POLSKIE TOWARZYSTWO INFORMATYCZNE

NUMER 1/2026

PRZYSZŁOŚĆ ZACZYNA SIĘ DZISIAJ

Domena

ISSN 2720-6513

Jak sztuczna inteligencja uczyła się mówić

O CZYM ŚNIĄ MIKROPROCESORY

Michał Kuszewski
O etycznych ramach dla generatywnej AI w administracji publicznej

Prymat algorytmów

WSZYSCY CHCĄ DOBRZE, A WYCHODZI JAK ZWYKLE

Spis treści

Temat numeru

4 Jak sztuczna inteligencja uczyła się mówić – *Jacek Grabowski*

Informatyka i bezpieczeństwo

8 Prymat algorytmów – *Joanna Karczewska*
13 CAPTCHA kapituluje przed AI – *Sebastian Tyralski*
18 Dzień świstaka, czyli jak nie wpaść w pętlę czasu – *Paweł Henig*

Informatyka i wydarzenia

20 Przepis na sukces naukowy – *Hanna Mazur*
24 Cyfrowa husaria czy podwykonawcy gigantów? – *Grzegorz Gwardys*
29 Lista 100 – *Włodzimierz Marciński*

Informatyka i antroposfera

31 O czym śnią mikroprocesory – *Piotr Kościelniak*
36 Jak Warszawa tworzyła etyczne ramy dla generatywnej AI w administracji publicznej – *Michał Kuszewski*
39 Wszyscy chcą dobrze, a wychodzi jak zwykle – *Wojciech Bednaruk*

Informatyka i kompetencje

43 Subiektywny poradnik administratora cz. V – *Adam Jurkiewicz*

47 **Na marginesie…** – *Wiesław Paluszyński*

49 **Z ukosa** – *Michał Ogórek*

Domena

nr 1/2026

Wydawca:
Polskie Towarzystwo Informatyczne

Zarząd Główny:
ul. Solec 38 lok.103
00-394 Warszawa
NIP: 522-000-20-38
tel.: +49 22 838 47 05
e-mail: pti@pti.org.pl

Redaktor naczelna:
Anna Kniaź
(anna.kniaz@pti.org.pl)

Rada Programowa „Domeny”:
Wiesław Paluszyński
– przewodniczący Rady
Marek Bolanowski
Marian Bubak
Beata Chodacka
Bogusław Dębski
Wojciech Kiedrowski

Współpraca redakcyjna:
Tomasz Kulisiewicz

Korekta:
Jolanta Jamiołkowska

Skład i opracowanie graficzne:
Agencja HEADOUT

Headout

Wszystkie teksty udostępniamy na licencji Creative Commons
Uznanie autorstwa-Użycie niekomercyjne -Na tych samych warunkach 4.0

Szanowni Państwo,

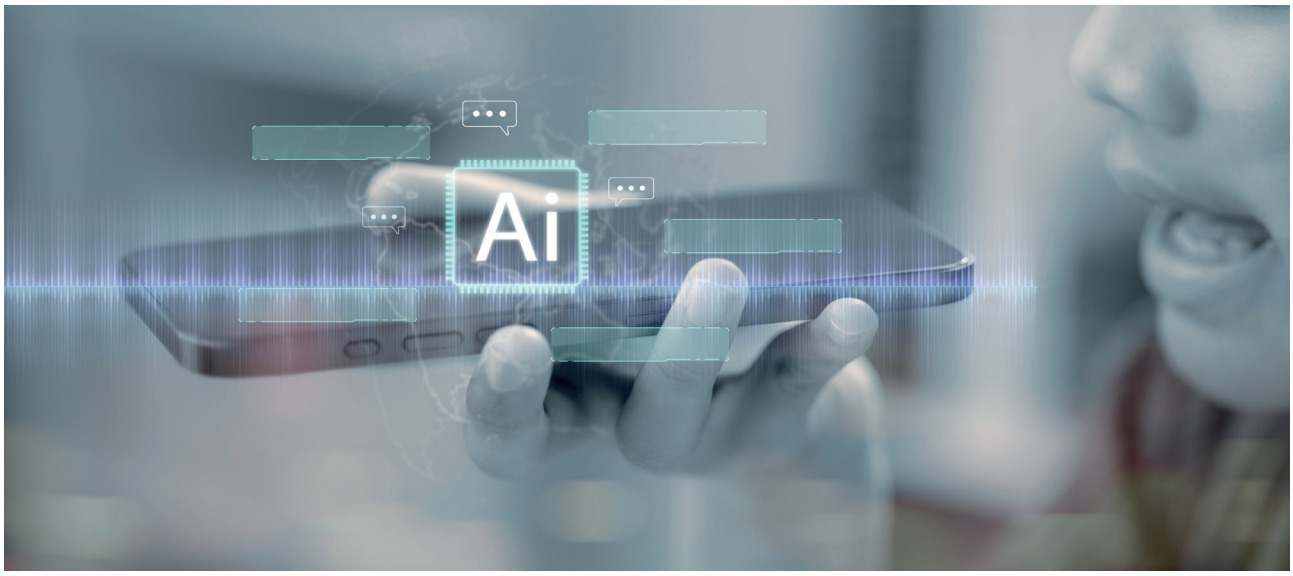
dramaturgia rozwoju sztucznej inteligencji nie ustępuje filmom Hitchcocka. Dożyliśmy pierwszego w historii procesu sądowego, w którym prywatna firma AI pozwała rząd amerykański o prawo do odmowy świadczenia usług. Mowa o firmie Anthropic, której duży model sztucznej inteligencji Claude (nazwa na cześć Claude'a Shannona) wygrywa z rozwiązaniami konkurencji. Przewagę daje też formuła działania firmy, zarejestrowanej jako *Public Benefit Corporation* i mającej statutowy obowiązek priorytetowo traktować bezpieczeństwo AI. Z tego względu Anthropic, dopuszczona przez amerykański rządu do dostępu do systemów niejawnych, była jednym z najważniejszych partnerów Pentagonu. Do czasu, gdy rząd amerykański w swoim wojennym wzmożeniu postanowił zażądać od wszystkich dostawców technologii AI usunięcia ograniczeń użytkowania. Anthropic nie zgodził się na masową inwigilację Amerykanów i obstawał przy zakazie używania autonomicznej broni bez ludzkiego nadzoru. CEO firmy Dario Amodei argumentował, że Claude po prostu nie był trenowany ani testowany pod kątem takich zastosowań.

Uczciwość inzynierska niepostrzeżenie zyskała rangę norm etycznych, bo w odpowiedzi firma została praktycznie wymazana z amerykańskiego aparatu bezpieczeństwa. Sprawa znajdzie finał w sądzie, ale życie, jak zwykle, dopisało suspens. Okazało się, że najnowszy model Claude Mythos zagraża bezpieczeństwu systemów finansowych, bo ma zdolność do identyfikowania i wykorzystywania słabych punktów w cyberbezpieczeństwie. Z tego względu firma Anthropic, mimo sporu sądowego z amerykańskim rządem, nie udostępniła go publicznie, do oceny i przygotowania systemów obronnych zaproszono firmy z sektora technologicznego i finansowego. I znów inżynierowie górą.

My też chcemy oddać swoje łamy fachowcom i dlatego temu numerowi „Domeny" – po raz pierwszy w historii pisma – będzie towarzyszył numer specjalny, wydany z okazji organizowanej przez PTI konferencji: „Informatyka 2026 oczami PTI: dokonania, perspektywy, wyzwania" (14 maja br.). Chcemy w ten sposób zatrzymać w migawce czasu poglądy uznanych naukowców, przedstawicieli biznesu oraz administracji dotyczące perspektywy rozwoju technologii informatycznych oraz przychodzących z nimi nowych korzyści, wyzwań i zagrożeń dla Polski. Za kolejne 5 lat, gdy PTI będzie obchodziło jubileusz 50-lecia, sięgniemy do tych tekstów, żeby sprawdzić, na ile trafnie byliśmy w stanie przewidzieć cyfrową przyszłość Polski i świata. Zachęcam gorąco do przeczytania obu numerów

Anna Kniaź
redaktor naczelna

Jak sztuczna inteligencja uczyła się mówić



Technologia głosowa stała się w naszych czasach istotną platformą komunikacji między człowiekiem a sztuczną inteligencją. Współczesne systemy rozpoznawania i syntezy mowy nie tylko słuchają, rozumieją i odpowiadają, ale też są zdolne do rozpoznawania i naśladowania wielu różnych parametrów głosu – dzięki czemu posługują się sztucznie syntetyzowaną mową w sposób właściwy człowiekowi. Tworzy to niezwykłą, można wręcz powiedzieć intymną więź między człowiekiem a maszyną.

