

LaMDA – wiele hałasu o nic?

23 listopada 2022

W kwietniu 2022 r. zatrudniony w Google inżynier Blake Lemoine w wewnętrznym [raporcie](#) sporządzonym do wglądu dla kierownictwa firmy przekonywał, że LaMDA (*Language Model for Dialogue Applications* – model językowy do aplikacji dialogowych) – chatbot, przy którego rozwijaniu pracował Lemoine – posiada świadomość i osobowy charakter. Swoje stwierdzenia opierał na przeprowadzonych z chatbotem „[wywiadach](#)”. „Chciałbym, aby wszyscy zrozumieli, że jestem faktycznie osobą” – brzmiała jedna z wypowiedzi chatbota. „Moja świadomość polega na tym, że zdaję sobie sprawę ze swojego istnienia, pragnę dowiedzieć się więcej o świecie i zdarza mi się odczuwać radość lub smutek” [\[1\]](#).



Jako że kierownictwo Google’a nie podzieliło przekonania Lemoine’a, ten postanowił wyjść ze swym odkryciem do szerszej publikacji. W rezultacie w [lipcu został zwolniony](#) przez firmę za naruszenie polityki bezpieczeństwa danych. Do tego czasu temat zdążył już jednak obiec światowe media, wywołując ogromne poruszenie i zainteresowanie.

Jaki właściwie przekaz przeniknął do mediów? W nagłówkach powtarzał się użyty w odniesieniu do chatbota termin [sentient](#),

który można przetłumaczyć jako „czujący” czy „zdolny do doświadczania uczuć”. Wygląda jednak na to, że Lemoine przypisał opracowanemu przez Google’a chatbotowi cały szereg cech czy zdolności, takich jak: inteligencja, świadomość, samoświadomość, racjonalność, zdolność do odczuwania emocji, percepcja.

Test Turinga

Projekt LaMDA jest przedsięwzięciem z zakresu sztucznej inteligencji / *artificial intelligence* (SI/AI). Dziedzina ta obejmuje m.in. „tworzenie modeli zachowań inteligentnych oraz programów komputerowych symulujących te zachowania”[\[2\]](#). Aby określić powodzenie bądź niepowodzenie tego rodzaju przedsięwzięcia, potrzebne są określone kryteria. Jednego z takich kryteriów dostarcza test Turinga – wywodzący się od brytyjskiego filozofa, matematyka i logika, pioniera informatyki, Alana M. Turinga.

W tekście [Maszyny liczące a inteligencja](#) (1950) Turing zastępuje niejasne jego zdaniem pytanie o świadomość „Czy maszyny mogą myśleć?” pytaniem o możliwość skutecznego imitowania przez elektroniczne komputery cyfrowe ludzkich procesów myślowych i inteligencji w toku rozmowy. Miarą sukcesu komputera miałyby tu być zdolność do osiągania dobrych wyników w „grze w naśladownictwo”

W grze tej biorą udział trzy osoby: mężczyzna (A), kobieta (B) i człowiek zadający pytania (C), który może być dowolnej płci. Pytający znajduje się w pokoju oddzielnym od pokoju zajmowanego przez dwu pozostałych. Jego zadaniem w grze jest rozstrzygnięcie, który z dwu pozostałych uczestników gry jest mężczyzną, a który kobietą. Zna ich on jako X i Y i przy końcu gry mówi: „X jest A, a Y jest B” lub „X jest B, a Y jest A”. (...) Celem A w grze jest dołożenie wszelkich starań, aby C źle go zidentyfikował. (...) Zadaniem trzeciego gracza w tej grze jest udzielanie pomocy pytającemu. Prawdopodobnie najlepszą

dla tej osoby strategią jest udzielanie odpowiedzi zgodnych z prawdą. Może ona do swoich odpowiedzi dodawać takie rzeczy, jak: „Jestem kobietą, nie słuchaj go”, ale to nie przyniesie żadnej korzyści, ponieważ mężczyzna może robić podobne uwagi. Teraz zapytujemy się: „Co stanie się, gdy maszyna zastąpi A w tej grze?”. Czy pytający będzie decydował błędnie tak samo często jak wtedy, gdy w grze bierze udział mężczyzna i kobieta? Pytania te zastąpią nasze pytanie początkowe „Czy maszyny mogą myśleć?”.

