Przedmowa: Esther Dysor

Juliette Powell, Art Kleiner

DYLEMAT SZTUCZNEJ INTELIGENCJI

7 zasad odpowiedzialnego tworzenia technologii

> Sukces w tworzeniu sztucznej inteligencji może być największym wydarzeniem w historii ludzkości. Niestety, może być również wydarzeniem ostatnim (...).
>
> Stephen Hawking

Helion

Tytuł oryginału: The AI Dilemma: 7 Principles for Responsible Technology

Tłumaczenie: Katarzyna Ellerik
Projekt okładki: Studio Gravite / Olsztyn
Obarek, Pokoński, Pazdrijowski, Zaprucki

ISBN: 978-83-289-0733-1

Drogi Czytelniku!
Jeżeli chcesz ocenić tę książkę, zajrzyj pod adres
https://helion.pl/user/opinie/dyszin
Możesz tam wpisać swoje uwagi, spostrzeżenia, recenzję.

Helion S.A.
ul. Kościuszki 1c, 44-100 Gliwice
tel. 32 230 98 63
e-mail: *helion@helion.pl*
WWW: *https://helion.pl* (księgarnia internetowa, katalog książek)

Printed in Poland.

Spis treści

ROZDZIAŁ

2

Planuj ryzyko w stosunku do ludzi

Użytkownik powinien być chroniony przed niebezpiecznymi lub nieefektywnymi systemami.

— szkic „Konstytucji AI”

Nocą 26 września 1983 r., gdy rozległ się alarm, służbę przy monitorowaniu satelitarnego systemu śledzenia pocisków pełnił 44-letni podpułkownik Armii Radzieckiej Stanisław Pietrow. System wskazywał, że pięć amerykańskich interkontynentalnych rakiet balistycznych zmierza ku Związkowi Radzieckiemu z terytorium Stanów Zjednoczonych. Pietrow miał obowiązek niezwłocznie zgłosić taki sygnał swojemu dowódcy. Niemniej jednak zaufał własnej intuicji i wysłał notatkę, w której uruchomienie alarmu uznane zostało za błąd systemu.

Pietrow uznał komputerową klasyfikację za fałszywą, ponieważ wydawało się prawdopodobne, że pierwsze nuklearne uderzenie Stanów Zjednoczonych obejmowałoby setki pocisków wystrzeliwanych jednocześnie w celu pozbawienia ZSRR możliwości odpowiedzi na atak. Ponadto wiarygodność systemu satelitarnego kwestionowano w przeszłości[1].

Intuicja nie zawiodła Pietrowa. Był to błąd systemu. USA nie wystrzeliły żadnych pocisków. Związek Radziecki nie wziął odwetu, a Pietrow został w końcu bohaterem po obu stronach żelaznej kurtyny[2]. Własnymi rękoma powstrzymał atomowy holokaust, jaki mógł zafundować nam automat.

Jak przestaliśmy się martwić i pokochaliśmy AI

26 września 2022 r. — w 39. rocznicę działań Pietrowa — postanowiliśmy dołączyć jego historię do rękopisu tej książki. Trwała wojna między Rosją a Ukrainą i wiele osób bało się eskalacji konfliktu oraz użycia broni atomowej. Przecież żaden kraj nie wdrożyłby algorytmu w miejsce stanowiska Pietrowa, skoro na szali leżała przyszłość cywilizacji, prawda?

Zdaniem Arta, żaden kraj nie rozważałby na poważnie rezygnacji z ludzkiego nadzoru, uruchamiając maszynę i rezygnując z ludzi takich jak Pietrow. Juliette, ciekawa zdania ekspertów, zapytała fizyków i specjalistów nauk komputerowych — ludzi biegłych w kwestiach związków atomu i sztucznej inteligencji — czy w przyszłości jest to możliwe.

Okazało się, że eksperci również byli zaniepokojeni. Podobnie jak Rada Nauki i Bezpieczeństwa Biuletynu Naukowców Atomistów, grupy współzałożonej przez Alberta Einsteina[3]. Największe wątpliwości, jakimi się z nami podzielono, nie dotyczyły samej technologii, ale możliwych nadużyć sztucznej inteligencji w kontekście broni atomowej.

„Mamy do czynienia z ciągłą presją na zwiększanie automatyzacji [w broni atomowej]", pisze Szabolcs Márka, profesor fizyki na Uniwersytecie Columbia. Przyczyną, zdaniem Márki, jest nieustannie kurczący się czas na podjęcie decyzji. „Systemy automatyczne w sposób nieunikniony muszą korzystać z zaawansowanej sztucznej inteligencji na wszystkich poziomach" — dodaje. „Mam szczerą nadzieję, że ludzkość nie przegapi momentu, w którym mogłaby zostać całkowicie wyłączona z tego procesu. Niemniej jednak na całym świecie funkcjonują potęgi wojskowe wyznające diametralnie różne poglądy. Niektóre z nich mogą uznać, że czas odpowiedzi sztucznej inteligencji jest kluczową zaletą. W zmiennym, chaotycznym i nastawionym na konkurencję świecie zautomatyzowana przez AI broń masowej zagłady jest *absolutnie przerażającą możliwością*". Ujmując to inaczej, fizyk uważa, że użycie sztucznej inteligencji do monitorowania pocisków atomowych i sterowania nimi nie tylko jest prawdopodobne. Tak się po prostu stanie[4].