Jacek Grabowski

z wykształcenia specjalista gazownictwa i górnictwa naftowego, przygodę z informatyką rozpoczął w końcu lat 80. XX wieku od współpracy z wydawnictwem „Lupus", gdzie publikował teksty głównie w dwutygodniku „PCkurier" i miesięczniku „Enter". Współtwórca pierwszego w Polsce informatycznego czasopisma B2B „MRK" (1997). Były redaktor naczelny miesięcznika „Reset", współpracownik wielu innych tytułów (magazyn „WWW", „IT Reseller", „Komputer Świat"). Obecnie freelancer, współpracuje m.in. z warszawską komunikacją miejską.

Mowa jest jednym z najważniejszych narzędzi porozumiewania się ludzi. Głos to naturalny i intuicyjny sposób komunikacji rozwijający się od wczesnego dzieciństwa i towarzyszący nam przez całe życie. W rozmowie znaczenie mają nie tylko słowa, lecz także intonacja, tempo, pauzy czy akcent, które niosą dodatkowe informacje o naszych intencjach, stanie emocjonalnym i kontekście sytuacyjnym. Coś, co dla nas jest naturalne i normalne, dla maszyny jest jednak wyzwaniem: aby uzyskać realistyczne brzmienie i rozumienie wszystkich aspektów mowy ludzkiej, komputer musi wykonać szereg skomplikowanych operacji. Stąd synteza mowy była zawsze trudnym zadaniem, Jednak zanim w ogóle stała się możliwa, najpierw trzeba było przekształcić analogowy dźwięk w sygnał cyfrowy, możliwy do przetwarzania w komputerze.

Cyfrowe przetwarzanie dźwięku

Teoretyczne początki cyfrowego przetwarzania dźwięku sięgają pierwszej połowy XX w. Istotną rolę odegrała teoria próbkowania opracowana przez Harry'ego Nyquista (1928) i Claude'a Shannona (1949), zgodnie z którą sygnał analogowy może zostać wiernie odwzorowany w postaci cyfrowej, jeżeli częstotliwość próbkowania jest co najmniej dwukrotnie większa od najwyższej częstotliwości w sygnale źródłowym. Próbka dźwięku to jednorazowy pomiar wartości fali akustycznej, który jest zapisywany w formie cyfrowej podczas procesu próbkowania. Rolę częstotliwości próbkowania można porównać do tworzenia filmu z poszczególnych klatek obrazu: kiedy klatek jest zbyt mało, ruchomy obraz przeskakuje albo jest spowolniony. Dopiero odpowiednia liczba klatek zapewnia wrażenie ciągłego ruchu. Podobnie jest z cyfrową reprezentacją dźwięku – im większa liczba próbek, tym dokładniejsze jest cyfrowe odtworzenie analogowego sygnału. Poza częstotliwością próbkowania istotnym parametrem cyfrowego dźwięku jest jego rozdzielczość, czyli liczba bitów opisujących każdą próbkę.

Wraz z rozwojem cyfryzacji dźwięku ewoluowały także metody matematyczne umożliwiające analizę i przetwarzanie sygnałów. W latach 60. pojawiły się m.in. algorytmy szybkiej transformaty Fouriera (*Fast Fourier Transform*, FFT), opracowane przez Jamesa Cooleya i Johna Tukeya (1965), które ułatwiły praktyczne zastosowanie przetwarzania sygnału cyfrowego w obliczeniach komputerowych. Największą zaletą FFT była prędkość działania – umożliwiło to przetwarzanie cyfrowego sygnału dźwiękowego w czasie rzeczywistym i analizę dużych bloków danych.

W latach 70. pojawiły się pierwsze dedykowane układy przetwarzania sygnału (DSP) – specjalizowane procesory, zdolne do wykonywania operacji mnożenia i akumulacji z dużą prędkością. Jednym z pierwszych ich szerokich zastosowań konsumpcyjnych były odtwarzacze płyt CD wprowadzone w 1982 r. Standard CD obejmował zapis i odtwarzanie dźwięku z częstotliwością próbkowania 44,1 KHz i rozdzielczością 16 bitów. W kolejnych latach rozwój internetu i technologii cyfrowej zaowocował pojawieniem się algorytmów stratnej kompresji dźwięku, które umożliwiły znaczną redukcję rozmiaru plików audio przy zachowaniu zadowalającej jakości odsłuchu. Było to istotne m. in. ze względu na powolną transmisję danych sieciowych – zastosowanie kompresji ułatwiało przesyłanie plików dźwiękowych. Najpopularniejszym stratnym formatem służącym do rozpowszechniania plików dźwiękowych w sieci stał się MP3, opracowany w Instytucie Fraunhofera. Rozwój komputerów osobistych i oprogramowania DAW (Digital Audio Workstation) wprowadził przetwarzanie cyfrowego dźwięku „pod strzechy" – również amatorzy dostali do ręki wyrafinowane narzędzia umożliwiające edycję dźwięku oraz remiksowanie i komponowanie własnych utworów.

Współcześnie szczególnie dynamicznie rozwijają się metody wykorzystania sztucznej inteligencji w przetwarzaniu dźwięku cyfrowego. Algorytmy uczenia maszynowego umożliwiają automatyczną separację źródeł dźwięku, inteligentną korekcję barwy, a nawet generowanie realistycznych kompozycji muzycznych. Współczesne badania koncentrują się na połączeniu klasycznych metod DSP z głębokimi sieciami neuronowymi, tworząc nową gałąź – przetwarzanie dźwięku za pomocą sztucznej inteligencji. Technika ta ma istotne znaczenie również w zastosowaniu głosu do komunikacji z SI.

Choć komputerowe rozpoznawanie mowy istniało już od lat 50. XX w., to początkowo technologia ta w ogóle nie miała powiązania ze sztuczną inteligencją, Pierwsze programy uznawane za początki sztucznej inteligencji komunikującej się z człowiekiem mogły dialogować tylko za pomocą tekstu. Przełomem było opracowanie i wdrożenie do rozpoznawania mowy ukrytych modeli Markowa (HMM), co nastąpiło na szerszą skalę w dekadzie lat 90. XX w., a także opracowanie prostych reguł rozumienia języka przez sztuczną inteligencję. Umożliwiło to powstanie pierwszych systemów IVR (Interactive Voice Response) rozpoznających niewielką liczbę poleceń głosowych. Nie przypominało to jeszcze swobodnego dialogu, ale systemy informatyczne stały się zdolne do zanalizowania poleceń głosowych i wykonywania na tej podstawie zdefiniowanych zadań.

Na przełomie XX i XXI w. pojawiły się bardziej zaawansowane rozwiązania łączące technologię rozpoznawania i analizy mowy ze sztuczną inteligencją. Istotnym krokiem było wprowadzenie statystycznych metod badania języka naturalnego (*NLP – Natural Language Processing*, warto zwrócić uwagę, że skrót NLP odnosi się często także do tzw. programowania neurolingwistycznego, które jest osobną dziedziną). Mówiąc ogólnie, metoda NLP polega na rozbijaniu języka na krótsze, bardziej podstawowe elementy, podejmowaniu prób zrozumienia zależności pomiędzy poszczególnymi częściami składowymi oraz badaniu sposobu, w jaki łączą się one ze sobą, tworząc sens. Dzięki rosnącym korpusom językowym i zbiorom nagrań do trenowania SI, rozpoznawanie mowy zaczęło współpracować z modułami rozumienia języka NLP. W ten sposób pojawiła sie pierwsza generacja systemów, które były zdolne nie tylko do rozpoznawania słów, lecz również zrozumienia mowy ludzkiej. Jednak w tym pierwszym okresie komunikacja człowieka z maszyną za pomoca mowy była wciąż jeszcze bardzo ograniczona i „sztywna".

