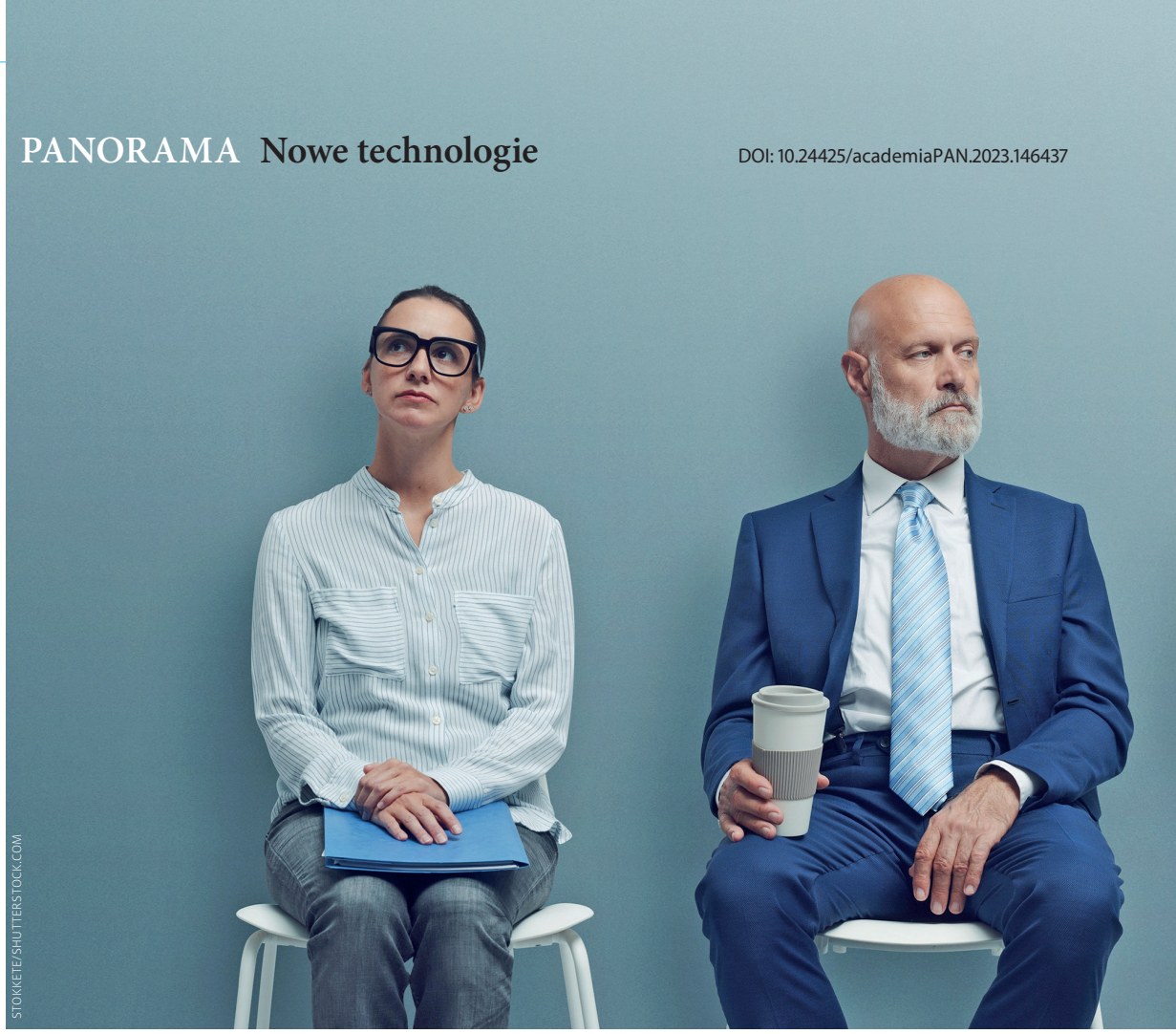
DOI: 10.24425/academiaPAN.2023.146437



STOKKETE/SHUTTERSTOCK.COM

MACIEK BERNAS

dr Piotr Kaczmarek-Kurczak
Wykładowca w Katedrze Przedsiębiorczości i Etyki Biznesu Akademii Leona Koźmińskiego oraz w Centrum Studiów Kosmicznych Akademii Leona Koźmińskiego – Kozminski ESA Lab. Członek rady programowej Space Entrepreneurship Institute. Ekspert zewnętrzny Fundacji Platforma Przemysłu Przyszłości w obszarze cyfrowych modeli biznesowych i wykorzystania technologii przyszłości. Ekspert i audytor transformacji firm przemysłowych ADMA (advanced manufacturing) w Polsce.
pkurczak@kozminski.edu.pl

Sztuczna inteligencja pomaga

Szybko rozwijająca się sztuczna inteligencja ma nam pomagać w różnych sektorach życia, jednak jej zastosowanie niesie również pewne zagrożenia.

Piotr Kaczmarek-Kurczak

Centrum Studiów Kosmicznych
Akademii Leona Koźmińskiego
– Kozminski ESA Lab w Warszawie

Sztuczna inteligencja (SI) przekształca nasz świat w bezprecedensowy sposób. Od autonomicznych dronów po spersonalizowane zabiegi medyczne – możliwości są nieograniczone. Jednak w miarę jak coraz bardziej polegamy na SI, pojawiają się obawy dotyczące bezpieczeństwa.

Job interview

W styczniu 2017 roku z wyrzutni na poligonie w południowej Arizonie wystartował dron rozpoznawczy armii USA Shadow RQ-7Bv2 w ramach typowych ćwiczeń organizowanych przez armię. Drony tego typu zwykle są używane do prostej obserwacji terenu i korygowania ognia artylerii i mają