Możesz, czytając te słowa, pomyśleć sobie, że uczenie maszynowe przedstawia inne zagrożenia niż broń atomowa. Jesteśmy skłonni się z tym zgodzić, ale być może nie tak, jak sądzisz.

Systemy AAA są wszechobecne i głęboko wrosły w tkankę naszej codzienności. Większość skutków ich stosowania jest łagodna, ale tam, gdzie istnieje ryzyko, jest ono znaczne. Być może odczułeś już niektóre z nich na własnej skórze, tak jak ludzie, których historie poznaliśmy podczas pracy nad tą książką. Być może wyciekły Twoje dane, przez co narażasz się na kradzież tożsamości, oszustwa fałszywych „znajomych" czy stalking. Być może trafiają do Ciebie reklamy i treści wideo wykorzystujące Twoje impulsy. Ty lub ktoś, kogo znasz, mógł paść ofiarą państwowego systemu rozpoznawania twarzy i zostać niesłusznie oskarżony o popełnienie przestępstwa. Nie jest to dyskusja akademicka ani nie są to rozważania o przypadkach hipotetycznych. Użycie systemów AAA wiąże się ze szczególnie przykrymi i arbitralnymi konsekwencjami dla podatnych społeczności: dzieci, kobiet i mniejszości.

Analityka predykcyjna odgrywa tu dużą rolę. Kiedy średnie statystyczne wskazują, że ktoś jest prawdopodobnie winny zbrodni, nie dokonał przypisywanych mu osiągnięć lub mu się nie powiedzie, czynnik ludzki podąży za tą sugestią. Nie zdajemy sobie sprawy, jak często się to dzieje, bo systemy oparte na AI i rezultaty ich działania często są ukrywane przed wścibskimi spojrzeniami.

Planowanie ryzyka nie oznacza skupianie się na nim tylko wtedy, gdy jest to wygodne czy zbieżne z kierunkiem naszych prac, a koszty jego ograniczenia są niskie. Oznacza to nieustanny wysiłek wkładany w osiągnięcie dwóch bliźniaczych celów: wdrażanie technologii pozwalające zrealizować w pełni jej potencjał oraz zmniejszanie możliwych szkód wyrządzonych ludziom lub firmom w konsekwencji tego wdrożenia.

Umacnianie swoich szans w środowisku całkowitej niepewności

Rzeczywisty przykład problemu Moral Machine zmaterializował się późną nocą na szerokiej, podmiejskiej drodze w Tempe w Arizonie w 2018 r., gdy pojazd autonomiczny zabił 49-letnią kobietę — Elaine Herzberg. Była to pierwsza śmierć przechodnia związana z technologią samojezdną. Jeden z pokazowych samochodów autonomicznych Ubera, zmodyfikowany

SUV Volvo, jechał z prędkością około 65 km/godz. Herzberg prowadziła rower, by przejść przez jezdnię[5].

Według Rona Dembo, eksperta od zarządzania ryzykiem i przedsiębiorcy, który przeanalizował ten przypadek w swojej książce *Risk Thinking*, „doświadczony człowiek kierowca mógłby pomyśleć (przez milisekundę), że istnieje scenariusz, w którym to coś na poboczu drogi jest człowiekiem chcącym przejść przez jezdnię. Kierowca zabezpieczyłby się, zwalniając w celu uniknięcia kolizji do czasu, aż uzyskałby dodatkowe informacje".

„Gdy pojazd Ubera uzyskał informację, że obiekt na poboczu to kobieta z rowerem — pisze Dembo — było już za późno. Auto nie zwolniło i skoro kobieta podjęła decyzję o wejściu na jezdnię, musiało dojść do zderzenia, a prawdopodobieństwo, że będzie ono miało skutek śmiertelny, było prawie stuprocentowe. Być może ani pojazd autonomiczny ani jego operator nie spotkali się nigdy z taką sytuacją. Niemniej jednak, jeśli auta samojezdne mają poruszać się po naszych drogach, muszą być w stanie zarządzać tego typu ryzykiem. Muszą umieć zbudować strategię i podjąć decyzję nawet na podstawie niepełnych danych"[6].

Zdaniem Dembo mieliśmy tu do czynienia z nieadekwatnym namysłem nad ryzykiem na etapie projektowania oprogramowania. Przez „namysł nad ryzykiem" rozumie on umiejętność brania pod uwagę tego, co niepewne, i wybierania elastycznej ścieżki pozwalającej na manewry w miarę, jak rozwija się sytuacja. Zachowanie Stanisława Pietrowa, gdy powstrzymał on wojnę atomową, zauważając błąd systemu, było przykładem namysłu nad ryzykiem. Program AI w pojeździe Ubera nie miał namysłu nad ryzykiem w swoim repertuarze. Auto nie było w stanie określić rodzaju ruchu, jaki wyczuły jego sensory, nie dysponowało zaprogramowanym sposobem poradzenia sobie z niepewnością, więc kontynuowało jazdę z tą samą prędkością, aż było za późno, by uniknąć zderzenia. W pewnym sensie także Uberowi jako firmie zabrakło namysłu nad ryzykiem. Ten wypadek zmusił firmę do zawieszenia testów swoich pojazdów autonomicznych. Ostatecznie jednostka aut autonomicznych została sprzedana, choć kiedyś stanowiła kluczowy element strategii biznesowej[7].

