniscience*).

Dlaczego?

Rozważmy zależność:

LO1: *jeśli* $\vdash A$, *to* $\vdash \otimes A$

Niech $\otimes$ oznacza "podmiot *a* wie, że".

Zależność **LO1** implikuje wówczas m.in., że podmiot a zna wszystkie tezy *KRZ* (ściślej: że dla każdej tezy *KRZ*, *A*, jest tak, że podmiot *a* wie, że *A*).

A teraz przypuśćmy, że $\otimes$ oznacza "podmiot *a* jest (całkowicie) przekonany, że *A*".

Wówczas, na mocy **LO1**, każda teza *KRZ* jest (całkowitym) przekonaniem podmiotu *a*.

Można zapytać, czy nie są to zbyt daleko idąca idealizacje.

Teraz rozważmy zależność:

LO2: *jeśli* $\vdash A \rightarrow B$, *to* $\vdash \otimes A \rightarrow \otimes B$

Jak pamiętamy, formuła postaci $A \rightarrow B$ (języka *KRZ*) jest tezą *KRZ* wtw formuła *B* wynika logicznie na gruncie *KRZ* z formuły *A*.

Przypuśćmy, że podmiot *a* wie/jest przekonany, że *A*. Wówczas, na mocy zależności **LO2**, do wiedzy/przekonań podmiotu *a* należą wszystkie sądy wynikające logicznie z *A*. W tym także te, które wprawdzie wynikają logicznie z *A*, ale są na tyle "odległymi konsekwencjami" sądu *A*, iż rozważany podmiot nie wie/nie jest przekonany/nie jest świadomy tego, że wynikają one z *A*.

Również tutaj można mieć uzasadnioną wątpliwość, czy nie jest to zbyt daleko idąca idealizacja.

Z kolei zależność:

LO3: *jeśli* $\vdash A \leftrightarrow B$, *to* $\vdash \otimes A \leftrightarrow \otimes B$

zdaje się orzekać o podmiocie *a* co następuje: gdy *A* oraz *B* są logicznie równoważne, to podmiot *a* wie/ jest przekonany, że *A* zawsze i tylko, gdy podmiot *a* wie/ jest przekonany, że *B*.

Jest oczywiste, że jest to daleko idąca idealizacja.

O zależnościach **LO4-LO10** mówiliśmy na konwersatorium, stąd też pominę tutaj ich analizę.

Zauważmy teraz, że obowiązywanie w rozważanych logikach epistemicznych zależności **LO1**, **LO2** i **LO3** jest prostym następstwem faktu, iż logiki te budujemy na bazie normalnych modalnych logik zdaniowych.

Z kolei, patrząc od strony semantycznej, gdy budujemy semantyki dla (zdaniowych) logik epistemicznych tak, jak robiliśmy to dotąd, zależności **LO1**, **LO2** i **LO3** obowiązywać muszą. Aby ich obowiązywanie uchylić, musimy zatem zmodyfikować semantykę.

Do powstających w logikach epistemicznych paradoksów wszechwiedzy logicznej można podejść dwojako.

Po pierwsze, można twierdzić, że w istocie wskazują one na przedmiot rozważań danej logiki epistemicznej, logiki wiedzy i/lub przekonań. Otóż przedmiotem tym nie są tylko odpowiednie nastawienia sądzeniowe (*propositional attitudes*) żywione aktualnie przez podmiot – nawet wyidealizowany – lecz również coś więcej: zobowiązania (*commitments*) podmiotu do żywienia tych nastawień.

Co to konkretnie znaczy? Cóż, zapraszam na wykład :)

Po drugie, można uznać, że występowanie paradoksów wszechwiedzy logicznej w poszczególnych logikach epistemicznych jest – jednak! – istotną wadą tych logik, a następnie starać się budować logiki epistemiczne wolne od tych paradoksów.

Wymaga to jednak zastosowania jakichś "nowych" środków semantycznych i/lub odejścia od normalnych modalnych logik zdaniowych.

Modele Rantali

Fiński logik Veikko Rantala podał w 1982 r. pewną semantyczną metodę "blokowania" paradoksów wszechwiedzy logicznej – i zarazem semantyczną metodę budowy logik epistemicznych wolnych od tych paradoksów.

Mówiąc najogólniej, główny pomysł Rantali polega na wzięciu pod uwagę obok światów możliwych (*possible worlds*) również tzw. niemożliwych światów możliwych (*impossible possible worlds*).

Pojęcie niemożliwego świata możliwego (*sic*!) było już, nawiasem mówiąc, używane wcześniej, gdy budowano semantyki typu Kripkego dla logik istotnie słabszych od ***K***.