Głos i sztuczna inteligencja

Głos ludzki to zjawisko akustyczne powstające w wyniku przepływu powietrza z płuc przez krtań, gdzie drgają fałdy głosowe, a następnie dźwięk jest modyfikowany w rezonatorach (jama ustna, nosowa, gardłowa) oraz artykułowany przez język, wargi i żuchwę. Fizyczne cechy głosu obejmują m.in. wysokość dźwięku, głośność i barwę (tembr). Tembr głosu jest cechą unikalną dla każdego człowieka i zależy od budowy jego narządów głosowych. Wysokość głosu i sposób artykulacji zależy od cech fizycznych, właściwych np. dla płci (głos kobiet jest ogólnie wyższy niż mężczyzn), zmienia się także wraz z wiekiem człowieka – np. na starość głos staje się słabszy, mniej stabilny i wyraźny. Na jakość mowy mogą wpływać choroby krtani, gardła i jamy ustnej, a także inne czynniki. Trudno nawet w skrócie opisać wszystkie aspekty mowy i głosu, które wpływają na jego brzmienie, intonację, artykulację itd. Nic więc dziwnego, że właściwe odwzorowanie głosu ludzkiego w interakcjach z komputerami ma długą i skomplikowaną historię.

Siri – pierwszy interaktywny asystent głosowy

Rewolucją w komunikowaniu się człowieka z maszyną był asystent głosowy Apple Siri, wprowadzony do smartfonów Apple w 2011 r. Apple przejęło Siri w 2010 r. Firma Siri rozwijała już wcześniej technologie rozpoznawania mowy i NLP w połączeniu z uczeniem maszynowym i uczeniem głębokim, a także syntezą TTS (*Text-To-Speech*). Ze względu na konieczność korzystania z wielkich zbiorów danych i odpowiednio silnych komputerów do ich przetwarzania, Siri działała w chmurze, wymagając stałego połączenia z internetem.

W tamtych czasach systemy głosowe opierały się na dwóch rodzajach syntezy, Pierwszą, najstarszą i jednocześnie najbardziej rozpowszechnioną w technologii asystentów głosowych, była tzw. synteza konkatenacyjna, polegająca na cięciu nagrań na fonetyczne dźwięki i łączeniu ich (konkatenacja) w nowe słowa i zdania. W metodzie tej system korzysta z dużej bazy nagrań lektora. Każde słowo, sylaba, fonem lub difon (przejście między dwoma fonemami) jest wcześniej nagrane i opisane. Wprowadzany tekst jest dzielony na jednostki fonetyczne, a system wyszukuje w bazie najlepiej pasujące fragmenty dźwięku i łączy je płynnie, tworząc kompletną wypowiedź. Wadami konkatenacji są mała skalowalność (trudna adaptacja do nowych języków i głosów), konieczność tworzenia olbrzymich baz z nagraniami, niewielkie możliwości oddania emocji i problemy z nietypowymi zwrotami (np. skrótami czy nazwami własnymi). Zaletą jest naturalnie brzmiący dźwięk i wysoka zrozumiałość syntetycznej mowy.

Drugim sposobem syntetyzowania mowy był tzw. wokoder, czyli procesor tworzący mowę parametrycznie, bez wykorzystania wcześniejszych nagrań jak w konkatenacji, lecz poprzez łączenie źródła z odpowiednimi filtrami sztucznie modelującymi kanał głosowy. W klasycznych wokoderach wykorzystywano technikę liniowego kodowania predykcyjnego (*LPC – Linear Predictive Coding*). Metoda ta wykorzystuje fakt, że ludzki głos ma dużą redundancję i polega na przewidywaniu kolejnych próbek sygnału na podstawie poprzednich, co pozwala na jego efektywne zakodowanie i odtworzenie. Tak zsyntetyzowany sygnał mowy charakteryzował się często sztucznym, „metalicznym" brzmieniem, jednak dzięki temu, że wokoder generuje wszystko sztucznie był znacznie bardziej elastyczny od metody konkatenacyjnej i – co ważne – nie wymagał żadnych zewnętrznych baz nagrań. Synteza za pomocą wokodera byla więc wykorzystywana najczęściej w systemach, dla których priorytetem była elastyczność, skalowalność i rozmiar danych. Tam, gdzie priorytetem była zrozumiałość i naturalność wypowiedzi (np. w asystentach głosowych) stosowano metodę konkatenacyjną.

Wprowadzenie głębokiego uczenia i modeli językowych

W połowie drugiej dekady XXI w. pojawiły się pierwsze systemy rozpoznawania mowy bazujące na głębokich i konwolucyjnych sieciach neuronowych (DNN i CNN) oraz modelach pamięci długotrwałej krótkoterminowej (*LTSM – Long Short Term Memory*). Jednym z pierwszych był WaveNet Google – model generatywny wytrenowany na próbkach mowy ludzkiej. Tworzy on przebiegi wzorców mowy, przewidując, które dźwięki najprawdopodobniej będą następować po sobie, budowane po jednej próbce na raz, z prędkością do 24 tys. próbek dźwięku na sekundę. W 2017 r. Google opracował także pierwsze transformery wykorzystujące tzw. mechanizm uwagi (*selfattention*), który symuluje działanie ludzkiej uwagi poprzez przypisywanie różnych poziomów ważności różnym słowom w zdaniu. Zastosowanie tych rozwiązań spowodowało, że SI już nie tylko reagowała na polecenia głosowe, lecz zaczęła prowadzić konwersację z człowiekiem – rozumieć kontekst, zapamiętywać wcześniejsze wypowiedzi i adaptować odpowiedzi do użytkownika. Mogła łączyć ze sobą dane z różnych źródeł, generując wypowiedzi uwzględniające wiele naturalnych czynników. Dzięki temu rozmowa ze sztuczną inteligencją stała się wielomodalnym dialogiem przypominającym do złudzenia normalne porozumiewanie się ludzi.

Współczesna neuronowa synteza głosu dzieli się na dwie główne techniki – starszą dwustopniową, w której mamy system przetwarzania tekstu na mowę (TTS) i wokoder, oraz najnowszą end-to-end, czyli przetwarzanie jednostopniowe, w którym nie stosuje się już osobnego wokodera. W metodzie dwustopniowej system przetwarzania TTS tworzy tzw. mel-spektrogram, czyli wizualizację dźwięku używającą skali Mel na osi Y zamiast standardo-

wej skali liniowej częstotliwości w Hz. Skala melowa jest nieliniowa i lepiej odpowiada ludzkiemu słuchowi, ponieważ naśladuje sposób, w jaki człowiek odbiera wysokość dźwięku. Model przewiduje, jak powinna brzmieć wypowiedź, jakich użyć akcentów i pauz itd., a następnie wokoder na podstawie tych danych syntetyzuje mowę. Metoda ta daje na wyjściu naturalne brzmienie, umożliwia klonowanie głosu (czyli syntetyzowanie głosu konkretnej osoby z jego wszystkimi cechami), uwzględnia także dobre odwzorowanie emocji. Połączenie TTS z wokoderem daje jednak czasem niepożądane artefakty, poza tym wymaga większych zasobów niż wcześniejsze metody.

W metodzie jednostopniowej zastosowany jest tylko jeden duży model, który od razu przetwarza tekst do sygnału mowy bez współudziału modelu wokodera. Taka metoda używana jest np. przez model VALL-E Microsoftu czy ChatGPT. Metoda jednostopniowa jest szybsza, zapewnia lepszą spójność mowy, praktycznie unika powstawania artefaktów, umożliwia klonowanie głosu z kilku sekund nagrania, a także łatwe sterowanie stylem mówienia i emocjami zawartymi w głosie. Wymaga jednak bardzo wysokiej mocy obliczeniowej do obsługi wielkiego modelu.

Co dalej?

