NAUKA 2/2024 • 91–94 DOI: 10.24425/nauka.2024.151207

MICHAŁ KLEIBER*

Niekontrolowany rozwój AI jest zagrożeniem dla ludzkości**

Wielu ekspertów pisze i mówi na temat sztucznej inteligencji, używając sformułowania „AI – korzyści czy zagrożenia?". Uważam takie alternatywne sformułowanie za nietrafne, bowiem właściwa według mnie jest koniunkcja „AI – korzyści i zagrożenia". W świetle dotychczasowych osiągnięć jest bowiem niezaprzeczalnym faktem, że AI przynosi nam już dzisiaj wiele korzyści, których będzie znacznie więcej już w nieodległej przyszłości – w gospodarce, ochronie zdrowia, badaniach naukowych, w istocie w prawie wszystkich dziedzinach naszego życia. Stosowanie AI z pewnością pozwoli także skutecznie stawiać czoła wyzwaniom dotychczas jeszcze nieznanym bądź uznawanym dzisiaj za niepokonywalne. Równocześnie jednak AI tworzy już obecnie widoczne olbrzymie zagrożenia, których niestety także będzie znacznie więcej w przyszłości. Przyszłe korzyści dużo łatwiej jest przy tym planować i przewidywać, bowiem dostrzeganie i pełne rozumienie zagrożeń, tworzonych z reguły w tajemnicy i najczęściej przez autorytarne państwa, jest niezwykle skomplikowane. Tezę o powadze zagrożeń potwierdza dobitnie ostatni raport zlecony grupie ekspertów przez amerykański State Department, czyli ministerstwo spraw zagranicznych. Raport bazuje na głębokich, merytorycznych rozmowach przeprowadzonych z ponad 200 ekspertami – szefami wiodących firm rozwijających AI, renomowanymi badaczami problemów cyberbezpieczeństwa, specjalistami badającymi konsekwencje użycia broni masowego rażenia oraz wysokiej rangi pracownikami ministerstwa. Wśród rozmówców byli m.in. szefowie i główne osoby odpowiedzialne za rozwój technologii w takich firmach jak OpenAI (właściciel systemu ChatGPT), Google DeepMind, Meta czy Anthropic. Zarysowany w raporcie obraz sytuacji jest prawdziwie alarmujący, a jego podstawową tezą jest stwierdzenie o zagrożeniu dla przetrwania nawet całej ludzkości, mogącym powstać w wyniku niekontrolowanego dalszego rozwoju AI. Obawy dodatkowo zwiększyła wypowiedź Geoffreya Hintona, uważanego za „ojca" AI, który stwierdził w niedawnym wywiadzie, że możliwe zbrodnicze wykorzystanie AI w nadchodzących trzech dekadach ma co najmniej 10% szansy na całkowite unicestwienie ludzkości. Odnosząc się do tych wydarzeń, rzecznik Białego Domu oświadczył zdumiewająco dobitnie, że prace nad uregulowaniem zasad

* Prof. dr hab. Michał Kleiber (michal.kleiber@pan.pl), członek rzeczywisty PAN, Instytut Podstawowych Problemów Techniki PAN, Warszawa

** Artykuł opublikowano na portalu „Wszystko co Najważniejsze" 5 kwietnia 2024 r.

www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

tworzenia i korzystania z AI stają się najważniejszym zadaniem dla amerykańskiego rządu i wszystkich jego zagranicznych partnerów.