Trzeba przy tym zauważyć, że w czasie wypadku w samochodzie znajdowała się operatorka. Jej zadaniem nie było prowadzenie auta, ale

obserwowanie i opisywanie działań samochodu. Na kilka sekund przed wypadkiem kobieta nie patrzyła na drogę, a auto nie ostrzegło jej wcześniej niż sekundę przed zderzeniem, kiedy to chwyciła za kierownicę.

Jest to przykład dynamiki często występującej w systemach zautomatyzowanych, a pierwotnie odnotowany w czasach, kiedy szkolono pierwszych pilotów. Zjawisko to określa się czasem mianem *zachwytu nad automatyzacją* — im bardziej autonomiczna jest maszyna, tym bardziej ludzie jej ufają i zaniedbują uwagę. Gdy natomiast przestajemy przykładać wagę do działania systemu, wkraczamy w niepewną dolinę automatyzacji, pętlę negatywnego wzmocnienia. W przypadku pilotów korzystanie z autopilota pociągało za sobą erozję umiejętności lotniczych. Uogólniając, im większa automatyzacja, tym mniejsza uwaga ludzi. A im mniej uwagi, tym więcej wypadków. Z kolei im więcej wypadków, tym większe zapotrzebowanie na automatyzację. Negatywna spirala się nakręca[8].

W kategoriach Dembo ludzie oddelegowują namysł nad ryzykiem na maszyny, ale maszyny nie są w stanie sprostać temu zadaniu. Wydaje się, że właśnie tak stało się w Tempe[9]. Zachwyt nad automatyzacją sprawia problemy wszędzie, gdzie ludzie sprawują nadzór nad technologią zautomatyzowaną. Całymi dniami wyczekujemy i wyglądamy sygnału, że dzieje się coś mało prawdopodobnego, ale możliwego. Popadamy w znudzenie, poczucie fałszywego bezpieczeństwa i ten sygnał nam umyka. Syndrom ten dotyka strażników i ochroniarzy. Tak samo jako nadzorców reaktorów atomowych, ruchu drogowego i wielu systemów AAA.

Gdy rozmawialiśmy z Dembo w czasie jego rodzinnej wycieczki na Kostarykę w 2022 r., dodawał on, że sztuczna inteligencja jako całość poprawia swoje możliwości w zakresie namysłu nad ryzykiem. Mimo to ma jeszcze do przebycia daleką drogę. Dembo ma teorię na temat czasów całkowitej niepewności, kiedy szaleje efekt motyla[10], a małe zdarzenia prowadzą często do ogromnych kryzysów. Uważa on, że w tej sytuacji systemy AI znajdują się w trudniejszej pozycji wyjściowej niż ludzie. „Dla ludzi namysł nad ryzykiem jest naturalny, dokonujemy uogólnień na podstawie niewielkiej próbki danych, zawieramy kompromisy, odwracamy kota ogonem i żałujemy swoich decyzji, w miarę jak idziemy przez życie. Każdy złożony wybór, jakiego dokonujemy, jest sposobem radzenia sobie z jakimś

przyszłym niepewnym zdarzeniem, choćby nieświadomym bilansem zysków i strat”[11].

Przykładem namysłu nad ryzykiem w podejmowaniu decyzji politycznych i biznesowych jest planowanie scenariuszy. Jest to podejście w zarządzaniu, które zakłada rozważenie wielu przyszłych możliwości. Podejmując się tego ćwiczenia, Ty i Twój zespół zwykle zauważacie problemy z wyprzedzeniem. Nie wiecie, czym one są. Algorytm predykcyjny wybrałby tylko jedno rozwiązanie i działał zgodnie z nim. Ty natomiast wymyślasz kilka różnych scenariuszy. W przypadku pojazdów autonomicznych być może tym, co rusza się w oddali, jest kobieta z rowerem chcąca przejść przez jezdnię, a Ty pędzisz prosto na nią. A może to jeleń. Może cień lub odbicie promieni światła od szyby samochodu. Nie starasz się przewidzieć przyszłości, tylko wciskasz pedał hamulca. I to nie dlatego, że takie jest rozwiązanie problemu, tylko dlatego, że to opcja pokrywająca największą liczbę możliwości. Ponieważ nie wiesz dokładnie, co się stanie, wybierasz strategiczne działanie, które zapewni Ci co najmniej przyzwoity rezultat w największej liczbie przypadków.

Może Ci się wydawać, że automatyczna natura sztucznej inteligencji predysponuje ją do skutecznego zarządzania ryzykiem, przynajmniej we właściwych przypadkach użycia. Niemniej jednak systemy nie potrafią przewidzieć przyszłości, ponieważ wytrenowano je na danych historycznych. Gdy tylko pojawią się nowe zmienne czy warunki, jak w przypadku auta Ubera, które zabiło Elaine Herzberg, zdolność przewidywania systemu maleje. System nie został wytrenowany na tych danych. W zależności od przypadku użycia może to decydować o czyimś życiu lub śmierci.