zasięg do 77 mil, czyli do mniej więcej 120 km od naziemnej stacji kontrolnej. Ten konkretny egzemplarz był jednak zmodyfikowany, miał rozszerzone możliwości samodzielnego działania (autonomiczność). Kiedy został wystrzelony z wyrzutni, cała komunikacja z urządzeniem została przerwana, a ono samo wkrótce rozpoczęło zaskakującą epopeję. Zaraz po starcie zmieniło kurs w stronę Gór Skalistych i zniknęło.

Operatorzy byli przekonani, że dron rozbije się na zboczach gór, ponieważ nie tylko miał ograniczony zasięg, lecz także nie był przeznaczony do operowania na dużych wysokościach. Jednak w jakiś sposób zdołał wznieść się na wysokość 4000 m i przelecieć nad całym łańcuchem górskim, przedostając się na jego drugą stronę i rozbijając się dopiero 600 km dalej – zapewne w momencie, w którym ostatecznie skończyło mu się paliwo. Nie wiadomo, ani dlaczego dron wykonał całą tę operację, ani jak mu się to udało. Podejrzewa się, że prawdopodobnie dużą rolę odegrały prądy powietrzne, które nie tylko wydłużyły zasięg urządzenia, lecz także pomogły mu się wznieść na niezbędną wysokość. Zagadką jest też kierunek lotu. Ponieważ jednostka rozpoznawcza stacjonowała wcześniej w bazie w tej części kraju, podejrzewa się, że w pamięci urządzenia pozostały koordynaty poprzedniej bazy i biedny Shady jak go nazwała prasa – „próbował wrócić do domu".

Nieprzewidywalność

Całe to wydarzenie dobrze ilustruje kilka kluczowych problemów, które tworzy dziś rozwój i zastosowanie sztucznej inteligencji.