Wspomniany raport analizuje dwa główne zagrożenia dotyczące zastosowania AI. Pierwsze dotyczy planowanego, niekoniecznie w globalnej skali, ale szeroko niszczycielskiego użycia AI w tworzonej różnorodnej broni, w szczególności broni całkowicie autonomicznej, a także biologicznej i chemicznej, drugie zaś utraty ludzkiej kontroli nad opracowywanymi w laboratoriach systemami o ogromnie niszczącym potencjale. W obu przypadkach naruszone może zostać globalne bezpieczeństwo, wywołując wyścig nowoczesnych zbrojeń o wręcz nieprzewidywalnych, dramatycznych konsekwencjach. Wśród zagrożeń o katastroficznych wymiarach raport wymienia także użycie AI do zaprojektowania i implementacji cyberataku całkowicie unieruchamiającego infrastrukturę państwa. A także możliwości przeprowadzenia w olbrzymiej skali kampanii dezinformacyjnej mającej potencjał destabilizacji całego społeczeństwa i podważenia podstawowych zasad funkcjonowania państwa. Wobec tak poważnych i przekonująco zidentyfikowanych zagrożeń raport apeluje o natychmiastowe podjęcie konkretnych działań, takich jak powołanie międzynarodowej agencji kontroli kierunków rozwoju AI, ustanowienie awaryjnych zasad działania w przypadku rozpoczęcia konfliktu z szerokim wykorzystywaniem AI oraz wprowadzenie międzynarodowych ograniczeń na charakter pozyskiwania wiedzy przez systemy AI. Ta ostatnia sprawa ma wielkie znaczenie, bowiem dynamika narastającej konkurencji głównych producentów systemów wyposażonych w AI wpływa na ograniczenie stosowanych zasad bezpieczeństwa, stwarzając możliwości kradzieży oprogramowania w celu zbrodniczego wykorzystania, w dodatku bez możliwości przewidzenia przez złoczyńców możliwych konsekwencji jego użycia. Fakt, iż poważny amerykański raport już w swym tytule ostrzega przed możliwością zniszczenia przez AI całej ludzkości nie powinien pozostawiać wątpliwości co do powagi problemu, przed którym stoi dzisiaj cały świat.

Przy tej tak dobitnie wyartykułowanej powadze wymienionych zagrożeń niełatwo pogodzić się z faktem, że stanowią one niestety tylko część problemów dotyczących naszego bezpieczeństwa w pełnym znaczeniu tego terminu. Niektóre z tych problemów ze względu na ich drugie, nie do zakwestionowania pozytywne oblicze, są w opinii publicznej praktycznie niezauważalne. Spośród naprawdę wielu spraw mających taki charakter przytoczmy choćby najnowsze osiągnięcie firmy Neuralink w postaci bezprzewodowego chipa wszczepionego do ludzkiego mózgu. Firma ta jest startupem założonym w 2017 roku przez Elona Muska, której głównym celem jest stworzenie interfejsu mózg-komputer, umożliwiającego w szczególności sterowanie za pomocą myśli urządzeniami takimi jak komputer czy telefon. Pełna realizacja tego pomysłu byłaby oczywiście bardzo przydatna dla osób sparaliżowanych po urazach bądź niesłyszących lub niewidzących. Urządzenie Neuralink zawiera ponad 1000 elektrod, które po wszcze-

www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

pieniu do mózgu człowieka za pomocą specjalnie zaprojektowanego chirurgicznego robota oddziałują na poszczególne neutrony. Dzięki bezprzewodowej komunikacji sygnały mózgowe przesyłane są do specjalnego programu dekodującego je na intencje i działania pacjenta. Po licznych próbach na zwierzętach, w fazie rekonwalescencji jest pierwsza osoba ze wszczepionym implantem. Efekt wydaje się być spektakularny – całkowicie sparaliżowany pacjent jest w stanie grać w szachy i korzystać z gier wideo, używając wyłącznie swych myśli. Nauralink planuje przeprowadzić w tym roku 11 operacji tego typu i doprowadzić do komercjalizacji urządzenia w ciągu paru nadchodzących lat. Wobec znaczących potencjalnych korzyści, takich jak: wspomaganie osób niepełnosprawnych, ułatwienie korzystania z Internetu, telefonu i wielu innych urządzeń, poprawie zdolności uczenia się i zapamiętywania nowych informacji, nie wolno jednak i tu zapomnieć o możliwych zagrożeniach. Należą do nich z pewnością możliwe i na razie nieznane komplikacje po wszczepieniu urządzenia, trudności w razie potrzeby jego naprawy bądź usunięcia, brak informacji o długoterminowych konsekwencjach zabiegu, w szczególności psychologicznych, takich jak trwałe uzależnienie, niepokój czy depresja oraz – co jest chyba szczególnie niepokojące – możliwości wykorzystywania przez osoby niepożądane dostępu do myśli człowieka w celu naruszenia jego prywatności, szerokiej kontroli zachowania bądź prowadzenia szantażu. Jeśli więc szeroka, akcentująca korzyści promocja doprowadzić miałaby do powszechności w korzystaniu z tego typ urządzeń – a są już inne firmy zaawansowane w procesie ich tworzenia – możliwości manipulowania opinią publiczną stałyby się wręcz nieograniczone, przynosząc konsekwencje, których nie sposób nawet dzisiaj przewidzieć.