Intuicja leżąca u podstaw pojęcia niemożliwego świata możliwego jest następująca: jest to taki świat, w którym wszystko jest możliwe – nawet sprzeczność! – ale nic nie jest konieczne.

Dla uproszczenia załóżmy, że rozważamy jednopodmiotową zdaniową logikę przekonań, w której jedynym operatorem epistemicznym jest **B** (*believes*).

Niech ***W*** będzie niepustym zbiorem, natomiast $\boldsymbol{W}^{\#}$ będzie podzbiorem zbioru ***W***. Intuicyjnie rzecz biorąc, $\boldsymbol{W}^{\#}$ to zbiór niemożliwych światów możliwych, różnica $\boldsymbol{W} \setminus \boldsymbol{W}^{\#}$ to zbiór "możliwych" światów możliwych, natomiast ***W*** – to zbiór "możliwych i niemożliwych" światów możliwych.

Zasadnicza idea jest teraz następująca: w światach "niemożliwych" wszystkie formuły są traktowane jak "atomy". W konsekwencji wartość logiczna formuły złożonej w świecie "niemożliwym" nie zależy od wartości logicznych jej formuł składowych w tym świecie. Powoduje to, że – przykładowo - formuła o schemacie $A \wedge \neg A$ może być prawdziwa w świecie "niemożliwym" ***w***; w świecie takim może być prawdziwa implikacja $A \rightarrow B$, chociaż jest w nim prawdziwy jej poprzednik i fałszywy jest następnik *etc*.

Model Rantali to czwórka uporządkowana:

($) $\quad <\boldsymbol{W}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{W}^{\#}>,$

gdzie $\boldsymbol{W}$ i $\boldsymbol{W}^{\#}$ są rozumiane jak wyżej, a ponadto:

- $\boldsymbol{v}$ jest funkcją wartościowania o następujących własnościach:
  - $\boldsymbol{v}$ przyporządkowuje każdej parze $<p_i, \boldsymbol{w}>$, gdzie p_i jest zmienną zdaniową, a $\boldsymbol{w} \in \boldsymbol{W} \setminus \boldsymbol{W}^{\#}$ (tj. $\boldsymbol{w}$ jest "możliwym" światem możliwym), dokładnie jedną z wartości logicznych: **1**, **0**;
  - $\boldsymbol{v}$ przyporządkowuje każdej parze $<A, \boldsymbol{w}>$, gdzie A jest (dowolną) formułą, a $\boldsymbol{w} \in \boldsymbol{W}^{\#}$ (tj. $\boldsymbol{w}$ jest niemożliwym światem możliwym) dokładnie jedną z wartości logicznych: **1**, **0**.
- $\boldsymbol{R} \subseteq \boldsymbol{W} \times \boldsymbol{W}$ jest relacją (epistemicznej) alternatywności taką, że:
  - $\boldsymbol{R}$ jest seryjna w zbiorze $\boldsymbol{W}$,
  - $\boldsymbol{R}$ jest przechodnia i euklidesowa w zbiorze $\boldsymbol{W} \setminus \boldsymbol{W}^{\#}$ (tj. w zbiorze "możliwych" światów możliwych rozważanego modelu),

- a ponadto $\boldsymbol{R}$ spełnia pewne dalsze warunki o których powiemy za chwilę.

Pojęcie prawdziwości formuły A w świecie $\boldsymbol{w}$ modelu Rantali $\boldsymbol{M} = \langle \boldsymbol{W}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{W}^{\#} \rangle$, symbolicznie $\boldsymbol{M} \models^{\boldsymbol{w}} A$, określa się tak:

- jeśli $\boldsymbol{w} \in \boldsymbol{W}^{\#}$, to: $\boldsymbol{M} \models^{\boldsymbol{w}} A$ *wtw* $\boldsymbol{v}(A, \boldsymbol{w}) = \mathbf{1}$;
- jeśli $\boldsymbol{w} \in \boldsymbol{W} \setminus \boldsymbol{W}^{\#}$, to:

(1) $\boldsymbol{M} \models^{\boldsymbol{w}} p_i$ *wtw* $\boldsymbol{v}(p_i, \boldsymbol{w}) = \mathbf{1}$;

(2) $\boldsymbol{M} \models^{\boldsymbol{w}} \neg A$ *wtw* $\boldsymbol{M}$ *non* $\models^{\boldsymbol{w}} A$;