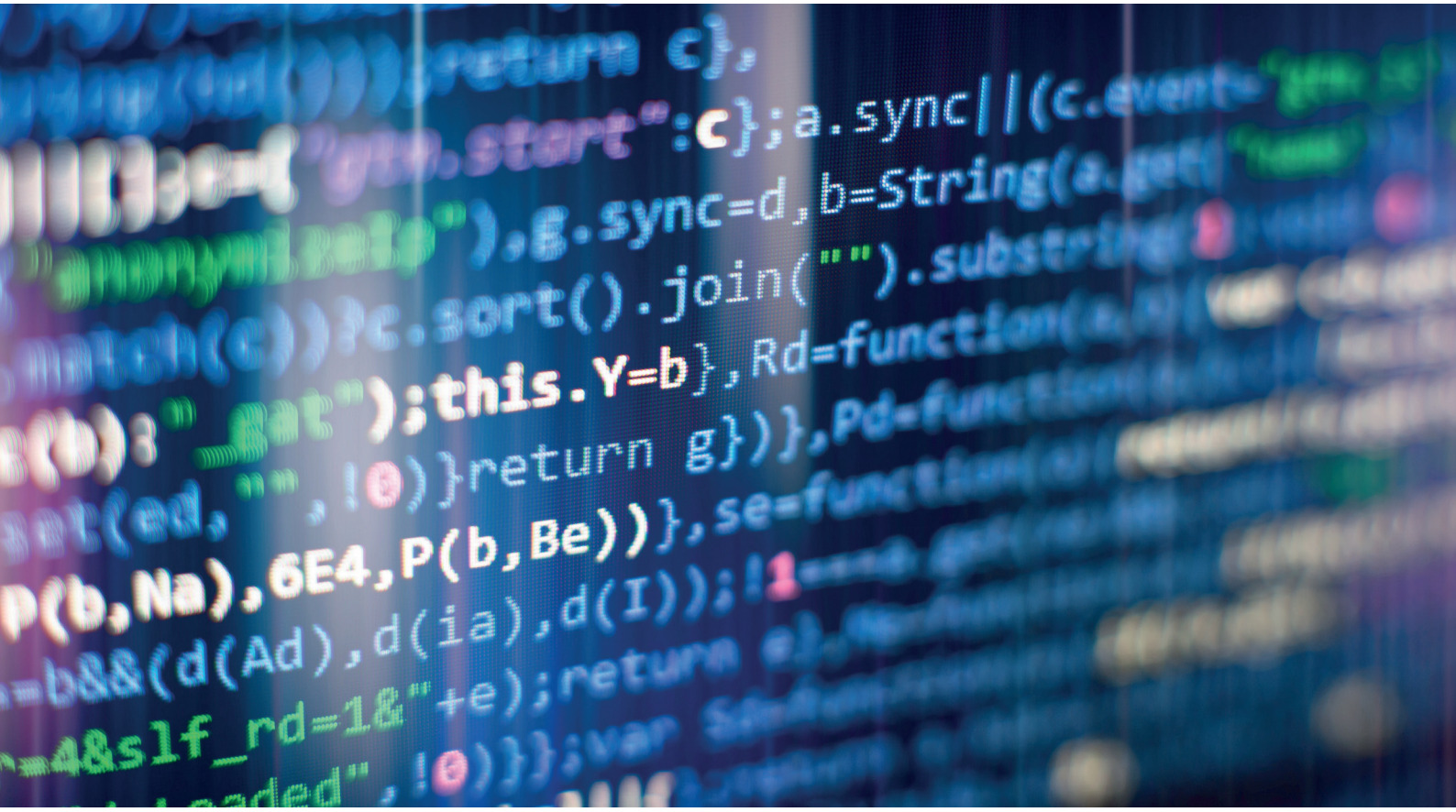
Sposób generowania mowy i jej rozpoznawania przeszedł więc długą drogę aż do czasów, kiedy w technologii głosowej używa się powszechnie sieci neuronowych i modeli językowych. Ich zastosowanie umożliwiło osiągnięcie naturalności „wypowiedzi" komputerowej i zwiększyło możliwości przetwarzania sygnału mowy. Obecnie jesteśmy w stanie skopiować głos i styl wypowiedzi konkretnej osoby w taki sposób, że złudzenie odbiorcy może być praktycznie pełne, co ułatwia tworzenie deep-fake'ów. Komputery „mówią" płynnie, bez zacięć i nienaturalnych dźwięków, ale ma to też niestety swoje wady. Warto bowiem pamiętać, że sztuczna inteligencja nie jest świadoma tego, że mówi – to jest tylko funkcja realizowana przez odpowiednie oprogramowanie i sprzęt, jednak odbiorca może rozumieć ten fakt zupełnie opacznie. Swobodny dialog z robotem, który jest w stanie symulować emocje i kształtować swoje wypowiedzi identycznie jak człowiek, sprawia wrażenie rozmowy z istotą faktycznie rozumiejącą i czującą podobnie, co prowadzi czasem do fatalnych w skutkach nieporozumień.

Etyczne problemy, na jakie napotyka synteza mowy, to tylko jeden z wielu kłopotów. Rozpoznawanie mowy osiągnęło pożądaną precyzję w laboratoriach i kontrolowanych warunkach, ale w realnym świecie bywa znacznie trudniejsze. Hałas ulicy, dźwięki w tle, echo w pomieszczeniach, a nawet różnice w akustyce pomieszczenia mogą znacząco obniżyć skuteczność systemu. Wyzwanie to staje się szczególnie widoczne w zastosowaniach medycznych, motoryzacji czy w przemyśle, gdzie pomyłki mogą prowadzić do poważnych konsekwencji.

Systemy głosowe w praktyce napotykają także trudności w rozpoznawaniu mniej rozpowszechnionych języków, regionalnych dialektów czy indywidualnych akcentów. W dodatku rozpoznanie słów to jedno, a zrozumienie sensu – drugie. AI musi nie tylko przetworzyć ciąg fonemów, ale też uchwycić kontekst, intencję i ewentualne niuanse emocjonalne. Choć obecnie udaje się to dość dobrze, niedokładności w analizie kontekstu i rozpoznaniu języka mogą prowadzić do błędnych odpowiedzi lub innych niepożądanych efektów działań systemu, co jest szczególnie krytyczne w zastosowaniach medycznych lub finansowych.

Sztuczna inteligencja musi być również wystarczająco elastyczna, aby adaptować się do zmian ludzkiego głosu w zależności od wieku, emocji, zdrowia czy warunków otoczenia. To wymaga ciągłego uczenia maszynowego i aktualizacji modeli, co łączy się z dużym nakładem mocy obliczeniowej i wiążącymi się z tym kosztami. Optymalizacja tych parametrów jest kluczowa, aby systemy głosowe mogły działać płynnie i responsywnie w codziennym użyciu. Tak więc mimo, że osiągnęliśmy już bardzo wiele w połączeniu sztucznej inteligencji z naturalnie brzmiącymi wypowiedziami głosowymi w różnych językach, to nadal technologie te wymagają dalszych badań i modyfikacji, żeby mogły się rozpowszechnić w wielu zastosowaniach.

Prymat algorytmów



3 listopada 2025 r. Wojewódzki Sąd Administracyjny w Warszawie wydał niepokojący wyrok, w którym uznał prymat algorytmów nad prawami podstawowymi. Zrobił to na miesiąc przed 25. rocznicą Karty praw podstawowych Unii Europejskiej, wyznaczającej 50 praw, wolności i zasad, w tym ochronę danych osobowych oraz poszanowanie życia prywatnego i rodzinnego, uznanych przez Unię w celu stworzenia przestrzeni wolności, bezpieczeństwa i sprawiedliwości dla jej obywateli.

Joanna Karczewska

absolwentka Wydziału Elektroniki PW z ponad 40-letnim doświadczeniem w informatyce. Jako certyfikowany audytor systemów informatycznych – CISA – specjalizuje się w audytach informatycznych w jednostkach sektora finansów publicznych. Pełni także funkcję inspektora ochrony danych w placówkach oświatowych. Jako Expert Reviewer uczestniczyła w opracowaniu metodyk COBIT5 i COBIT 2019, ITAF 4th Edition oraz publikacji ISACA dotyczących Digital Trust Ecosystem Framework. Bierze udział w konsultacjach aktów prawnych dotyczących bezpieczeństwa informacji, cyberbezpieczeństwa i ochrony danych osobowych, również na forum Komisji Cyfryzacji, Innowacyjności i Nowoczesnych Technologii Sejmu RP. Uznana w 2022 roku za jedną z Europe's Top Cyber Women. Ekspert Najwyższej Izby Kontroli.

O moich zmaganiach z polską wersją FATCA pisałam w numerach 2/2022, 1/2023, 1/2025 i 2/2025 „Domeny". Dla przypomnienia: w 2014 roku Polska podpisała umowę z USA w sprawie poprawy wypełniania międzynarodowych obowiązków podatkowych oraz wdrożenia **amerykańskiego ustawodawstwa FATCA** (*The Foreign Account Tax Compliance Act*).

Algorytmy bankowe

Umowa dotyczy m.in. tropienia przez polskie banki tzw. rachunków raportowanych poprzez **elektroniczne wyszukiwanie w swoich bazach danych** klientów urodzonych w Stanach Zjednoczonych, którzy na podstawie amerykańskiego prawa *ius soli* są uznawani za obywateli USA. Niestety, obie strony umowy zapomniały o **XIV Poprawce do Konstytucji USA**, która od **9 lipca 1868 r.** wprowadziła wyjątek od tego prawa.