Pierwszym z nich jest niepewność. W maszynach prostych, takich jak rower czy huśtawka, łatwo jest zrozumieć zarówno zasady ich działania, jak i do pewnego stopnia przewidzieć ich zachowanie (w codziennym życiu nic nie jest do końca deterministyczne, więc elementu chaosu nigdy się nie da do końca wykluczyć). Jednak maszyny oparte nawet na względnie prostych procedurach działania (np. algorytmach) bardzo szybko generują kolejne stopnie niepewności – każda dodana kolejna reguła powoduje wzrost komplikacji całego systemu i ryzyka, że działanie tego systemu nie będzie zgodne z naszymi oczekiwaniami.

W przypadku programowania rozwiązaniem tego problemu jest rygorystyczne testowanie – sprawdza się dany program tak długo i w takim stopniu, żeby uzyskać wysoki poziom pewności, że w najbardziej typowych sytuacjach program będzie się zachowywał w przewidywalny sposób. Choć wiele algorytmów SI jest testowanych i walidowanych przed wprowadzeniem ich

do użytku, to czasami zdarza się, że błędy pojawiają się dopiero w momencie, gdy system jest już w użyciu. Przykładem może być algorytm rozpoznawania obrazów stosowany w systemie kamer miejskich w Stanach Zjednoczonych. Okazało się, że algorytm ten zawierał pomyłki, przez co system nieprawidłowo identyfikował niektórych ludzi jako podejrzanych i wprowadzał policję w błąd. Na dodatek wiele systemów SI ma charakter eksperymentalny i jest wprowadzanych bez przeprowadzania odpowiednio rozbudowanych testów, uwzględniających również nietypowe sytuacje. Ryzyko z ich zastosowaniem ujawnia się w bardzo rzadkich statystycznie przypadkach, których wystąpienie w długim okresie czasu staje się jednak coraz bardziej prawdopodobne. W konsekwencji może dochodzić do takich niezwykłych zdarzeń jak wyprawa drona Shady, który w szczególnych okolicznościach dokonał wyczynu, uznawanego przez samych jego twórców za niemożliwy z przyczyn technicznych, a co samo w sobie mogło stanowić niespodziewane ryzyko np. dla ruchu lotniczego, infrastruktury krytycznej itd.

> Całe grupy społeczne mogą zostać wykluczone z dostępu do niektórych usług lub zasobów.

Błędy

Drugim problemem braku przejrzystości procesów podejmowanych przez systemy SI są algorytmy. Sztuczna inteligencja jest tylko tak dobra, jakie są jej algorytmy. A te są tworzone przez ludzi, a więc nie są one doskonałe i mogą zawierać błędy. Działanie algorytmu z jednej strony jest wynikiem ograniczeń wynikających z konstrukcji samych maszyn i specyfiki procesu ich programowania (nawarstwianie się poprawek, reedycji, zmiany wprowadzane w ostatniej chwili bez weryfikacji ich wpływu na działanie całości). Oprócz tego każdy algorytm jest w gruncie rzeczy zapisem wiedzy eksperckiej, która posłużyła do jego konstrukcji. Można powiedzieć, że algorytmy są regułami opracowanymi na podstawie obecnej wiedzy na dany temat i zapisanymi w postaci wykonywanego kodu – instrukcji dla maszyny lub człowieka (np. skrypt rozmowy kwalifikacyjnej). Mamy zatem dwa źródła ryzyka.

Jednym jest aktualność i adekwatność wiedzy eksperckiej, z której korzystamy. Czy kryteria oceny, którymi się kierują specjaliści poproszeni przez nas o pomoc, są właściwe w danej sytuacji? Drugim jest poziom zrozumienia tych zasad przez programistów i ich zdolność odzwierciedlenia tej wiedzy w instrukcjach (kodzie). Najlepsza wiedza może zostać nieprawidłowo zaimplementowana, prowadząc do wypaczenia intencji ekspertów, których wiedza została wykorzystana. Efektem może być stronniczość SI. Dochodzi do niej, gdy systemy SI są projektowane lub szkolone w sposób, który prowadzi do niesprawiedliwych lub dyskryminujących wyników. Jest to problem, ponieważ SI jest coraz częściej wykorzystywana do podejmowania decyzji, które mają wpływ na życie ludzi, takich jak decyzje o zatrudnieniu lub ocena zdolności kredytowej.

