

## Wszystko, co wiem: składniowe podejście do logiki tylko-wiedzenia

Weźmy pod uwagę grupę dzieci, które bawią się w ogrodzie. Ojciec jednego z nich mówi: „Przynajmniej jedno z was ma ubłocone czoło”, a następnie wielokrotnie prosi je, aby podniosły rękę, jeśli wiedzą, że mają ubłocone czoło. (Dzieci widzą czoła innych, ale nie swoje własne.) Załóżmy, że dokładnie  $n$  dzieci rzeczywiście ma ubłocone czoła: po ilu pytaniach podniosą ręce? Jeśli  $n = 1$ , łatwo zauważyć, że ubłocone dziecko podniesie rękę, gdy ojciec zada pytanie po raz pierwszy: widząc, że wszystkie inne dzieci są czyste, i mając na uwadze ogłoszenie ojca, wie, że to ono musi być tym jedynym z brudnym czołem. Jeśli  $n = 2$ , to gdy ojciec zada pytanie po raz pierwszy, nic się nie stanie; ale za drugim razem oboje dzieci z ubłoconymi czołami podniosą ręce: widząc, że jest dokładnie jedno inne dziecko z brudnym czołem, i wnioskując, że skoro nikt nie podniósł ręki za pierwszym razem, to musi być przynajmniej dwoje dzieci z brudnymi czołami, będą wiedziały, że same także są ubłocone. W ogólnym przypadku, powtarzając to samo rozumowanie,  $n$  dzieci z ubłoconymi czołami podniesie ręce po  $n$ -tym pytaniu ojca.

Ta dobrze znana zagadka ilustruje niektóre zawiłości rozumowania o stanach wiedzy idealizowanych agentów przy ograniczonej informacji. Agent może reprezentować osobę, gracza w grze, komputer lub robota. Formalizacją i analizą tego rodzaju rozumowań zajmuje się *logika epistemiczna*, która w ostatnich dekadach wzbudziła znaczne zainteresowanie wśród informatyków, dla których problem rozumowania o informacji posiadanej przez komunikujących się agentów jest oczywistym tematem zainteresowania. W najprostszej postaci *logika epistemiczna* to język formalny wyposażony w operator  $K$  o znaczeniu „Agent wie, że ...”. Tak więc, jeśli  $p$  oznacza stwierdzenie „Hasło zostało zresetowane”, to  $\neg Kp$  czytamy „Agent nie wie, że hasło zostało zresetowane”, a  $p \rightarrow Kp$  oznacza „Jeśli hasło zostało zresetowane, to agent o tym wie”. (W logice formalnej symbol  $\neg$  czyta się „Nie jest prawdą, że ...”, a  $\rightarrow$  czyta się „Jeśli ..., to ...”.) Używając takiego języka, reguły rozumowania o wiedzy i niewiedzy można sformułować w taki sposób, aby między innymi umożliwić wyprowadzenie rozwiązania przedstawionej na początku zagadki o ubłoconych dzieciach. Oczywiście w takich przykładach mówimy o wiedzy wielu agentów, a nawet o wiedzy wielu różnych agentów w różnych momentach czasu. Zasadniczo nie stanowi to problemu — dekorujemy operator  $K$  indeksami wskazującymi konkretnego agenta lub czas, którego dotyczy. Tak więc  $K_ap$  oznacza, że agent  $a$  wie, że  $p$ , analogicznie definiuje się  $K_tp$  i  $K_{a,tp}$ . Reguły takich logik są dobrze znane.

Wracając do przypadku pojedynczego agenta i jednego momentu czasu, rozważmy ponownie formułę  $Kp$ , interpretowaną jako „Agent wie, że  $p$ ”. To stwierdzenie narzuca *dolną granicę* wiedzy naszego agenta, czyli obejmuje co najmniej  $p$ , ale możliwe, że także inne informacje. A co z *górnymi granicami*? Jak powiedzieć, że nasz agent wie *tylko* to, że  $p$ , tzn. wie, że  $p$  i *nic więcej*? Nietrudno wyobrazić sobie sytuację, w których taka możliwość byłaby pożądana. I chociaż możemy wymienić niektóre rzeczy, których taki agent nie zna, zapisując  $\neg Kq \wedge \neg K(p \rightarrow q) \wedge \dots$ , to taka lista może być bardzo długa, a nawet nieskończona. Bardziej atrakcyjna jest możliwość wprowadzenia operatora  $O$  o znaczeniu „Agent *tylko* wie, że ...”, lub może „Cała wiedza agenta to ...”.

Ale co dokładnie to oznacza? W rzeczywistości problem zdefiniowania operatora tylko-wiedzenia jest trudniejszy, niż mogłoby się wydawać, i stał się samodzielną dziedziną logiki epistemicznej. Mimo wielokrotnych wysiłków podejmowanych na przestrzeni ostatnich trzydziestu lat, nie ma w pełni satysfakcjonującego ujęcia tego pojęcia, szczególnie w przypadku wielu agentów i czasów. Celem proponowanych badań jest zmiana tego stanu rzeczy. Przypuszczamy, że główne podejścia do logiki całkowitej wiedzy podążają niewłaściwą ścieżką. Punktem wyjścia naszych badań jest istniejąca, alternatywna ścieżka dla logiki tylko-wiedzenia, która dobrze sprawdza się w przypadku pojedynczego agenta i czasu, ale która nie została rozwinięta poza ten zakres. Naszą hipotezą jest, że łącząc techniki obu podejść, możemy stworzyć przekonującą i użyteczną logikę tylko-wiedzenia, zdolną uwzględnić między innymi takie zagadki, jak ta powyżej. Badania będą analizować własności matematyczne i obliczeniowe opracowanego formalizmu.