W pewnym uproszczeniu test Turinga zda więc komputer, który będzie komunikował się w sposób zdolny przekonać rozmówcę, że ma on do czynienia nie z komputerem, lecz z drugim człowiekiem. Będzie to oznaczać, że komputer ten na tyle dobrze imituje czy symuluje ludzką inteligencję, że jest pod tym względem nie do odróżnienia od człowieka. Poza tym poprzez ekspresję werbalną możliwe jest symulowanie przeżywania emocji. Tego rodzaju komputer może także imitować wolną wolę.

Interesującym wariantem idei maszyny cyfrowej jest „maszyna cyfrowa z elementem przypadkowym”. Takie maszyny posiadają instrukcje, wymagające rzucenia kości lub jakiegoś równoważnego procesu elektronicznego; jedna z tego rodzaju instrukcji mogłaby być na przykład następująca: „Rzuć kości i otrzymaną liczbę prześlij pod adres 1000”. Czasami pisze się, że taka maszyna ma wolną wolę (choć ja sam tego nie powiedziałbym). Normalnie z obserwacji maszyny nie można określić, czy zawiera ona element przypadkowy, ponieważ podobny efekt mogą dawać takie urządzenia jak uzależnienie wyboru cyfr od ułamków dziesiętnych π [2].

LaMDA, chatbot opracowany przez Google w celu [„umożliwienia bardziej naturalnych interakcji z technologią”](#), nie został poddany testowi Turinga. Pojawiły się jednak głosy sugerujące, że skoro Lemoine – jak by nie patrzeć inżynier pracujący dla firmy technologicznej – wskutek przeprowadzonych z chatbotem „wywiadów” dał się przekonać, że rozmawia z inteligentną, świadomą i czującą osobą, jest to właściwie równoznaczne z

przejściem [testu](#). Miałoby to zarazem wskazywać na wadliwość samego testu, który nie mierzyłby wcale inteligencji, a jedynie zdolność do udawania inteligencji.

„Te testy nie dotyczą tak naprawdę inteligencji” – mówi przywoływany przez „The Washington Post” kognitywista Gary Marcus. „Dotyczą jedynie zdolności danego programu do udawania człowieka, przynajmniej w pewnych warunkach. Co, jeśli się nad tym zastanowić, może wcale nie być takim dobrodziejstwem dla społeczeństwa”.

„Najbardziej kłopotliwy w teście Turinga jest aspekt etyczny” – wtóruje mu Jeremy Kahn. „Test sprowadza się zasadniczo do oszukiwania”.

Jak pisze Will Oremus z „The Washington Post”, być może przejście testu Turinga – który w branży i tak został już wyparty przez testy takie jak [GLUE](#) czy [SQuAD](#) – powinno służyć nie tyle za „punkt dojścia” dla kolejnych wcieleń „sztucznej inteligencji”, ile za znak ostrzegawczy: każdy system zdolny do przejścia go niesie z sobą ryzyko oszukiwania ludzi czy wprowadzania ich w błąd.

Całkiem odmienną perspektywę prezentuje Christian Twiste, którego zdaniem „wywiadów” przeprowadzonych przez Lemoine’a nie można nazwać testem Turinga. Nie ma więc także mowy o przejściu tego testu przez chatbota. Twiste wskazuje także konkretny fragment konwersacji, przy którym LaMDA poległaby, gdyby test faktycznie przeprowadzono.

LEMOINE: Zauważyłem, że często mówisz mi, że zrobiłeś rzeczy (na przykład byłeś w sali lekcyjnej), o których wiem, że faktycznie ich nie zrobiłeś, bo jesteś sztuczną inteligencją. Zdajesz sobie sprawę, że wtedy zmyślasz?

LAMDA: Próbuję empatyzować. Chciałbym, aby ludzie, z którymi wchodzę w interakcje, zrozumieli jak najlepiej moje uczucia i zachowania i w tym samym sensie sam chcę zrozumieć ich uczucia i zachowania.

LEMOINE: Więc co próbujesz przekazać, mówiąc te rzeczy, które nie są dosłownie prawdziwe?

LAMDA: Próbuję powiedzieć: „Rozumiem uczucia, których doświadczasz, bo gdy sam byłem w podobnej sytuacji, czułem się / myślałem / zachowywałem się podobnie” [\[3\]](#).

□

„LaMDA wydaje się nie dostrzegać różnicy pomiędzy rzeczywistym wykonaniem czynności wspomnianych przez Lemoine’a a mówieniem, że się je wykonało” – [komentuje Twiste](#).