Wracając na koniec do wspomnianej kwestii niezbędnych regulacji mających na celu bezpieczny rozwój AI i korzystne dla społeczeństw jej zastosowania, nie można nie wspomnieć o pierwszej na świecie takiej próbie, podjętej ostatnio przez Unię Europejską. Przyjęte rozporządzenie koncentruje się na problemach etycznych, zakazując zastosowań AI zagrażających prawom obywateli takich, jak manipulowanie ludźmi na bazie wygenerowanych z ich danych słabości, wykorzystywanie danych biometrycznych dotyczących cech wrażliwych czy pobieranie wizerunków z dostępnych źródeł w celu użycia ich do rozpoznawania twarzy. Istotne obowiązki nałożone są także na instytucje stosujące systemy AI w obszarach mogących stanowić zagrożenia dla zdrowia, bezpieczeństwa bądź podstawowych praw obywatelskich. Dotyczy to w szczególności edukacji, ochrony zdrowia i sposobów zatrudniania pracowników. Przyjęte regulacje są próbą znalezienia remedium na niektóre z możliwych zagrożeń, ale nie są oczywiście kompletne ze względu na wiele innych, ciągle słabo rozpoznanych problemów. Trzeba przy tym także pamiętać, że niekorzystną konsekwencją tych i przyszłych, podobnych regulacji może okazać się spowolnienie rozwoju AI w globalnych firmach, powodujące ograniczenie innowacyjności w światowej gospodarce. Pogodzenie tych dwu spraw, ochrony

www.czasopisma.pan.pl PAN POLSKA AKADEMIA NAUK www.journals.pan.pl

bezpieczeństwa z jednej strony, a niezbędnej dla rozwoju innowacyjności z drugiej, jawi się dzisiaj jako wyzwanie o olbrzymich konsekwencjach dla nas wszystkich. Nie powinno to jednak podważać przekonania, że prace nad uregulowaniem problematyki AI muszą być szybko i zdecydowanie prowadzone we wszystkich krajach, oczywiście z uwzględnieniem szerokiego kontekstu międzynarodowego.

Niekontrolowany rozwój AI jest zagrożeniem dla ludzkości

W artykule omówiono konkluzje raportu amerykańskiego Departamentu Stanu dotyczącego zagrożeń wynikających z nieuregulowanego rozwoju sztucznej inteligencji. Raport ostrzega bardzo dobitnie przed najróżniejszymi przyszłymi zagrożeniami stojącymi przed całym światem w wyniku niekontrolowanego rozwoju nowych technologii informatycznych, mocno akcentując w szczególności możliwość zniszczenia całej ludzkości przez globalnie niekontrolowane zastosowania AI. Druga część artykułu dotyczy problematyki mającej potencjalnie również wielkie znaczenie dla naszej przyszłości, a mianowicie nowego urządzenia łączącego mózg z komputerem. Urządzenie to o nazwie Neuralink przesyła bezprzewodowo sygnały mózgowe do specjalnego programu, który je dekoduje, rozpoznając ludzkie myśli. Przy wielu niezaprzeczalnych zaletach olbrzymim zagrożeniem są możliwości wykorzystywania przez osoby niepożądane dostępu do myśli człowieka w celu naruszenia jego prywatności. Konkluzją artykułu jest stwierdzenie o konieczności prowadzenia szerokich prac dotyczących regulacji mających na celu bezpieczny rozwój AI i korzystne dla społeczeństw jej zastosowania.

Słowa kluczowe: sztuczna inteligencja, globalne zagrożenia, interfejs mózg-komputer, potrzeba globalnych regulacji

The uncontrolled development of AI is a threat to humanity

The article discusses the conclusions of the U. S. State Department report on the threats resulting from the unregulated development of artificial intelligence. The report warns very emphatically of various future threats facing the world as a result of the uncontrolled development of new information technologies, strongly emphasizing in particular the potential for the destruction of all humanity by globally uncontrolled applications of AI. The second part of the article deals with an issue that is potentially also of great importance for our future, namely a new device that connects the brain with the computer. This device, called Neuralink, wirelessly transmits brain signals to a special program that decodes them to recognize human thoughts. With many undeniable advantages, a huge threat is the ability of undesirable persons to exploit access to a person's thoughts in order to violate his privacy. The conclusion of the article is that it is necessary to carry out extensive work on regulations aimed at the safe development of AI and its use beneficial for society.

Key words: Artificial Intelligence, Global Threats, Brain-Computer Interface, Need for Global Regulation