(3) $\boldsymbol{M} \models^{\boldsymbol{w}} (A \wedge B)$ *wtw* $\boldsymbol{M} \models^{\boldsymbol{w}} A$ *oraz* $\boldsymbol{M} \models^{\boldsymbol{w}} B$;

(4) $\boldsymbol{M} \models^{\boldsymbol{w}} (A \vee B)$ *wtw* $\boldsymbol{M} \models^{\boldsymbol{w}} A$ *lub* $\boldsymbol{M} \models^{\boldsymbol{w}} B$;

(5) $\boldsymbol{M} \models^{\boldsymbol{w}} (A \rightarrow B)$ *wtw* $\boldsymbol{M}$ *non* $\models^{\boldsymbol{w}} A$ *lub* $\boldsymbol{M} \models^{\boldsymbol{w}} B$;

(6) $\boldsymbol{M} \models^{\boldsymbol{w}} (A \leftrightarrow B)$ *wtw* $\boldsymbol{M} \models^{\boldsymbol{w}} A$ *zawsze i tylko, gdy* $\boldsymbol{M} \models^{\boldsymbol{w}^*} A$;

(7) $\boldsymbol{M} \models^{\boldsymbol{w}} \mathbf{B}A$ *wtw dla każdego* $\boldsymbol{w}^* \in \boldsymbol{W}$ *takiego, że* $\boldsymbol{w}\boldsymbol{R}\boldsymbol{w}^*$: $\boldsymbol{M} \models^{\boldsymbol{w}^*} A$.

Wspomniane wcześniej dodatkowe warunki nakładane na relację alternatywności ***R*** są następujące:

- dla każdego $w \in W \setminus W^{\#}$, dla każdego $w^* \in W^{\#}$: jeśli wRw^* oraz $v(\neg\mathbf{B}\neg A, w^*) = \mathbf{1}$, to dla pewnego $w^{**} \in W$ takiego, że wRw^{**} zachodzi $M \models^{w^{**}} A$.
- dla każdego $w \in W \setminus W^{\#}$, dla każdego $w^* \in W^{\#}$: jeśli wRw^* oraz $v(\mathbf{B}A, w^*) = \mathbf{1}$, to dla każdego $w^{**} \in W$ takiego, że wRw^{**} zachodzi $M \models^{w^{**}} A$.

Komentarz: zapraszam na wykład :)

Prawdziwość formuły w modelu Rantali postaci $\langle W, R, v, W^* \rangle$ jest definiowana (nieco) niestandardowo: formuła *A* jest prawdziwa w modelu Rantali $\langle W, R, v, W^{\#} \rangle$ wtw dla każdego $w \in W \setminus W^{\#}$ jest tak, że $M \models^{w} A$

-- tj. gdy *A* jest prawdziwa w każdym "możliwym" świecie możliwym rozważanego modelu.

Teraz możemy rozważyć klasę formuł, które są prawdziwe w każdym modelu Rantali. Będą to prawa logiki epistemicznej (tutaj: logiki przekonań) wyznaczonej przez daną klasę modeli Rantali.

Otóż do klasy tej należą m.in. formuły:

$\mathbf{4_B}$: $\mathbf{B}p \rightarrow \mathbf{BB}p$

$\mathbf{5_B}$: $\neg\mathbf{B}p \rightarrow \mathbf{B}\neg\mathbf{B}p$

Dla nas jednak istotniejsze jest to, które formuły *nie są* prawami logiki epistemicznej wyznaczonej przez analizowaną klasę modeli Rantali.

Rozważmy dla przykładu formułę:

$$\mathbf{B}p \rightarrow \mathbf{B}(p \vee q)$$

Weźmy model Rantali $\boldsymbol{M} = \langle \boldsymbol{W}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{W}^{\#} \rangle$ spełniający następujące warunki:

- $\boldsymbol{w} \in \boldsymbol{W} \setminus \boldsymbol{W}^{\#}$, $\boldsymbol{w}' \in \boldsymbol{W}^{\#}$ oraz $\boldsymbol{wRw}'$;
- $\boldsymbol{v}(p, \boldsymbol{w}^*) = \mathbf{1}$ dla każdego $\boldsymbol{w}^* \in \boldsymbol{W}$ takiego, że $\boldsymbol{wRw}^*$;
- $\boldsymbol{v}(p \vee q, \boldsymbol{w}') = \mathbf{0}$.