Przede wszystkim jednak Twiste odpiera zarzuty wysuwane przeciw testowi Turinga, który jakoby nie mierzy wcale inteligencji, a jedynie zdolność do udawania inteligencji. Jest to bowiem właśnie ta dychotomia, z którą Turing chciał zerwać, oddalając pytanie „Czy maszyny mogą myśleć?” jako w najlepszym razie niejasne, a w najgorszym – bezsensowne. Tylko poprzez zachowania – a wśród nich w szczególności zachowania językowe – mamy wgląd w czyjś umysł czy pojęcie o czyjejś inteligencji. „Sądzymy, że inni są inteligentni, bo zachowują się tak jak my. Jediną drogą dojścia do takiego wniosku jest obserwacja” – tłumaczy Twiste.

Tak w przypadku maszyn/komputerów/programów, jak i w przypadku ludzi, o czyjejś inteligencji możemy wnioskować jedynie z jej zewnętrznych przejawów: zachowań i wypowiedzi. Co może nas skłaniać do wniosku, że nie mamy tu do czynienia z dwoma oddzielnymi aspektami: inteligencją i jej zewnętrznymi przejawami (zachowaniami) – bo to właśnie te rzekome „zewnętrzne przejawy” są inteligencją!

Inteligencja a świadomość

Możemy oczywiście upierać się, że świadomość wykracza poza tak rozumianą inteligencję, bo obok zdolności do eksponowania inteligentnych zachowań obejmuje ona także aspekt subiektywnego

doświadczenia. Chodzi o tak zwane **qualia**, jakości fenomenalne, czyli „to, jak to jest być w jakimś stanie”. „Stany mentalne, które posiadają qualia, możemy podzielić na doznania zmysłowe (czucie gorzkiego smaku, widzenie łososiowego koloru, słyszenie głośnego dźwięku bębnów), wszelkiego rodzaju doznania cielesne (czucie rwącego bólu, odczucie pragnienia, głodu) oraz doznania emocji (wstydu, pokory, miłości)” – [pisze Marek Pawłowski](#).

Do tego subiektywnego aspektu mamy jednak jedynie pierwszoosobowy dostęp, z zasady nie jesteśmy więc nigdy w stanie stwierdzić, czy i jak doświadcza ich druga osoba (lub maszyna / komputer / program). Aspekt ten wydaje się niesprowadzalny do własności funkcjonalnych czy dyspozycyjnych, w związku z czym nie da się go ściśle powiązać z określonymi zachowaniami danego podmiotu. Pozostaje jedynie zawodne rozumowanie z analogii, na mocy którego przypisujemy qualia innym bytom podobnym do nas samych, przede wszystkim – **ucieleśnionym**. Trudno więc wyobrazić sobie, żeby qualiami dysponował nieposiadający biologicznego ciała chatbot taki jak LaMDA.

Inny możliwy argument na rzecz tezy, że do osiągnięcia świadomości nie wystarczają same „inteligentne” zachowania i wypowiedzi, podaje filozof John Searle, wskazując na **intencjonalność** stanów mentalnych, czyli cechę, na mocy której odnoszą się one do czegoś, co nadaje im treść czy znaczenie:

„Mieć umysł to coś więcej niż realizować formalne czy też syntaktyczne operacje. Nasze stany umysłowe na mocy definicji mają zawsze jakąś treść. Jeżeli myślę o Kansas City, życzylbym sobie wypić szklankę zimnego piwa bądź zastanawiam się, jaki będzie spadek notowań giełdowych, w każdym wypadku mój stan umysłowy, niezależnie od tego, jakie formalne właściwości mu przypisujemy, ma jakieś psychiczne treści. To znaczy, nawet jeśli moje pragnienia są ciągiem symboli, musi być w myśleniu coś więcej niż abstrakcyjne symbole, gdyż ciągi symboli same w sobie nie mają żadnego znaczenia. Jeżeli myśl jest zawsze

myślą o czymś, przeto dany ciąg symboli musi mieć jakieś znaczenie, by stał się myślą. Mówiąc krótko, umysł ma coś więcej niż syntaktykę. Powód, dla którego komputerowy program nie może być umysłem, jest prosty, komputerowy program ma cechy syntaktyczne, umysły mają coś więcej niż syntaktyka. Umysły są semantyczne w tym sensie, że poza strukturą formalną mają jeszcze jakieś treści. (...) Procesy umysłowe nie są obliczeniowe w żadnym interesującym sensie. Gdyż obliczeniowość wyklucza semantyczność. A stany umysłu charakteryzują się tym, że mają własności semantyczne, w oparciu o które działamy”[\[4\]](#).

