Systemy certyfikacji AI jeszcze nie istnieją. Prof. Marta Kwiatkowska o granicach zaufania do algorytmów

Kategoria: Aktualności
Opublikowano: czwartek, 10, wrzesień 2026 07:29
Joanna Gryboś-Chechelska

Czy potrafimy udowodnić, że system oparty na sztucznej inteligencji zachowa się bezpiecznie w sytuacji granicznej – na drodze, w szpitalu, wobec cyberataku? Wokół tego pytania toczyła się rozmowa o matematycznej weryfikacji AI, o granicach zaufania do algorytmów i o braku systemów certyfikacji dla technologii, która rozwinęła się szybciej niż narzędzia jej kontroli. Debata odbyła się podczas XXXV Forum Ekonomicznego w Karpaczu.

Spotkanie prowadziła Magdalena M. Baran, filozofka i publicystka, która przedstawiła swoją rozmówczynię i wskazała, że badania jej gościa wykraczają poza popularne pytanie o to, jak korzystać z narzędzi AI, i dotyczą tego, czy sztuczna inteligencja może być bezpieczna, odporna i przewidywalna nawet w tym świecie, który jest bardzo chaotyczny.

Gościem była prof. Marta Kwiatkowska, informatyczka z Uniwersytetu Oksfordzkiego, członkini Polskiej Akademii Nauk, światowa liderka w dziedzinie weryfikacji systemów losowych sztucznej inteligencji opartych na sieciach neuronowych. Wyjaśniła, że niemal całą karierę poświęciła weryfikacji formalnej systemów. Badała między innymi protokoły Bluetooth i systemy informatyczne, sprawdzając matematycznymi narzędziami, czy spełniają zadaną specyfikację – na przykład czy transmisja nastąpi w wymaganym czasie. Od 2016 roku przenosi te metody na sztuczną inteligencję, zaczynając od systemów rozpoznawania obrazów, w tym rozpoznawania twarzy, a nie od modeli językowych. Podkreśliła, że jej narzędzia mają szczególne znaczenie tam, gdzie wdrażający system ponosi odpowiedzialność prawną za jego zachowanie.

Zapytana o zastosowanie systemu w sytuacjach kryzysowych, Kwiatkowska zastrzegła, że opiera się na scenariuszach modelowych. Stawka jest jednak wysoka i całkowicie realna – ryzyko może pociągać za sobą konsekwencje prawne, a nawet zagrażać ludzkiemu życiu. Badaczka wyodrębniła dwa kluczowe problemy: reakcję systemu na bezpośredni atak (gdzie właściwym rozwiązaniem bywa bezpieczne, automatyczne wyłączenie) oraz szerzej pojętą odporność całej struktury.

Baran przełożyła te rozważania na konkretne przykłady – samochody autonomiczne oraz medycynę – pytając, jak system powinien się zachować w sytuacji utraty danych. Kwiatkowska wskazała motoryzację jako modelowy obszar, w którym krzyżują się awarie i zagrożenia. Przypomniała o udokumentowanych przypadkach włamań do wewnętrznych sieci pojazdów, co pokazuje, że cyberbezpieczeństwo stanowi realne wyzwanie. Omówiła wprowadzane w Europie systemy wsparcia kierowcy (ADAS), które wykrywają przeszkody i wymuszają zatrzymanie pojazdu. Zwróciła też uwagę na funkcję rozpoznawania znaków drogowych, która – jak zauważyła na podstawie własnych doświadczeń z wynajmowanymi samochodami – wciąż nierzadko zawodzi.

Prowadząca nawiązała do testów pojazdów autonomicznych w Niemczech oraz do bezzałogowych systemów wojskowych. Kwiatkowska potwierdziła obecność technologii AI na polu walki, natomiast w transporcie cywilnym wskazała firmę Waymo założoną przez Google. Jej pojazdy – wyposażone w systemy LIDAR, liczne sensory i kamery – kursują już bez kierowcy w San Francisco, Los Angeles i Teksasie. Z kolei w Londynie, jak zauważyła, wdrożenie tej technologii wymaga jeszcze obecności człowieka. Brytyjskie prawo nie dopuszcza pełnej autonomii, dlatego na pokładzie musi znajdować się kierowca, którego rola ogranicza się do monitorowania jazdy.