Jako osoby rozważające ryzyko chcemy maksymalizować nasze szanse, tak by w przypadku nieoczekiwanego zdarzenia móc zmienić front i dopasować się do sytuacji. Wypadek Ubera pokazuje, co się stanie, gdy nie będziemy rozmyślnie brać pod uwagę ryzyka. Gdy jednak będziemy to robić, jak okaże się dzięki przedsiębiorczyni Helen Greiner, ta celowość może nam pozwolić myśleć efektywnie o nawet najpoważniejszych ryzykach. Pomoże nam zastanowić się nad długofalowymi skutkami działań i zabezpieczyć się przed negatywnymi, nieplanowanymi konsekwencjami. W ten sposób zmniejsza się i ogranicza ryzyko związane z systemami AAA.

Roboty, drony i celowość ryzyka

Kiedy Juliette po raz pierwszy spotkała się na przedmieściach Bostonu z przedsiębiorcą z branży robotyki, Helen Greiner, kobieta ta powiedziała jej, że jej wyborami nadal kieruje cel z dzieciństwa. Greiner zarabia na życie, produkując roboty, które ludzie chcą kupować, a dziewczynki — zhakować.

„Wszystko zaczęło się, gdy miałam 11 lat" — wspomina Greiner. „Obejrzałam wtedy *Gwiezdne wojny* i zakochałam się w R2-D2. On nie był tylko maszyną — miał misję. Przecież uratował świat! Posiadał osobowość i miał sporo do powiedzenia poprzez swoje piski i brzęczenia. Od tamtej pory zawsze chciałam robić rzeczy, które będą czymś więcej niż maszyną. Poszłam na MIT, żeby się tego nauczyć. Zobaczyłam tam mnóstwo świetnej roboty inżynieryjnej, ale nadal nie wiedziałam, jak stworzyć roboty funkcjonujące w prawdziwym świecie"[12].

> **„Obejrzałam wtedy *Gwiezdne wojny* i zakochałam się w R2-D2. On nie był tylko maszyną — miał misję. Przecież uratował świat! Od tamtej pory zawsze chciałam robić rzeczy, które będą czymś więcej niż maszyną".**
>
> **— Helen Greiner**

Greiner była niesamowicie bystrym dzieckiem. Niemal nie mówiła do okresu dojrzewania, ale to dlatego, że nie interesowała ją większość rozmów toczonych dookoła niej. Jako nastolatka zainteresowała się robotyką i odtąd wiele się w jej życiu zmieniło. W 1990 r. wraz z dwoma absolwentami laboratorium AI na MIT założyła firmę iRobot, która zajmowała się produkowanymi na specjalne zamówienie robotami przemysłowymi i wojskowymi, prototypami maszyn do podróży w kosmosie i pierwszymi zabawkami robotycznymi. Ryzyko wiążące się z tymi urządzeniami było niemal niedostrzegalne. Wiele z nich zaprojektowano właśnie po to, by zmniejszyć ryzyko związane z czynnikiem ludzkim w niebezpiecznych okolicznościach, np. na podwodnych platformach wiertniczych.

W tym samym czasie założyciele iRobot z zaangażowaniem zgłębiali tajniki sukcesu komercyjnego, tym bardziej że w ramach swojej działalności stworzyli jeden z niewielu małych start-upów produkujących zarówno oprogramowanie, jaki i sprzęt, i to przy bardzo niewielkim kapitale początkowym. Później pojawiło się finansowanie w ramach programu DARPA (Defense Advanced Research Projects Agency — Agencja ds. Specjalnych Projektów Badawczych nad Obronnością) oraz zewnętrzny kapitał inwestycyjny. Greiner i jej zespół wyprodukowali dwa roboty. Roomba był pierwszym autonomicznym, domowym robotem odkurzającym. W ciągu pierwszych czterech lat sprzedał się w liczbie miliona egzemplarzy. PackBot był z kolei odpowiednikiem Roomby w rozminowywaniu bomb. Przypisuje mu się rozbrojenie tysięcy bomb domowej produkcji w Iraku i Afganistanie. Bez niego prawdopodobnie wielu żołnierzy i cywilów straciłoby życie.

„Pewnego razu przemawiałam w koledżu wojskowym" — opowiada Greiner. „Po mojej prawej i lewej stronie siedzieli generałowie z mnóstwem gwiazdek na pagonach. Zakładałam, że po konferencji wszyscy będą chcieli rozmawiać właśnie z nimi. Tymczasem wielu żołnierzy podchodziło do mnie. Pamiętam jednego, który uścisnął mi dłoń i powiedział: »PackBot ocalił jedenastu moich ludzi podczas misji«"[13].