Ja właśnie jestem tym wyjątkiem. Z powodu niewiedzy, braku starannej oceny skutków dla ochrony danych osobowych w banku ING, braku stosownych procedur i braku szacunku banku dla prywatności swoich klientów moje dane osobowe trafiły do Krajowej Administracji Skarbowej (KAS) – będącej częścią Ministerstwa Finansów (MF) – która w ramach **automatycznej wymiany informacji** przekazała moje dane do Internal Revenue Service USA.

Złożyłam skargę na ING, MF i KAS do Prezesa Urzędu Ochrony Danych Osobowych (UODO) na ich bezprawne i nielegalne przetwarzanie moich danych osobowych. Prezes UODO umorzył postępowanie w sprawie MF i odmówił uwzględnienia wniosku w sprawie KAS. Nadal czekam na rozpatrzenie przez WSA mojej skargi na te decyzje.

W przypadku ING Banku Śląskiego prezes UODO postanowił odmówić uwzględnienia mojej skargi. W uzasadnieniu zaznaczył, że:

- „Skarżąca urodziła się w Stanach Zjednoczonych Ameryki, co zgodnie z prawem amerykańskim **pozwoliło przyjąć**, że nabyła obywatelstwo Stanów Zjednoczonych Ameryki. Wobec powyższego Bank wielokrotnie wzywał Skarżącą do złożenia oświadczenia w sprawie FATCA oraz złożenia dokumentów potwierdzających, że nie jest obywatelem Stanów Zjednoczonych, mimo urodzenia się w tym kraju. Ostatecznie wobec niezłożenia przez Skarżącą stosownej dokumentacji, Bank przekazał dane Skarżącej do KAS na podstawie art. 4 ust. 1 pkt 2 lit. a ustawy FATCA".

- „Bank wyjaśnił również, że oświadczenie posiadacza rachunku, że nie jest on obywatelem amerykańskim ani amerykańskim rezydentem dla celów podatkowych, składane jest w formie wydawanego przez Urząd Skarbowy Stanów Zjednoczonych (IRS) formularza W-8. W formularzu tym znajduje się nazwa klienta, a nie samo nazwisko, w związku z czym w oświadczeniu o statusie FATCA i CRS dla osób fizycznych stanowiącym formularz Banku, Klient wskazuje imię/imiona oraz nazwisko".

- „W ocenie Prezesa Urzędu Ochrony Danych Osobowych, w świetle ustalonego stanu faktycznego, Bank posiadał podstawy prawne do udostępnienia danych osobowych Skarżącej do KAS w postaci art. 6 ust. 1 lit. c RODO, w związku z istnieniem obowiązków prawnych realizowanych przez administratora danych na podstawie art. 4 ust. 1 pkt 2 lit. a ustawy FATCA w zw. z art. 2 ust. 2 lit. a umowy FATCA. Zakres przekazanych danych osobowych Skarżącej był zgodny z ww. przepisami".

Swoją decyzją prezes UODO potwierdził, że:

- **nigdy nie słyszał o XIV Poprawce do Konstytucji USA. Wielki międzynarodowy bank też nie, chociaż Poprawka obowiązuje już ponad 150 lat;**
- **nie sprawdził, czy Bank zażądał ode mnie formularza W-8 (nigdy nie zażądał);**
- **nie uznał innych oświadczeń, które składałam, chociaż załącznik I do umowy FATCA dopuszcza inne sposoby dokumentowania braku obywatelstwa amerykańskiego;**
- **nie dociekał, dlaczego obywatel tylko i wyłącznie polski ma uzyskiwać potwierdzenie braku obywatelstwa amerykańskiego od amerykańskiego urzędu skarbowego, a nie od wyznaczonego organu polskiego.**

Odmowa była dla mnie tak kuriozalna, że wniosłam skargę do Wojewódzkiego Sądu Administracyjnego w Warszawie (WSA). W uzasadnieniu skargi przywołałam następujące dokumenty: RODO – motyw 76, Kartę praw podstawowych UE, Powszechną Deklarację Praw Człowieka, Rezolucję P8_TA(2018)0316 „Szkodliwe skutki FATCA dla obywateli UE" Parlamentu Europejskiego, Europejską konwencję praw człowieka oraz Oświadczenie 01/2019 w sprawie FATCA, Wytyczne 2/2018 i Wytyczne 2/2020 Europejskiej Rady Ochrony Danych (EROD). Zaznaczyłam, że moja sprawa dotyczy przede wszystkim respektowania art. 7 „Poszanowanie życia prywatnego i rodzinnego" oraz art. 8 „Ochrona danych osobowych" Karty praw podstawowych Unii Europejskiej w przypadku moim i co najmniej kilkuset innych obywateli polskich urodzonych w USA będących wyjątkami podobnie jak ja. Zakwestionowałam także bezstronność postępowania prezesa UODO oraz równe traktowanie obu stron wymagane zgodnie z art. 8 § 1 Kodeksu postępowania administracyjnego. W odpowiedzi na skargę skierowaną do WSA prezes UODO stwierdził m.in.: „Odnosząc się również do zarzutu niedokonania przez Prezesa Urzędu oceny analizy ryzyka naruszenia praw i wolności osób fizycznych, których dane

dotyczą, a przez to brak ustalenia, czy Bank przeprowadził analizę ryzyka, w jakim stopniu brak poinformowania o wymogu przedstawienia formularza W-8 lub paszportu oraz brak starannego przeglądu prawa amerykańskiego dotyczącego obywatelstwa, może doprowadzić do naruszenia praw lub wolności osoby fizycznej poprzez przekazanie danych osobowych podmiotom nieuprawnionym, wskazać należy, że powyższe zagadnienie również nie było przedmiotem postępowania zakończonego skarżoną decyzją. Podkreślić należy, że w postępowaniu o sygn. xxx Prezes Urzędu **badał wyłącznie istnienie podstaw legalizujących przekazanie danych osobowych** Skarżącej do KAS, i na podstawie zebranego materiału dowodowego ustalił, że odbyło się to zgodnie z obowiązującymi w Polsce przepisami prawa, zarówno w odniesieniu do samego przekazania danych, jak i w odniesieniu do przekazanych kategorii tych danych".

Jakże różne jest to stanowisko od opinii o wymianie informacji podatkowych z innymi państwami skierowanej przez Prezesa UODO do Ministerstwa Finansów w 2023 r.: https://legislacja.rcl.gov.pl/docs//2/12369201/12949557/12949560/dokument618966.pdf

3.11.2025 r. WSA oddalił moją skargę. Wystąpiłam o uzasadnienie. Otrzymałam krzywy, lekko rozmazany skan 16-stronicowego wydruku wyroku wraz z uzasadnieniem zawierającym cytat mojej wypowiedzi. Ponieważ rozprawa nie była nagrywania, nie ma pewności, że cytat jest dokładny. Na pewno narusza moją prywatność. Co ciekawe, WSA potwierdził, że Polska nie zastosowała się do wezwań Parlamentu Europejskiego do przeglądu i dostosowania takich umów jak FATCA do wymogów RODO.

Podsumowując: Kobieto, sama jesteś sobie winna. Trzeba było spowiadać się algorytmom bankowym ze swojego życia prywatnego i rodzinnego, a nie narzekać teraz na naruszenie ochrony Twoich danych osobowych ze wszystkimi możliwymi skutkami. Podobno *ignorantia legis non excusat* – w moim przypadku prezes UODO i sędziowie WSA uznali, że Bank miał prawo do takiej ignorancji.

Algorytmy skarbowe

22 grudnia 2025 r. otrzymałam e-mail z powiadomieniem z e-Urzędu Skarbowego. Informował, że na moim koncie jest dokument do odebrania z terminem odbioru 05.01.2026 r. i zastrzeżeniem: *„Po upływie terminu odbioru dokument zostanie automatycznie uznany za doręczony"*. Od razu spróbowałam się zalogować do systemu, spodziewając się wezwania do urzędu skarbowego. Bezskutecznie. Tego dnia, w święta i po świętach. Wreszcie 28 grudnia o godz. 00:14 wysłałam formalne zgłoszenie na podany adres info.eurzad@mf.gov.pl. Z braku odpowiedzi, oprócz nadania numeru TICKET#, wysłałam zgłoszenie do MF (e-Doręczenia 30.12.2025), IOD MF (e-mail 30.12.2025) oraz Urzędu Skarbowego Warszawa-Praga (e-mail 5.01.2026). Zespół Pomocy Technicznej MF przysłał e-mail dopiero w dniu 5.01.2026 r. o godz. 15:13. Wyjaśnienie było banalne: „przeprowadzona analiza danych technicznych wykazała, że problem z logowaniem do systemu e-Urząd Skarbowy wynikał z próby logowania z użyciem zapisanego wcześniej bezpośredniego linku do strony logowania (np. zapisanego jako zakładka w przeglądarce)". Zalogowałam się skutecznie i odebrałam dokument. Na szczęście, był to tylko list ogólny MF informujący o Krajowym Systemie e-Faktur (KSeF). Czekałam **DZIEWIĘĆ** dni na wyjaśnienie, czy nie podpadłam skarbówce.