Najwięcej uwagi rozmówczynie poświęciły medycynie. Kwiatkowska zauważyła, że w diagnostyce rzadko trafiają się sytuacje czarno-białe, podczas gdy standardowo trenowane modele podają tylko jedno, kategoryczne rozstrzygnięcie. Choć sieci neuronowe analizują obrazy radiologiczne szybciej i dokładniej

niż ludzie, to właściwe rozpoznanie zmiany nowotworowej wciąż wymaga od lekarza znajomości całego kontekstu pacjenta. Aby zapobiec nadmiernej, nieuzasadnionej pewności algorytmów, zespół badaczki modeluje niepewność za pomocą systemów losowych. Najlepsze efekty przynosi jednak współpraca: połączenie sztucznej inteligencji z wiedzą jednego lub dwóch lekarzy pozwala na sprawny i bezpieczny triaż. Baran podsumowała ten mechanizm, nazywając człowieka niezbędnym „zaworem bezpieczeństwa".

Kwiatkowska dodała, że badania nad współpracą medyków z AI ujawniają wyraźny deficyt zaufania, wynikający z faktu, że systemy nie potrafią uzasadnić swoich decyzji. Z tego powodu dynamicznie rozwija się nurt zajmujący się interpretowalnością sztucznej inteligencji. Kwestia zaufania dotyczy zresztą nie tylko lekarzy, ale też pasażerów autonomicznych aut czy dzieci wchodzących w interakcje z robotami.

Zapytana o certyfikację, ekspertka odpowiedziała stanowczo: narzędzia kontroli nie nadążają za technologią i systemy certyfikacji AI po prostu jeszcze nie istnieją. Stworzenie takich standardów wymaga współdziałania prawników, inżynierów systemów krytycznych oraz wprowadzenia niezależnych testów, prowadzonych przez podmioty zewnętrzne, a nie samych producentów. Wskazała także na potrzebę zaangażowania filozofów, przywołując jako przykład oksfordzki kierunek studiów łączący informatykę z filozofią – co uzasadnia rosnąca niezależność autonomicznych agentów AI.

W kwestii odpowiedzialności człowieka Kwiatkowska przyznała, że nie jest to prosta sprawa. Przedstawiła AI jako narzędzie, powołując się na amerykański raport Arvinda Narayanana i Sayasha Kapoora. Autorzy opracowania traktują AI jak każdą inną standardową technologię – jest to zarazem opis stanu faktycznego, prognoza oraz zalecenie, by to człowiek zachował nad nią pełną kontrolę. Zgodnie z modelem dyfuzji innowacji, badaczka umieściła sztuczną inteligencję głównie na etapie laboratoriów i wynalazków, przed fazą masowych wdrożeń. Analogii szukała w lotnictwie: początkowej ekscytacji nowym wynalazkiem również towarzyszył brak regulacji, a surowe przepisy pojawiły się dopiero po pierwszych katastrofach. W przypadku AI nie zaczęliśmy jeszcze myśleć o tym na serio – oceniła.

Ostatni wątek dotyczył dezinformacji. Kwiatkowska uznała ten problem za nierozwiązany. Tłumaczyła, że modele językowe nie weryfikują faktów, ponieważ ich nie znają, a halucynacje i błędne odpowiedzi mogą wynikać z samego procesu uczenia, a nie tylko ze złośliwej manipulacji. Częściowym rozwiązaniem są mechanizmy zabezpieczające – cyfrowi „strażnicy", którzy weryfikują generowane treści i blokują te nieodpowiednie. Narzędzia te wymagają jednak stałego korygowania.

Rozmówczynie zgodziły się, że sztuczna inteligencja pozostaje jedynie narzędziem zależnym od ludzkiego nadzoru, a warunkiem jej bezpiecznego rozwoju są weryfikacja, interpretowalność i niezależne testy. Kwestia odpowiedzialności pozostaje otwarta: brak systemów certyfikacji, ram prawnych i spójnych rozstrzygnięć etycznych sprawia, że technologia wciąż wyprzedza regulacje.

Więcej informacji na stronie www.forum-ekonomiczne.pl

Związek Powiatów Polskich jest partnerem instytucjonalnym wydarzenia.

Źródło: www.forum-ekonomiczne.pl