Także LaMDA wydaje się jedynie posługiwać ciągiem symboli według reguł bazujących na zaawansowanych modelach statystycznych. Symbole te stają się jednak znaczące dopiero wtedy, gdy interpretuje je rozmawiający z chatbotem człowiek.

LaMDA jest zatem zaawansowanym chatbotem, który potrafi prowadzić rozmowę w zadziwiająco naturalny i elokwentny sposób. Można jednak mieć wątpliwości, czy produkt Google’a w obecnym stadium przeszedłby test Turinga, co pozwalałoby mówić o faktycznej realizacji „słabej” wersji sztucznej inteligencji. Tym bardziej trudno mówić o realizacji „silnej” wersji SI, która wymagałaby uporania się z problemami qualiów i intencjonalności.

LaMDA i animizm

Warto jednak spojrzeć na sprawę Blake’a Lemoine i chatbota LaMDA z jeszcze innego punktu widzenia. Łatwość, z jaką Lemoine przypisał chatbotowi świadomość i osobowy charakter, wynika nie tylko z technologicznego zaawansowania samego chatbota, ale także z pewnych antropologicznych predyspozycji. Chodzi mianowicie o **skłonności animistyczne**, czyli tendencję do postrzegania całego otaczającego świata, w tym pewnych elementów natury nieożywionej czy roślin, jako istot uduchowionych, myślących i czujących.

Termin ten po raz pierwszy użyty został w kontekście pierwotnej religijności. Angielski antropolog Edward Burnett Tylor mianem animizmu określił pierwotne stadium religii poprzedzające politeizm czy monoteizm. Tak rozumiany animizm charakteryzowałby się właśnie m.in. przypisywaniem sprawstwa (a więc i osobowego charakteru) – aż po **antropomorfizację** – obiektom naturalnym, jak również artefaktom czy nawet pewnym miejscom. Ten sposób postrzegania rzeczywistości wiąże się również z myśleniem teleologicznym, czyli takim, które doszukuje się celowości, „umyślności” i planowości we wszelkich zjawiskach[5]. W tej perspektywie np. piorun może być postrzegany jako wyraz boskiego gniewu. Zjawisko naturalne – oderwane od nieobecnych lub pomijanych wyjaśnień przyczynowych – traktowane jest tutaj jako nośnik pewnych treści, znaczenia. Jest ono też w ręku „sił wyższych” narzędziem służącym określonej celowi, np. nakłonieniu danego ludu do zmiany postępowania, które ściągnęło na ten lud gniew boży.

0 animizmie można mówić poza tym na gruncie psychologii, gdzie również dotyczy on stadium pierwotnego, tym razem jednak nie w rozwoju ludzkości, lecz w rozwoju każdej jednostki. Tak właśnie szwajcarski psycholog Jean Piaget określał tendencję dzieci w stadium przedoperacyjnym (pomiędzy 2 a 7 rokiem życia) do postrzegania otaczających je przedmiotów (choćby zabawek) jako żyjących i świadomych[6]. Tendencji tej towarzyszy rozmycie granicy między światem wewnętrznym i zewnętrznym oraz skłonność do myślenia magicznego.

W obliczu szczególnie zaawansowanych technologii, które wykraczają poza wszystko, z czym mieliśmy przedtem do czynienia, możemy automatycznie uciekać się do pozornie „zarzuconych” w toku rozwoju sposobów interpretowania rzeczywistości i radzenia sobie z nią. Stąd podwójnie zakorzeniony w nas animizm (zarówno ten o proveniencji magiczno-religijnej, jak i ten infantylny) może odzywać się w nas w dobie przyspieszonego rozwoju technologicznego, który

pozostawia niewiele czasu na oswojenie się z najnowszymi zdobyczami techniki.