Jeśli te decyzje są stronnicze, mogą mieć znaczące negatywne konsekwencje dla jednostek lub grup. Całe grupy społeczne mogą zostać wykluczone z dostępu do niektórych usług lub zasobów tylko na podstawie błędnej implementacji jakiejś całkiem rozsądnej reguły. Pośpiech, ograniczenia techniczne, próby zmniejszenia kosztów w pracach na poziomie programistycznym mogą prowadzić do uproszczeń w regułach proponowanych przez ekspertów, a w konsekwencji do pojawienia się bardzo nieprzyjemnych konsekwencji. Przykładem jest działanie jednego z algorytmów rozpoznawania obrazów, który czarnoskóre osoby identyfikował systematycznie jako małpy człekokształtne, ponieważ programiści dla uproszczenia i skrócenia kodu pominęli w regule możliwość posiadania przez ludzi innych odcieni skóry niż biały. W takim przypadku brak przejrzystości i zrozumienia tego, jak algorytm działa, może prowadzić do dyskryminacji i niesprawiedliwości. W tym kontekście odpowiedzialne i etyczne podejście do sztucznej inteligencji wymaga wypracowania standardów transparentności i przetestowania algorytmów pod kątem potencjalnych skutków ubocznych.

Militarne zastosowania

Trzecim problemem są zagrożenia związane z użyciem sztucznej inteligencji w sektorze militarnym, w którym systemy SI są wykorzystywane do wykonywania zadań bojowych. Możliwe jest także zastosowanie sztucznej inteligencji do działań szpiegowskich lub manipulacji informacjami, co może prowadzić do poważnych problemów w dziedzinie bezpieczeństwa państwowego. Jednak w tym obszarze konsekwencje niespodziewanych zdarzeń mogą być dużo poważniejsze niż np. „bunt" inteligentnej pralki. Można sobie zadać pytanie: a co by było, gdyby Shady był zautomatyzowanym bombowcem z bronią nuklearną na pokładzie? Co by było, gdyby taka maszyna postanowiła „wrócić do domu", a potencjalne próby jej zawrócenia uznałaby za atak ze strony przeciwnika? Wojna jest zjawiskiem trudnym do zrozumienia dla maszyn. W przypadku systemów SI, które są w stanie samodzielnie podejmować decyzje, istnieje ryzyko,

że rozwój wojskowych systemów SI wyprzedzi możliwości ich kontroli i doprowadzi do powstania sytuacji niebezpiecznych dla ludzi i środowiska. Dotyczy to przede wszystkim zdarzeń, w których systemy SI nie będą w stanie zrozumieć kontekstu swoich działań i podejmować decyzji na podstawie prostych założeń, zamiast uwzględniać aspekty społeczne, ekonomiczne lub etyczne. W latach II wojny światowej alianci – mimo że szczycili się swoim wyższym poziomem etyki – podjęli decyzje o burzeniu miast, w tym np. Drezna, które nie miało w końcu wojny znaczenia strategicznego. Podjęli również decyzję o użyciu broni nuklearnej przeciwko japońskim miastom, mimo że domyślali się, że decyzja o kapitulacji jest nieuchronna i w zasadzie już zapadła. Innym pytaniem jest kwestia identyfikacji wroga. Jak rozróżnić swoich od obcych? Nawet ludzie mają z tym nieustannie problem – co często prowadzi do przypadków bratobójczej wymiany ognia między jednostkami tych samych sił. Skąd inteligentna broń ma wiedzieć, że wojna się skończyła? Zautomatyzowane, inteligentne miny morskie mogą nie tylko atakować statki, które wydają się im statkami wroga, lecz także unikać wykrycia i zniszczenia przez okręty przeznaczone do ich likwidacji, co może nieść zagrożenie dla ludzi przez wiele dziesiątków lat po zakończeniu konfliktu zbrojnego.