Wówczas mamy $\boldsymbol{M} \models^{w} \mathbf{B}p$ oraz $\boldsymbol{M}$ *non* $\models^{w} \mathbf{B}(p \vee q)$, czyli:

$$\boldsymbol{M} \textit{ non} \models^{w} (\mathbf{B}p \rightarrow \mathbf{B}(p \vee q)).$$

Skoro $w \in W \setminus W^{\#}$, znaczy to, że formuła:

$$\mathbf{B}p \rightarrow \mathbf{B}(p \vee q)$$

nie jest prawdziwa w pewnym modelu Rantali. Tak więc możemy powiedzieć, że zależność o schemacie:

LO4: $\vdash \otimes A \rightarrow \otimes(A \vee B)$

nie obowiązuje w logice epistemicznej wyznaczanej przez (rozważaną) klasę modeli Rantali.

W podobny sposób można pokazać, że w logice tej nie obowiązuje żadna z zależności o schematach **LO5**-**LO10**. Pozostawiam to Państwu jako ćwiczenie :)

Rozważmy teraz przypadek zależności o schemacie:

LO1: *jeśli* $\vdash A$, *to* $\vdash \otimes A$

Niech A ma postać $p \rightarrow p$. Bierzemy model Rantali $\boldsymbol{M} = \langle \boldsymbol{W}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{W}^{\#} \rangle$ i świat $\boldsymbol{w}' \in \boldsymbol{W}^{\#}$ taki, że $\boldsymbol{v}(p \rightarrow p, \boldsymbol{w}') = \mathbf{0}$. Jest oczywiste, że formuła $p \rightarrow p$ jest prawdziwa w każdym "możliwym" świecie możliwym modelu $\boldsymbol{M}$, tj. że $\boldsymbol{M} \models^{\boldsymbol{w}} (p \rightarrow p)$ dla każdego $\boldsymbol{w} \in \boldsymbol{W} \setminus \boldsymbol{W}^{\#}$. Jest niemniej oczywiste, że formuła $p \rightarrow p$, <u>podobnie jak każdy inny aksjomat rachunkowozdaniowy</u>, jest prawdą w każdym modelu Rantali, czyli wszystkie aksjomaty rachunkowozdaniowe są tezami logiki epistemicznej wyznaczonej przez (rozważane) modele Rantali. Przypuśćmy teraz, że $\boldsymbol{w}$ jest dowolnym ale ustalonym światem z $\boldsymbol{W} \setminus \boldsymbol{W}^{\#}$ oraz że $\boldsymbol{wRw}'$. Dostajemy:

$$\boldsymbol{M} \; non \models^{\boldsymbol{w}} \mathbf{B}(p \rightarrow p)$$

czyli kontrprzykład dla zależności o schemacie **LO1**.

W podobny sposób można pokazać, że w logice epistemicznej wyznaczonej przez (rozważane) modele Rantali nie obowiazują zależności o schematach **LO2** i **LO3**. I to pozostawiam Państwu jako ćwiczenie :)

Modele sitowe

Wśród formuł wynikających logicznie z danej formuły mamy zarówno takie, których podmiot jest świadomy (*aware*), jak i takie, których podmiot świadomy nie jest. Kolejna semantyczna metoda "blokowania" paradoksów wszechwiedzy logicznej nawiązuje do tej intuicji, zaliczając do wiedzy/ przekonań podmiotu tylko te konsekwencje logiczne aktualnie żywionych przekonań/ posiadanej wiedzy, których podmiot jest świadomy.

Metoda ta została zaproponowana przez Fagina i Halperna w 1988 r.

Podobnie jak poprzednio, dla uproszczenia rozważamy jednopodmiotową logikę przekonań z operatorem **B** (*believes*), ale bez operatora dualnego wobec **B**. Do języka wprowadzamy natomiast nowy operator **Ś**; formułę o schemacie **Ś***A* czytamy "podmiot jest świadomy tego, że *A*".

Modelem sitowym nazywamy czwórkę uporządkowaną:

($$) $\langle \boldsymbol{W}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{f} \rangle$

gdzie $\boldsymbol{W}$ jest niepustym zbiorem ("światów możliwych"), $\boldsymbol{R} \subseteq \boldsymbol{W} \times \boldsymbol{W}$ jest relacją epistemicznej alternatywności, która jest seryjna, przechodnia i euklidesowa w zbiorze $\boldsymbol{W}$, a $\boldsymbol{v}$ jest funkcją wartościowania, przyporządkowującą każdej parze postaci $\langle p_i, \boldsymbol{w} \rangle$, gdzie p_i jest zmienną zdaniową oraz $\boldsymbol{w} \in \boldsymbol{W}$, dokładnie jedną z wartości logicznych: **1** lub **0**, natomiast $\boldsymbol{f}$ jest funkcją o argumentach w zbiorze $\boldsymbol{W}$ i wartościach będących zbiorami formuł (intuicyjnie: funkcją, przyporządkowującą światom zbiór tych wszystkich formuł, których podmiot jest świadomy w tych światach), spełniającą następujący warunek:

- dla każdego $\boldsymbol{w} \in \boldsymbol{W}$: elementami $\boldsymbol{f}(\boldsymbol{w})$ są wszystkie podstawienia następujących formuł:
  - $\mathbf{B}p \rightarrow \neg\mathbf{B}\neg p$
  - $\mathbf{B}p \rightarrow \mathbf{BB}p$
  - $\neg\mathbf{B}p \rightarrow \mathbf{B}\neg\mathbf{B}p$.

Komentarz: oczywisty, ale i tak zapraszam na wykład :)

Prawdziwość formuły A w świecie $\boldsymbol{w}$ modelu sitowego $\boldsymbol{M} = \langle \boldsymbol{W}, \boldsymbol{R}, \boldsymbol{v}, \boldsymbol{f} \rangle$, symbolicznie: $\boldsymbol{M} \models^{\boldsymbol{w}} A$, definiujemy następująco:

(1) $\boldsymbol{M} \models^{\boldsymbol{w}} p_i$ *wtw* $\boldsymbol{v}(p_i, \boldsymbol{w}) = \mathbf{1}$;

(2) $\boldsymbol{M} \models^{\boldsymbol{w}} \neg A$ *wtw* $\boldsymbol{M}$ *non* $\models^{\boldsymbol{w}} A$;

(3) $\boldsymbol{M} \models^{\boldsymbol{w}} (A \wedge B)$ *wtw* $\boldsymbol{M} \models^{\boldsymbol{w}} A$ *oraz* $\boldsymbol{M} \models^{\boldsymbol{w}} B$;

(4) $\boldsymbol{M} \models^{\boldsymbol{w}} (A \vee B)$ *wtw* $\boldsymbol{M} \models^{\boldsymbol{w}} A$ *lub* $\boldsymbol{M} \models^{\boldsymbol{w}} B$;

(5) $\boldsymbol{M} \models^{\boldsymbol{w}} (A \rightarrow B)$ *wtw* $\boldsymbol{M}$ *non* $\models^{\boldsymbol{w}} A$ *lub* $\boldsymbol{M} \models^{\boldsymbol{w}} B$;

(6) $\boldsymbol{M} \models^{\boldsymbol{w}} (A \leftrightarrow B)$ *wtw* $\boldsymbol{M} \models^{\boldsymbol{w}} A$ *zawsze i tylko, gdy* $\boldsymbol{M} \models^{\boldsymbol{w}} B$;

(7) $\boldsymbol{M} \models^{\boldsymbol{w}} \acute{\mathbf{S}} A$ *wtw* $A \in \boldsymbol{f}(w)$;

(8) $\boldsymbol{M} \models^{\boldsymbol{w}} \mathbf{B} A$ *wtw* $A \in \boldsymbol{f}(w)$ *oraz dla każdego* $\boldsymbol{w^*} \in \boldsymbol{W}$ *takiego, że* $\boldsymbol{wRw^*}$: $\boldsymbol{M} \models^{\boldsymbol{w^*}} A$.

Prawdziwość w modelu sitowym to prawdziwość w każdym świecie tego modelu. Tezami logiki epistemicznej wyznaczonej przez (rozważaną) klasę modeli sitowych są formuły prawdziwe w każdym modelu sitowym.

Jak łatwo zauważyć, tezami analizowanej logiki są m.in. wszystkie aksjomaty rachunkowozdaniowe oraz wszystkie podstawienia formuł $\mathbf{D_B}$, $\mathbf{4_B}$ i $\mathbf{5_B}$. W logice tej *nie zachodzą* natomiast zależności o schematach **LO1**-**LO5** i **LO7**-**LO10**. Dlaczego? Cóż, zapraszam na wykład :)

***

Poza przedstawionymi wyżej, istnieje rzecz jasna wiele innych metod "blokowania" paradoksu wszechwiedzy logicznej – oraz wiele logik epistemicznych, w których nie występują co najmniej niektóre formy tego paradoksu. Logiki te są jednak zwykle bardziej skomplikowane – syntaktycznie i/lub semantycznie – niż logiki epistemiczne przedstawiane na tym cyklu wykładów.

Informacje na temat wielu z takich logik mogą Państwo zaczerpnąć np. z artykułu przeglądowego Gocheta i Gribomonta "Epistemic Logic".