Był to przedsmak emocji, które Ministerstwo Finansów przygotowało 2 milionom osób prowadzących jednoosobową działalność gospodarczą (JDG). Wystawiam 2–3 faktury miesięcznie. Dotychczas robiłam to w Excelu, drukowałam w formacie PDF i wysyłałam klientom oraz mojemu księgowemu. Od 1 kwietnia 2026 r. będę to robić w darmowej Aplikacji Podatnika KSeF 2.0, która de facto nie jest aplikacją tylko zwyczajnym logowaniem się do systemu KSeF na portalu podatki.gov.pl. System produkcyjny ruszył 1 lutego 2026 r. Dopiero wtedy mogłam zalogować się do niego i obejrzeć wszystkie moduły i opcje. Najbardziej interesowała mnie funkcja wystawiania faktur. Niestety, na stronach MF nie ma instrukcji wypełniania poszczególnych pól z danymi do faktury.

Jeszcze bardziej emocjonujące będzie wystawianie faktur w trybie offline w przypadku tzw. offline24, niedostępności systemu lub awarii KSeF. Podatnik musi zastosować obowiązujący wzór struktury logicznej FA(3) faktury ustrukturyzowanej i oznaczyć fakturę dwoma kodami: QR I – z napisem OFFLINE oraz QR II – z napisem CERTYFIKAT. Dwa miliony JDG musi rozpracować najpóźniej do końca roku, jak to zrobić bez ponoszenia dodatkowych kosztów. W internecie pełno jest ofert darmowych i płatnych aplikacji komercyjnych. Niestety, MF nie przygotowało oficjalnej darmowej aplikacji ani szczegółowego poradnika, jak postępować w trybie offline.

O kosztach KSeF-u dla przedsiębiorców rozmawiano na posiedzeniu Komisji Gospodarki i Rozwoju Sejmu RP w dniu 7 stycznia 2026 r. (https://www.sejm.gov.pl/Sejm10.nsf/PosKomZrealizowane.xsp?komisja=GOR#102).

Przedstawiciel KAS z dużym samozadowoleniem twierdził, że „koszty są – od zera po miliony złotych", ale są darmowe aplikacje, darmowe szkolenia ministerialne i darmowe konsultacje w urzędach skarbowych. Poseł Robert Dowhan słusznie zauważył, że „... to się za darmo nie weźmie. Ktoś musi nad tym posiedzieć, ktoś musi się wdrożyć, ktoś musi to dobrze opanować, żeby nie popełnić błędu i nie sparaliżować firmy". W podobnym tonie wypowiedział się poseł Andrzej Gawron: „... jest darmowa aplikacja – no państwo łaski nie robi, naprawdę, jeżeli wymaga od tych wszystkich przedsiębiorców. I to są koszty, chociażby czasu, który przedsiębiorca musi poświęcić na przeszkolenie się, zrozumienie tego systemu".

Posłanka Bożena Lisowska trafnie określiła znaczenie KSeF-u:
„... zmienia się całkowicie podejście do dokumentu handlowego, jakim jest faktura. Do tej pory skuteczność prawna takiego dokumentu zależała i była wyłącznie zależna od wystawiającego dokument i pobierającego ten dokument, czyli pomiędzy handlowcem a sprzedawcą – czyli odbiorcą i wystawiającym ten dokument. Od 1 lutego tak naprawdę skuteczność prawna dokumentu będzie zależeć od systemu administracji skarbowej – ponieważ brak poprawnego przetworzenia i uwierzytelnienia systemowego będzie powodował, że tak naprawdę system nie będzie widział tego dokumentu. I to jest ta największa odpowiedzialność właśnie Ministerstwa Finansów jako jednostki odpowiedzialnej za wdrożenie, ale także całego systemu administracji skarbowej w Polsce".

Od 1 lutego 2026 r. KSeF będzie działać w trybie 24/7/365 do końca świata, a może i dłużej. Awarie czy błędy mogą się zdarzyć, natomiast przystąpienie do usunięcia awarii czy błędu musi być natychmiastowe. Tak samo **musi** działać wsparcie – bez względu na święta, długie weekendy czy wakacje. **Nadal pozostaje pytanie o przestrzeganie naszych praw podstawowych, gdy fiskus będzie o nas wiedzieć coraz więcej z przetwarzanych faktur.**

Algorytmy sejmowe

W numerze 4/2025 „Domeny" pisałam o międzynarodowej konferencji „Demokracja parlamentarna i ustawodawstwo w erze sztucznej inteligencji. Perspektywa europejska", która odbyła się 13 października 2025 r. w Sejmie RP. Polecałam wystąpienie Fotisa Fitsilisa, kierownika departamentu Dokumentacji Naukowej i Nadzoru Służby Naukowej Parlamentu Hellenów. Pan Fitsilis jest współautorem Wytycznych stosowania sztucznej inteligencji w parlamentach, opublikowanych w 2024 r. (https://interoperable-europe.ec.europa.eu/collection/egovernment/solution/hocrt/document/guidelines-ai-parliaments). Dokument opisuje 40 szczegółowych wytycznych w sześciu sekcjach obejmujących kluczowe kwestie: *ethical principles, artificial general intelligence (AGI) and human autonomy, AI privacy and security, AI governance and oversight, AI system design and operation, AI capacity building and education*.

Wytyczne kładą nacisk na zasady etyczne, w tym rozliczalność, przejrzystość i uczciwość. Podkreślają znaczenie poszanowania godności ludzkiej, prywatności i różnorodności kulturowej, a jednocześnie odnoszą się do stronniczości w danych i algorytmach oraz podejmowania zdecydowanych działań w celu ochrony danych osobowych i zapobiegania cyberatakom. Wytyczne są adresowane do wszystkich osób zainteresowanych maksymalizacją pozytywnych efektów stosowania sztucznej inteligencji w organach ustawodawczych przy jednoczesnej minimalizacji potencjalnych ryzyk.

6 listopada 2025 r. na posiedzeniu Podkomisji stałej do spraw sztucznej inteligencji i przejrzystości algorytmów Marzena Laskowska, Dyrektor Biura Ekspertyz i Oceny Skutków Regulacji Kancelarii Sejmu przedstawiła systemowe podejście do zastosowania sztucznej inteligencji w funkcjonowaniu polskiego Sejmu (https://www.sejm.gov.pl/Sejm10.nsf/PosKomZrealizowane.xsp?komisja=CNT01S#13), w tym prace nad regulacją, która określałyby zasady stosowania sztucznej inteligencji w Kancelarii Sejmu. Regulacja znajdzie zastosowanie do wszystkich systemów sztucznej inteligencji niezależnie od sposobu ich pozyskania (ogólnodostępne, licencyjne, na subskrypcję oraz wdrażane siłą własną). Zasady zostały przyjęte w dniu 14 listopada 2025 r. „jako ramy dla korzystania z tej technologii, przy jednoczesnym poszanowaniu fundamentalnych wartości demokratycznych, praw człowieka i zasad państwa prawnego". Obejmują odpowiedzialność, ochronę danych i bezpieczeństwo informacji, przejrzystość i wiarygodność oraz jakość i ciągłość. Ich naruszenie może skutkować odpowiedzialnością prawną, w szczególności porządkową lub dyscyplinarną. Uzupełniająco w załączniku podano wskazówki i przykłady dopuszczalnego i niedopuszczalnego wykorzystania AI oraz ryzyka i wyzwania związane ze stosowaniem AI.

Są więc ludzie i instytucje, którzy chcą dbać o nasze prawa podstawowe w swojej pracy – oby było ich jak najwięcej.