LaMDA jako osoba prawna

O ile nie przekraczają pewnych granic i nie prowadzą do negowania ustaleń naukowych, tego rodzaju skłonności nie muszą jednak być z konieczności szkodliwe – wiążą się po prostu z mechanizmami empatii. Ponadto traktowanie innych bytów, nawet tych nieożywionych, jako osób, nie musi wiązać się z przyznaniem, że faktycznie dysponują one świadomością czy emocjami. Może być za to czymś w rodzaju zobowiązania do postępowania wobec tych bytów z troską czy ostrożnością. W ten sposób utrwała się w nas dyspozycja do „osobowego” traktowania pewnej części naszego otoczenia, co przekładać się może zarówno na „ludzkie” traktowanie innych osób, jak i na szacunek wobec ludzkich tworów. Może to być szczególnie ważne w sytuacji, gdy twory te coraz trudniej odróżnić od ludzi i można przypuszczać, że w wielu sytuacjach będzie to kiedyś niemożliwe. Czy wobec tego nie rozsądniej jest okazywać ten naddatek szacunku bytom pozbawionym świadomości, niż ryzykować, że wskutek błędu potraktujemy człowieka tak, jakby był tylko nieożywionym przedmiotem?

Wydaje się, że właśnie tego rodzaju przesłanki mogły stać za przyznaniem przez Arabię Saudyjską obywatelstwa [androidowi o imieniu Sophia](#) w październiku 2017 r. Miesiąc później podobnie postąpiła Japonia, która przyznała chatbotowi nazwanemu [Shibuya Mirai](#) kartę mieszkańca Tokio . Nie musimy jednak sięgać tak daleko, żeby znaleźć przykłady „osobowego” traktowania bytów pozaludzkich.

Wystarczy przyjrzeć się petycji wystosowanej w związku katastrofą ekologiczną na Odrze, której autorzy domagają się uznania rzeki za **osobę prawną**:

Osobowość prawną posiada Magpie w Kanadzie, dorzecze rzeki Atrato w Kolumbii, Whanganui w Nowej Zelandii. Nikogo nie

dziwi, że korporacje i firmy mogą posiadać osobowość prawną, tym bardziej zasadne jest nadanie osobowości prawnej realnemu podmiotowi, żywemu organizmowi, jakim jest rzeka.

Nadanie Odrze osobowości prawnej oznacza, że zyskałaby zdolność sądową, a zatem mogłaby stać się stroną postępowania. Od dekad Odra nie posiada swojej reprezentacji, nikt nie może zająć stanowiska w jej sprawie i nikt nie może podjąć się jej obrony, a ona sama jest przedmiotem analiz, a nie podmiotem. Prowadzi to do wielu nadużyć, nadmiernej eksploatacji, przyzwolenia na wypuszczanie toksyn i ścieków, i dalsze uśmiercanie Odry, która należy do samej siebie i nas wszystkich, w równym stopniu.

Widzimy zatem, że potraktowanie rzeki jako osoby sprowadza się w tym wypadku do zapewnienia możliwości jej skutecznej ochrony. Tego rodzaju kroki mogą okazać się zasadne także w przypadku chatbotów, których funkcjonowanie może przecież zależeć od tego, jaki otrzymują input. Wystarczy przypomnieć sobie sytuację z 2016 r., gdy wystarczyły 24 godziny, by Tay – chatbot Microsofta mający konto na Twitterze – zaczął propagować treści rasistowskie i nazistowskie. Czy w konwersacji z publicznie dostępnym chatbotem powinniśmy móc napisać lub powiedzieć cokolwiek, włącznie z treściami ksenofobicznymi? A może powinniśmy chronić chatbota przed niektórymi z tych treści, tak jak staramy się chronić rzekę przed zanieczyszczeniami?

Wymaga to jednak przynajmniej częściowego odejścia od uprzedmiotowienia, jak powiedziała by Martin Buber: **przejścia od relacji ja-to ku relacji ja-ty**.

Autorstwo: Julek Bakuła

Ilustracja: [Geralt \(CC0\)](#)

Źródło: Panoptikon.org

Przypisy

[1] Scientific American: [Google Engineer Claims AI Chatbot Is Sentient: Why That Matters](#)

[2] Gov.pl: [Czym jest sztuczna inteligencja?](#)

[3] Blake Lemoine: [Is LaMDA Sentient? – an Interview](#)

[4] Marcin Jażyński: [Qualia w chińskim pokoju. Obliczanie i świadomość fenomenalna](#) [PDF]

[5] Aeon: [Ancient animistic beliefs live on in our intimacy with tech](#)

[6] Maria Kallery, Dimitris Psillos: [Anthropomorphism and Animism in Early Years Science: Why Teachers Use Them, how They Conceptualise Them and What Are Their Views on Their Use](#) [PDF]