Bezpieczeństwo

Czwartym problemem jest kwestia bezpieczeństwa i ryzyko włamania do systemów SI. Systemy sztucznej inteligencji są podatne na hakowanie tak samo jak każdy inny system komputerowy. Jeśli system SI zostałby zahakowany, mógłby zostać wykorzystany do złych celów. Im bardziej złożony jest dany system, tym trudniej jest go zabezpieczyć przed działaniem z zewnątrz. Luki w algorytmach dają również większą szansę na wprowadzenie danego systemu w błąd. Przykładem jest program do rozpoznawania zmian w płucach. Badacze zmodyfikowali zdjęcia rentgenowskie, wprowadzając do nich zdjęcia głowy goryla. Dla algorytmu była to niepokojąca zmiana we wrażliwym obszarze płuc. Dla ludzi analizujących zdjęcie – podejrzany artefakt sugerujący, że ktoś zaingerował w dane. Co się stanie, jeśli jakiś dowcipniś włamie się do bazy zdjęć rentgenowskich pacjentów i pozmienia je dla żartu? Ilu pacjentów może otrzymać błędne diagnozy? Co się stanie, jeśli przez przypadek lub celowo nieupoważniona osoba wprowadzi do bojowego drona błędne współrzędne celu?

Żeby rozwiązać problemy związane z bezpieczeństwem, naukowcy i decydenci pracują nad opracowaniem środków bezpieczeństwa dla SI. Jednym z podejść jest projektowanie systemów SI z myślą o bezpieczeństwie od samego początku. Oznacza to, że bezpieczeństwo powinno być kluczową kwestią na każdym etapie procesu rozwoju SI, od projektu do wdrożenia.

ALEX YUZHAKOV/SHUTTERSTOCK.COM

Dron z profesjonalną kamerą robi zdjęcia mglistych gór o zachodzie słońca

Innym podejściem jest tworzenie systemów SI, które są przejrzyste i możliwe do wyjaśnienia. Oznacza to, że systemy sztucznej inteligencji powinny być zaprojektowane w sposób umożliwiający ludziom zrozumienie sposobu ich działania i podejmowania decyzji.

> Systemy sztucznej inteligencji są podatne na hakowanie tak samo jak każdy inny system komputerowy.

Może to pomóc w zmniejszeniu ryzyka nieuczciwości i stronniczości SI.

Kolejne podejście polega na uregulowaniu rozwoju i wdrażania systemów SI. Może to obejmować ustanowienie norm bezpieczeństwa i wytycznych, których twórcy SI muszą przestrzegać. Mogłoby to również obejmować ustanowienie organów regulacyjnych, które nadzorowałyby rozwój i wdrażanie systemów SI.

Sztuczna inteligencja ma potencjał, by pozytywnie przekształcić nasz świat. Jednak jak w przypadku każdej potężnej technologii istnieją obawy dotyczące bezpieczeństwa, którymi należy się zająć. Stronniczość, nieuczciwość czy hakowanie SI to tylko kilka z obaw dotyczących bezpieczeństwa, którymi należy się zająć. Żeby zapewnić, że SI jest bezpieczna dla ludzi, naukowcy i decydenci muszą współpracować w celu opracowania środków bezpieczeństwa, które ograniczą te zagrożenia. Tylko w ten sposób możemy uwolnić pełny potencjał SI i zapewnić lepszą przyszłość dla wszystkich. ■

W gromadzeniu danych do tego tekstu pomagała sztuczna inteligencja – głównie nowa wyszukiwarka MS Bing.

Chcesz wiedzieć więcej?

Bostrom N., *Superinteligencja: scenariusze, strategie, zagrożenia*, 2016.

Brockman J. (red.), *Człowiek na rozdrożu. Sztuczna inteligencja – 25 punktów widzenia*, 2020.

Tegmark M., *Życie 3.0. Człowiek w erze sztucznej inteligencji*, 2019.