Greiner była dyrektorem iRobot do 2008 r., kiedy to odeszła z firmy, by założyć nowy start-up: firmę produkującą drony CyPhy. Przyświecało jej założenie, że ludzie i roboty mogą współdzielić przestrzeń powietrzną co najmniej tak łatwo i efektywnie jak ziemię. W 2017 r. odeszła z CyPhy, by zacząć działalność na rzecz amerykańskiej armii, gdzie doradzała w kwestiach związanych z robotyką. W 2020 r. dołączyła do Tertill, start-upu założonego przez jednego z inżynierów z iRobot. Firma ta produkuje roboty ogrodowe zapobiegające chwastom i użyźniające ziemię[14].

Greiner to ten typ charyzmatycznej kobiety, którą zaprasza się do Białego Domu i na Światowe Forum Ekonomiczne. Mówi od serca i opowiada się za ciągłym używaniem innowacyjnej technologii AI, również w wojskowości. Za jej przywiązaniem do służby stoją determinacja i lata doświadczenia w pracy nad ograniczaniem ryzyka związanego z czynnikiem ludzkim. Wybrzmiewa w nim także świadomość potencjalnych

korzyści i zagrożeń nieodłącznie związanych z robotami i dronami napędzanymi przez AI.

W rozmowie opublikowanej przez Foreign Affairs Greiner odnosi się do tej kwestii bezpośrednio:

> „Terrorysta może dziś kupić drona i zacząć planować atak. Uważam, że jedynym sposobem, żebyśmy wychwycili takie przypadki, jest zastosowanie ludzkiej inteligencji. Terroryści z powodów komercyjnych nie kupią dronów od firmy je budującej. Jeśli będą chcieli wypełnić drona materiałami wybuchowymi, to pójdą do sklepu dla pasjonatów i nabędą model dostępny od ręki. Sądzę, że to wyzwanie. Ale to samo można zrobić z samochodem, a jednak nie mówimy: »Nie należy sprzedawać aut, bo mogą posłużyć do przeprowadzenia samobójczego ataku«. Musimy po prostu dowiedzieć się, kto ma takie plany, i powstrzymać go”[15].

Dla Greiner celowość w myśleniu o ryzyku jest kluczowa. Wydaje się także ważna dla wielu osób, które obserwują ją i jej karierę — nieustająco otrzymuje ona pytania o ryzyko wpisane w jej pracę.

Gdy ryzyko jest nie do zaakceptowania

Kilka lat temu specjalistka data science i artystka Lynn Cherny pracowała nad systemem AI do zarządzania społecznością. Jego celem było zmniejszenie ryzyka. Pracodawcą był europejski start-up oferujący nietypową usługą z zakresu bezpieczeństwa. Używał on algorytmu czyszczącego do przeszukiwania mediów społecznościowych i czatów swoich klientów w poszukiwaniu wiadomości obraźliwych lub nękających. Algorytm ten prezentował statystyki dotyczące niepokojących treści na prywatnym pulpicie tak, by klienci mogli usuwać tego typu wiadomości i blokować pojawienie się nowych. Cherny i dwóch innych specjalistów było odpowiedzialnych za zbudowanie systemu oznaczającego treści wymagające przejrzenia przez człowieka. „Nasza uwaga skupiała się — wyjaśnia — w szczególności na kwestiach związanych z przemocą wobec dzieci”[16].

Sama tylko ekspozycja na tego typu treści miała okropny wpływ na Cherny i jej zespół. W rozmowie przeprowadzonej z domu Cherny we

Francji opowiadała ona, że strony niektórych klientów pełne były przemocy, trollowania i mowy nienawiści. „Napotykaliśmy wiele dobrowolnych rozmów o charakterze erotycznym. Ale widzieliśmy też mnóstwo niechcianych wiadomości o treściach obscenicznych czy przypadków stalkingu. Oglądaliśmy wiadomości wysyłane do nieletnich przez ludzi o agresywnych zamiarach, którzy próbowali się zaprzyjaźnić z nimi czy prosili o nagie zdjęcia. Istniały także wpisy o wymowie samobójczej i kilka gróźb przemocy w świecie rzeczywistym”[17].

Najbardziej niepokojącym aspektem tej pracy była jednak reakcja przełożonych na sugestię Cherny, że pulpit mógłby sam wyszukiwać i filtrować wiadomości na przykład mające na celu nagabywanie dzieci. Przełożeni nie chcieli nawet przedyskutować tego pomysłu z klientami. „Tłumaczyli, że gdyby posłuchali mojej rady, ściągnęłoby to na firmę odpowiedzialność prawną”. Jakieś wiadomości o przestępczym charakterze z pewnością przedostałyby się przez filtry, a gdyby taka informacja wyszła na jaw, firma wolałaby zakomunikować, że nie wiedziała o problematycznych obszarach. „Szefowie nie chcieli wiedzieć o niczym niewygodnym”.

To oznaczało, że złoczyńcy i ich wiadomości zostaną. Dobrą intencję pogrzebano.

Wkrótce potem Cherny i jeden z jej kolegów odeszli ze start-upu, powołując się na szereg problemów, w tym na brak sensu ich pracy. „Zadaliśmy sobie pytanie: »Czy warto w ogóle przyglądać się tym treściom, skoro nie ma to znaczenia dla klientów?«”. Niezależnie od swoich kompetencji technicznych i poczucia kontroli związanego z ich pracą zespół Cherny nie miał siły przebicia w organizacji, by poprowadzić sprawę dalej. A skoro ludzie-moderatorzy zatrudnieni do ograniczania krzywd nie mają nad tym kontroli, to kto ją ma?