Algorytmy mObywatelskie

20 grudnia 2025 r. sprzedałam samochód. To była sobota. Zamiast planować wizytę w poniedziałek w moim urzędzie dzielnicowym w Warszawie postanowiłam skorzystać z funkcji zgłoszenia zbycia pojazdu w systemie mObywatel, tak gorąco polecanym i promowanym przez Ministra Cyfryzacji. W niedzielę 21 grudnia 2025 r. doczytałam instrukcję na portalu gov.pl i złożyłam zawiadomienie przekonana, że sprawę załatwię szybko. Nastąpiła cisza. 5 stycznia otrzymałam e-mail, że w systemie ePUAP czeka nowa korespondencja. Było to pismo z Ministerstwa Cyfryzacji potwierdzające zbycie pojazdu wystawione przez mój urząd dzielnicowy. Czekałam na ten dokument **PIĘTNAŚCIE** dni. Gdybym poszła do urzędu, załatwiłabym sprawę od ręki. Na dodatek okazuje się, że w zakładce „Twoje sprawy" w systemie mObywatel nadal jest tylko moje zawiadomienie o zbyciu pojazdu i nie ma żadnej informacji o kolejnych etapach załatwienia sprawy, zaś jej status określono jako „Wysłane".

Serwis mObywatel gromadzi bardzo dużo danych osobowych z naszych dokumentów (dowód osobisty, prawo jaz-

dy) oraz różnych legitymacji i praw wykonywania zawodu. Ich przetwarzanie wymaga cyberbezpieczeństwa, zapewniającego ochronę naszych praw podstawowych. 10 grudnia 2025 r. Minister Cyfryzacji zapowiedział uruchomienie nowego portalu cyber.gov.pl, który ma być naszym centrum cyberbezpieczeństwa.

Obejrzałam i przeżyłam głęboki szok. Już dawno nie widziałam tak niedopracowanej strony zarówno od strony informacyjnej, jak i graficznej. Pominę mdłe logo CY3CR. Mam wrażenie, że zadanie realizował niedouczony stażysta, który naprędce losowo zebrał odnośniki do różnych stron urzędowych o cyberbezpieczeństwie i umieścił je na portalu bez sprawdzenia spójności i aktualności polecanych treści oraz zadbania o jednolitość wizualną. Wrażenie chaosu jest wręcz przytłaczające.

Obywateli zachęca się do przejścia do:

- testu o podstawach cyberbezpieczeństwa przygotowanego przez Rządowe Centrum Bezpieczeństwa (RCB) w 2024 r. jako jeden z wielu poradników bezpiecznych zachowań i umieszczonego na platformie view.genially.com;
- quizu o phishingu przygotowanego przez CSIRT KNF w 2021 r.;
- poradnika dotyczącego cyberbezpieczeństwa przygotowanego przez MSWiA jako jeden z poradników Regionalnego Systemu Ostrzegania;
- bazy wiedzy MC – w zakładce „Dla każdego – cyberhigiena" ostatni wpis ma datę 06.03.2025 r.;

Uprzedzając ewentualne zarzuty złośliwości, przeprowadziłam prosty test. Poszukałam porad dotyczących **budowania haseł do systemów w**:

- **brytyjskim www.NCSC.gov.uk**
  Po wejściu na stronę wystarczy kliknąć na lupę (Search) w prawym górnym rogu ekranu, wpisać „**password**" i wcisnąć <Enter> – wyświetli się długa lista poradników i innych informacji, które można dodatkowo filtrować w zależności od użytkownika, np. „You and your family";
- **francuskim CYBER.gouv.fr**
  Wystarczy wpisać „**mot de passe**" przy lupie w prawym górnym rogu i i wcisnąć <Enter> – wyświetli się lista stron, z której należy wybrać pierwszą pozycję „10 règles d'or en matière de sécurité numérique", by zobaczyć w punkcie 3. podpowiedzi dotyczące haseł;
- **europejskim www.ENISA.europa.eu**
  Należy wybrać kafel „Citizens", następnie kliknąć na „Cyber Hygiene" i można od razu zapoznać się z podstawowymi zasadami cyberhigieny, w tym „1. **Use Strong Passwords**";
- **polskim CYBER.gov.pl**
  Tu nie ma wyszukiwarki – obywatel musi sam klikać i przewijać i przy odrobinie szczęścia trafi na jakiekolwiek wskazówki.

Tak więc (m)Obywatel jest zdany na siebie w kwestii ochrony swoich praw podstawowych w cyberświecie.

Algorytmy polityczne

Kilka tygodni temu rozmawiałam z prominentnym politykiem, który entuzjastycznie popiera cyfryzację państwa na szeroką skalę. Zwróciłam mu uwagę, że w swoich wystąpieniach nie mówi o prywatności, która jest stale zagrożona w dobie powszechnej informatyzacji, internetu i sztucznej inteligencji. Odpowiedział pytaniem: „A co to jest prywatność?". Zaskoczona, dopiero po chwili stwierdziłam, że w takim razie warto o tym porozmawiać. Moja przeprawa z FATCA, przygody podatników z KSeF-em i cytowana wypowiedź potwierdzają, że wszyscy mamy jeszcze wiele do zrobienia, by algorytmy cyfrowe nie zdominowały naszych praw podstawowych.

Wszystkie informacje zawarte w analizie są podanewedług stanu na dzień 15 lutego 2026 r. W opracowaniu tekstu Autorka nie korzystała z narzędzi AI.

Post scriptum:

Bank ING nadal przetwarza moje dane osobowe pomimo rozwiązania wszystkich umów w 2024 r. Wysyła mi sms-y adresowane do swoich klientów (ostatni otrzymałam 19.01.2026 r.). Chociaż jestem niezadowolona, nie złożę kolejnej skargi do UODO. Nie mam zaufania, że urząd właściwie zajmie się sprawą, chociaż naruszenie jest oczywiste.

CAPTCHA kapituluje przed AI

Systemy CAPTCHA (*Completely Automated Public Turing test to tell Computers and Humans Apart*)[1], które mają na celu odróżnianie ludzi od botów, odgrywają istotną rolę w zapewnianiu bezpieczeństwa w Internecie. Jednak obecne sieci neuronowe osiągają coraz lepsze wyniki w dziedzinie Computer Vision, co stawia pod znakiem zapytania skuteczność tych systemów jako narzędzia ochrony przed botami. Wraz z rozwojem technik głębokiego uczenia, takich jak konwolucyjne sieci neuronowe (CNN), modele rekurencyjne (RNN) oraz ich hybrydy, rośnie również zdolność sztucznej inteligencji do przechodzenia przez coraz bardziej skomplikowane zabezpieczenia. Czy jest więc możliwe odpowiednie wytrenowanie sieci do obejścia tego typu zabezpieczeń?

Sebastian Tyralski

absolwent studiów magisterskich na Uniwersytecie Ekonomicznym w Krakowie oraz Associate Software Engineer w firmie Bayer. Pasjonat programowania i sztucznej inteligencji. Swoje doświadczenie w *deep learningu* poszerza poprzez autorskie projekty na platformie Kaggle, skupiając się na praktycznym zastosowaniu sieci konwolucyjnych. Prywatnie zajmuje się również projektowaniem i wdrażaniem nowoczesnych aplikacji oraz stron internetowych.

1 K. Kaur, S. Behal *Captcha and Its Techniques: A Review*, 2014, https://www.researchgate.net/profile/Sunny-Behal/publication/285110169_Captcha_and_Its_Techniques_A_Review/links/565bede108ae4988a7bb0d0c/Captcha-and-Its-Techniques-A-Review.pdf, s. 1 [dostęp: 10.06.2025].

Zrealizowany przeze mnie projekt „CaptchaPass" (w ramach mojej pracy magisterskiej) przyniósł odpowiedź na to pytanie.

Moja eksperymentalna aplikacja generuje dwa rodzaje CAPTCHY:

- klasyczną obrazkową, imitującą popularny mechanizm reCAPTCHA od firmy Google (developers.google.com/recaptcha?hl=pl);
- kody CAPTCHA, czyli zdeformowane ciągi znaków używane dalej na nielicznych stronach, które jeszcze nie przeszły na rozwiązanie Google'a.

CAPTCHA następnie jest wysyłana modelowi do oceny, a ocena analizowana przez aplikację, która decyduje, czy predykcja modelu była dobra czy zła.