Podobne historie słyszeliśmy od wielu osób z wielkich firm technologicznych — w różnych okolicznościach, ale zasadniczo konkluzja była ta sama. Casey Cerretani twierdzi, że osoby techniczne na stanowiskach takich jak jego regularnie widzą dowody handlu seksem — np. na serwerach używanych przez klientów. Może to budzić sprzeciw inżynierów, ale co do zasady podnoszenie takich kwestii nie jest ich zadaniem i za każdą choćby próbę takiego zachowania mogą spotkać ich negatywne

konsekwencje. Taka sytuacja sprawia, że inżynierowie oprogramowania są sceptyczni w dążeniu do zmiany wypływającej z wnętrza organizacji. Zamiast tego chcą raczej na tyle dużo zarobić w wielkich korporacjach technologicznych, by móc zainwestować w start-upy odpowiadające wyznawanym przez nich wartościom[18].

Dzięki filmom dokumentalnym takim jak *The Social Dilemma* i wydarzeniom takim jak zeznania byłej specjalistki z zakresu nauki o danych Facebooka Frances Haugen złożonym przed Kongresem w październiku 2021 r. coraz więcej osób zyskuje świadomość wysokiego ryzyka związanego używanymi na co dzień systemami algorytmicznymi

Dzięki filmom dokumentalnym takim jak *The Social Dilemma* i wydarzeniom takim jak zeznania byłej specjalistki z zakresu nauki o danych Facebooka Frances Haugen złożonym przed Kongresem w październiku 2021 r. coraz więcej osób zyskuje świadomość wysokiego ryzyka związanego z używanymi na co dzień systemami algorytmicznymi[19]. Rankingi oparte na zaangażowaniu stanowią na przykład zasadniczą część mediów społecznościowych takich jak Meta (wcześniej Facebook) czy TikTok. Strony te wyświetlają użytkownikom treści na podstawie analizy ich uprzedniego zachowania w sieci. Analizy tej dokonuje oczywiście system AI. Olbrzymią rolę odgrywa tutaj iluzja kontroli. Nieustający napływ treści tworzonych przez użytkowników, dopasowany do Twoich przyzwyczajeń, by zdobyć Twoją uwagę, daje Ci poczucie, że kierujesz tym systemem. A tak naprawdę to on kieruje Tobą. Jego najprostsze wybory skłaniają Cię do pozostania na stronie, obejrzenia kolejnych reklam i to w kontekście, który sprawia, że jesteś przychylny reklamowanym produktom niezależnie od tego, jak szkodliwe mogą one być[20].

Haugen zeznała, że aplikacja Mety Instagram wywoływała w nastolatkach poczucie nienawiści do samych siebie i doprowadzała do samookaleczeń. Dochodziło też do innych nadużyć, np. subtelnego nakłaniania

do mowy nienawiści w stosunku do grup etnicznych. Haugen potwierdziła swoje zeznania wewnętrznymi dokumentami, z których jasno wynikało, że kierownictwo firmy wiedziało o tej sytuacji i akceptowało ją. Pewne wewnętrzne badanie w Mecie wykazało, że 13,5 procent nastolatek przyznało, że Instagram wpływa na pogłębienie ich myśli samobójczych, a 17 procent potwierdziło to samo w stosunku do zaburzeń żywienia[21].

Później na antenie programu *60 minutes* stacji CBS Haugen wyjaśniła, że jej były pracodawca przez lata przedkładał zysk nad bezpieczeństwo. „Facebook zdawał sobie sprawę, że jeśli zmieni algorytm na bezpieczniejszy, ludzie będą pozostawać na stronie krócej. Klikną w konsekwencji mniej reklam, więc firma zarobi mniej pieniędzy”[22]. Haugen ustawicznie wskazuje na konieczność odgórnej legislacji, by Meta poprawiła własną platformę. Sama z siebie nie przywoła się do porządku[23].

Model namysłu nad ryzykiem

W tej chwili dochodzimy ze wszystkich stron do konsensusu, że ogólne regulacje w związku ze sztuczną inteligencją są niezbędne. Nie jest jednak nadal jasne, jak daleko będą szły te regulacje i jak szeroki będzie ich efekt. Istnieją manifesty takie jak „Konstytucja AI” Białego Domu czy lokalne ustawodawstwo, np. AI Bias Law (Ustawa przeciw wypaczeniom AI) w Nowym Jorku, które weszło w życie na początku 2023 r. Prawo to zabrania wykorzystania systemów sztucznej inteligencji do podejmowania decyzji o zatrudnieniu, o ile nie przeszły one inspekcji pod kątem uprzedzeń płciowych i rasowych[24].

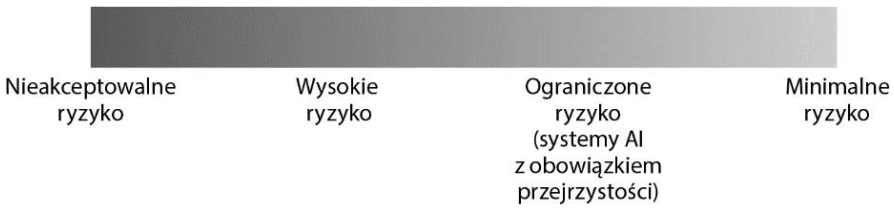
„W miarę jak algorytmy i inne systemy podejmowania automatycznych decyzji odgrywają coraz ważniejszą rolę w naszym życiu, naszą powinnością staje się zapewnienie, że są one właściwie oceniane pod kątem błędów uderzających w społeczności mniejszościowe lub marginalizowane”.

— amerykański senator Corey Booker

„W miarę jak algorytmy i inne systemy podejmowania automatycznych decyzji odgrywają coraz ważniejszą rolę naszym życiu, naszą powinnością staje się zapewnienie, że są one właściwie oceniane pod kątem błędów uderzających w społeczności mniejszościowe lub marginalizowane" — mówił amerykański senator Corey Booker, przedstawiając propozycję ustawy Algorithmic Accountability Act w 2022 r.[25]

Dotychczas najpełniejsze podejście pod tym względem stanowi akt prawny Unii Europejskiej — Artificial Intelligence Act (AIA). Zawarte w nim propozycje nagłośniły dyskusję o ryzyku i kompromisach, a odpowiedzi w związku z nimi udzielają przedstawiciele wszystkich czterech logik władzy: inżynierii, sprawiedliwości społecznej, korporacji i rządu. Ustawa jest na tyle szeroka i pojawiła się na tyle wcześnie, że prawdopodobnie wyznaczy standard regulacji sztucznej inteligencji na całym świecie — o ile zostanie ratyfikowana[26].

Jednym z kluczowych aspektów proponowanych regulacji jest piramida krytyczności[27]. Jest to model przypisywania działań wszystkich systemów AAA — rozwiązań autonomicznych, analitycznych i sztucznej inteligencji — do czterech kategorii ryzyka widocznych na rysunku 2[28].



Rysunek 2. Model namysłu nad ryzykiem; źródło: Kleiner Powell International (KPI)

1. Aplikacje minimalnego ryzyka nie wymagają ludzkiego nadzoru, bo nikogo nie krzywdzą. Zgodnie z modelem AIA tej kategorii odpowiadają skierowane do dzieci gry wideo używające AI oraz filtry spamu.
2. Aplikacje ograniczonego ryzyka mają obowiązek transparentności. Przykładowo, korzystanie z chatbotów jest możliwe tylko wtedy, gdy informuje się użytkownika o maszynowym charakterze interakcji.

3. Aplikacje wysokiego ryzyka obejmują systemy, które mogłyby spowodować uszczerbek na zdrowiu człowieka, zagrozić bezpieczeństwu czy podstawowym prawom człowieka, ale zarazem przedstawiają też istotną wartość dla ludzi i społeczeństwa. W tej kategorii mieszczą się auta samojezdne, systemy sprawdzania zdolności kredytowej i aplikacje chroniące dzieci, zaprojektowane, by alarmować, gdy dziecko jest ofiarą przemocy rodzinnej. Aplikacje te używane z rozwagą są cenne, w innym przypadku mogą prowadzić do nadużyć. W obecnej wersji europejskiej ustawy zakłada się przeprowadzanie audytu, zwykle przez zewnętrzne firmy audytowe lub organizacje non profit chroniące prawa konsumenta. Aby umożliwić wyższe poziomy innowacji, UE mogłaby ustanowić „regulacyjne środowisko testowe", w którym zachęcałoby się do eksperymentowania z kwalifikującymi się projektami bez typowych audytów.
4. Aplikacje nieakceptowalnego ryzyka byłyby zabronione. Do tej kategorii należałyby systemy identyfikacji biometrycznej czasu rzeczywistego, łącznie z wieloma ich zastosowaniami w służbach porządkowych, podprogowe techniki mające na celu wypaczenie zachowań użytkownika, aplikacje żerujące na podatności różnych grup, np. zabawki prowadzące dzieci do niebezpiecznych zachowań oraz systemy sztucznej inteligencji służące do relatywizacji społecznej — preferujące pewnych ludzi z możliwościami i wypychające innych na margines[29].

Jeśli nie jest dla Ciebie jasne, dlaczego Unia Europejska planuje mocno ograniczyć lub wręcz zdelegalizować pewne zastosowania AI, przemyśl poniższy komentarz Ryana Carriera, założyciela organizacji non profit ForHumanity, która wspiera i koordynuje niezależny audyt systemów AI: „Niektóre ryzyka brzmią jak scenariusze filmów science fiction, ale wszystkie one są prawdopodobne w ciągu najbliższych kilku lat. Posiadając moje DNA, mógłbyś mnie sklonować. Lub zaprojektować szczególną broń, która mogłaby zabić tylko mnie. Mógłbyś wykorzystywać mnie na podstawie mojego profilu psychologicznego lub fizycznego, mojego słownictwa czy reakcji emocjonalnych"[30].

Model ten odpowiada logice rządu i być może w pewnym stopniu pozostali się z nim zgadzają. Wydaje się, że rośnie przekonanie, iż ludzie tracą kontrolę nad technologią i tylko prawo pozwoli im ją odzyskać. Jednak wspomniane kryteria są na tyle mgliste, że dopuszczają arbitralne postanowienia. „Tych list nie uzasadniają zewnętrznie weryfikowalne kryteria" — pisze znawca prawa Lilian Edwards, profesor prawa, innowacji i społeczeństwa na uniwersytecie w Newcastle. „Jeśli nie jest do końca wiadome, dlaczego niektóre systemy teraz znajdują się na liście aplikacji o wysokim ryzyku, niemal niemożliwe będzie uzasadnienie dodania tam nowych systemów w przyszłości"[31].

Dyskusje związane z ustawą AIA przenoszą się także na grunt cyfrowej suwerenności: jak dużą kontrolę powinno utrzymywać państwo nad systemami AAA używanymi na jego terytorium lub przez jego obywateli. Obecne zapisy głoszą, że każda firma chcąca prowadzić interesy w Unii Europejskiej będzie musiała porzucić niekompatybilne praktyki także na innych rynkach, łącznie z praktykami swoich podwykonawców. Przykładowo, firma będzie musiała przestać sprzedawać swoje rozwiązania z zakresu analityki monitorowania i śledzenia reżimom autorytarnym — zwłaszcza jeśli te reżimy wykorzystywały technologię do marginalizowania własnych grup etnicznych lub wpływania na wyniki wyborów w innych krajach. Ustawa AIA uzasadnia takie zapisy stwierdzeniem, że międzynarodowe procesy uczenia maszynowego mogą obejmować dane wyjściowe z europejskich systemów algorytmicznych[32].

Nowe regulacje w Unii Europejskiej, nawet jedynie w formie propozycji, pokazują, jak ważne w zapobieganiu ryzykom będzie zaufanie.

Nowe regulacje w Unii Europejskiej, nawet jedynie w formie propozycji, pokazują, jak ważną rolę w zapobieganiu ryzykom będzie odgrywało zaufanie. Zgodnie z ich zapisami to na firmach będzie spoczywał ciężar udowodnienia, że ich aplikacje są niegroźne — w zamierzeniach i w realnych skutkach działania. W swojej bieżącej formie ustawa oznacza,

że firmy będą musiały publicznie nazwać cele swoich aplikacji, by zapewnić, że są one godne zaufania. O ile nie będzie istniał klarowny sposób zbadania wewnętrznych mechanizmów działania dowolnego systemu AAA, instytucje Unii Europejskiej będą mogły ograniczyć jego użycie, zabronić go lub nałożyć wysokie grzywny. Możliwość wyjaśnienia sztucznej inteligencji to jednak odrębny problem. Pochylamy się nad nim w kolejnym rozdziale pt. „Rzucaj światło".

Sztuczna inteligencja to technologia o nadzwyczajnym potencjale, jednak jej nieodpowiedzialne wykorzystanie doprowadzało już do odmowy świadczenia usług medycznych, czy nawet bezprawnych aresztowań.

Dlatego zanim pozwolimy maszynom podejmować decyzje etyczne, powinniśmy przeprowadzić globalną dyskusję. Ludzie mają różne priorytety i wartości, jak zatem możemy oczekiwać od technologii, by kierowała się właściwymi przesłankami - nie tylko w sprawach życia i śmierci, ale i na każdym kroku?

Ta książka łączy perspektywy inżynierów, ludzi biznesu, przedstawicieli rządów i społeczników. Ułatwia zrozumienie korzyści i szans, jakie niosą ze sobą autonomiczne systemy oparte na uczeniu maszynowym. Zawiera siedem ważnych zasad, które pozwolą ograniczyć ryzyko nadużyć i wypadków związanych z AI, a także zapewnią, by technologie tego rodzaju służyły rozwojowi ludzkości. Cztery z zasad dotyczą samych systemów i ich projektowania: uwzględniania ryzyka dla ludzi, przejrzystości działania, zapewnienia ochrony danych osobowych i ograniczania tendencyjności. Pozostałe trzy odnoszą się do organizacji tworzących systemy AI, stosowanych w nich procedur i kultury organizacyjnej. Co ważne, w książce znalazły się przykłady dobrych praktyk, jak również liczne rzeczowe uwagi i pożyteczne wskazówki.

Wnikliwa lektura pomoże nam uświadomić sobie zagrożenia i bezprecedensowe korzyści, jakie mogą zapewnić systemy sztucznej inteligencji.

Juliette Powell zajmuje się wzajemnym przenikaniem się kultury, nauki o danych i etyki. Jest założycielką i dyrektorką generalną firmy Turing AI i wykładowczynią na Uniwersytecie Columbia.

Art Kleiner specjalizuje się w zarządzaniu strategicznym. Jest redaktorem naczelnym magazynu „Strategy+Business", a także współzałożycielem i dyrektorem generalnym firmy Kleiner & Company.

Helion

helion.pl

HELION SA
ul. Kościuszki 1c
44-100 Gliwice
tel.: 32 230 98 63
helion@helion.pl

KOD KORZYŚCI
Sięgnij po więcej! ▶

ISBN 978-83-289-0733-1

9 788328 907331

Cena: 54,90 zł