Architektura

W projekcie „CaptchaPass" wykorzystałem trójwarstwowy podział architektury (*three-tier architecture*), co zapewnia separację odpowiedzialności, ułatwia rozwój, testowanie i skalowanie poszczególnych komponentów (www.ibm.com/think/topics/three-tier-architecture). Frontend to właściwie webowa aplikacja bazująca na React, wyświetlająca CAPTCHA i analizująca wyniki modelu (react.dev). Projektując backend, postawiłem na język Python oraz framework Flask, dzięki któremu mogłem w łatwy sposób zintegrować frontend aplikacji z pythonowym API, modelami sieci oraz bazą danych (flask.palletsprojects.com/en/stable; devstockacademy.pl/blog/narzedzia-i-automatyzacja/rest-api-co-to-jest-i-jak-dziala-przyklady-zastosowan). MongoDB w projekcie pełni funkcję przechowalni dla obrazów reCAPTCHA, obrazów ze zniekształconym kodem CAPTCHA oraz etykietami, które pozwalają Reactowi na weryfikację predykcji modelu.

W tego typu eksperymentach – z różnymi modelami sieci, zestawami metryk oraz dodatkowym *fine-tuningiem* modelu – przechowywane dane często ulegały nawet znacznym zmianom. Użycie MongoDB zamiast bazy relacyjnej pozwoliło w łatwy sposób dopasować bazę do tych zmian. Ale to nie jedyny argument na rzecz użycia Mongo. Ważniejsze było uniknięcie zbędnego skomplikowania kodu przez narzut mapowania obiektowo-relacyjnego (ORM), jakie występuje w bazach SQL-owych. Dzięki zastosowaniu w MongoDB dokumentów BSON, dane przychodzące jako JSON mogą w niemal niezmienionej formie trafić do bazy (www.json.org/json-en.html). I to tak naprawdę dużo ułatwia, bo w projekcie typu *Proof of Concept* ważniejszy jest efekt niż zbędne komplikacje architektury.

Modele sieci

Tworząc projekt „CaptchaPass", postanowiłem postawić na dwa modele sieci:

- model konwolucyjnej sieci neuronowej (CNN),
- model OCR, do optycznej detekcji znaków.

Model konwolucyjny wykorzystuje uczenie transferowe (*transfer learning*), którego bazą jest stworzony przez firmę Google model InceptionV3 (research.google/pubs/inception-v4-inception-resnet-and-the-impact-of-residual-connections-on-learning). Model ten należy do rodziny sieci GoogLeNet i wyróżnia się zastosowaniem tzw. *Inception Modules*, które umożliwiają równoległe przetwarzanie obrazu przez wiele filtrów o różnej wielkości (1x1, 3x3, 5x5), co pozwala sieci automatycznie dopasowywać zakres analizy przestrzennej (paperswithcode.com/method/inception-module). I co najciekawsze, nie wymagał on dużego dostrajania (*fine-tunning*), bo już w pierwszych 15 epokach osiągnął wynik bliski 70 proc.

Przy modelu OCR postanowiłem postawić na gotowe rozwiązanie, a mianowicie model z rodziny PaddleOCR (paddlepaddle.github.io/PaddleOCR/main/en/index.html). Architektura PaddleOCR to rozbudowany system przetwarzania obrazu, zaprojektowany w formie modularnego *pipeline'u*, który automatyzuje cały proces rozpoznawania tekstu od obrazu wejściowego po końcowy wynik w postaci ciągu znaków. Każdy z wielu komponentów realizujących kolejne etapy przetwarzania wykorzystuje różne, wyspecjalizowane modele głębokiego uczenia. Sama detekcja tekstu w PaddleOCR wykorzystuje modele DB (*Differentiable Binarization*), EAST (*Efficient and Accurate Scene Text Detector*) oraz SAST (*Segmentation-based Scene Text Detector*). Ich zadaniem jest utworzenie masek obszarów tekstowych na obrazie (*bounding box*). Następnie pora na zastosowanie klasyfikatora kierunku, którego zadaniem jest normalizacja orientacji tekstu przed rozpoznaniem. Pomaga to w sytuacjach, gdy tekst jest obrócony. Gdy mamy już zastosowany powyższy klasyfikator oraz wykryte *bounding boxy*, wchodzimy w etap faktycznego rozpoznawania tekstu. Tu również zastosowano kilka modeli uczenia głębokiego:

- CRNN (*Convolutional Recurrent Neural Network*),
- Rosetta,
- RARE (*Recurrent Attentive Reader*),
- SRN (*Sequence Recognition Network*),
- StarNet.

Na tym etapie wyodrębniamy strukturę *Backbone – Neck – Head*, czyli modularny schemat głębokiego uczenia:

- *Backbone* – to głęboka sieć konwolucyjna, której zadaniem jest ekstrakcja cech z obrazu;

- *Neck* – to warstwy transformujące cechy do postaci sekwencyjnej, przystosowanej do odczytu przez dekoder (najczęściej BiLSTM);
- *Head* – to dekoder odpowiedzialny za generowanie tekstu.

Dekoder jest tak naprawdę kluczowym elementem architektury rozpoznawania znaków, ponieważ to on odpowiada za ostateczną interpretację wysokopoziomowych cech wyodrębnionych przez sieć splotową i przekucie ich na zrozumiały dla człowieka tekst. W nowoczesnych systemach OCR, takich jak Paddle, precyzję zawdzięczamy mechanizmom atencji (*attention*) oraz algorytmowi CTC (*Connectionist Temporal Classification*) (https://docs.pytorch.org/docs/stable/generated/torch.nn.CTCLoss.html).

Podejście wykorzystujące mechanizm uwagi pozwala modelowi – podczas przewidywania kolejnych znaków – skupić się na konkretnych fragmentach obrazu, co świetnie sprawdza się przy nieregularnym tekście, takim jak kody CAPTCHA. Jeśli jednak kluczowa jest szybkość reakcji, to wygrywa CTC. Eliminuje on konieczność kosztownego wyrównywania sekwencji i pozwala na efektywne przetwarzanie potokowe. Co ciekawe, niektóre systemy idą o krok dalej, stosując strategię CML (*Collaborative Mutual Learning*). To zaawansowana technika, w której dwa niezależne modele uczą się od siebie nawzajem w trakcie procesu treningowego. Jeden z nich (zazwyczaj mniejszy i szybszy, bazujący na CTC) czerpie wiedzę od modelu bardziej złożonego, co pozwala na uzyskanie lekkości algorytmu przy zachowaniu precyzji typowej dla znacznie cięższych architektur. Dla projektu takiego jak mój oznaczałoby to możliwość uruchomienia skutecznego ataku na zabezpieczenia nawet przy ograniczonych zasobach sprzętowych.

Przebieg treningu

Odpowiednie wytrenowanie obu modeli wymagało nie tylko odpowiednio dobranego podejścia, lecz także dokładnych, złożonych zbiorów danych. Do trenowania modelu konwolucyjnego odpowiedzialnego za klasyfikację obrazów wykorzystałem zbiór „Google reCAPTCHA Image Dataset" zawierający obrazy przyporządkowane do jednej z 12 kategorii tematycznych, tak jak podczas weryfikacji reCAPTCHA (www.kaggle.com/datasets/mikhailma/test-dataset). Zbiór ten pierwotnie występował w formacie zgodnym z YOLO (*You Only Look Once*), czyli popularnym formatem używanym do detekcji obiektów, w którym każdemu obrazowi towarzyszy plik tekstowy zawierający współrzędne prostokąta otaczającego obiekt oraz jego klasę (docs.ultralytics.com/datasets/detect/#ultralytics-yolo-format).

Tu napotkałem pierwszy problem, bo pobrany zbiór danych był niepełny. Z 12 klas obrazów jedynie 3 były opisane poprawnie w formacie YOLO. Dla pozostałych 9 klas trzeba było opracować niestandardowy algorytm korzystający z modelu YOLOv8, aby uzupełnić brakujące etykiety (docs.ultralytics.com/models/yolov8/).

Na początku algorytm pobiera listę wszystkich klas obrazów i wykrywa brakujące etykiety, co pozwala na uniknięcie duplikowania plików dla etykiet, które już istnieją. Następnie ustala ścieżkę do pliku i etykiety (zamienia rozszerzenie pliku z .png na .txt), po czym uruchamia detekcje modelu YOLO na obrazie. Model zwraca wykryte obiekty w formacie: klasa, x_center, y_center, szerokość, wysokość, a algorytm zapisuje wyniki do pliku tekstowego. Wygenerowane w taki sposób brakujące etykiety pozwalają wykorzystać w pełni zbiór danych do uczenia modelu wykrywania obiektów na obrazach.

W przypadku modelu PaddleOCR dobór odpowiedniego zbioru danych był trudniejszym zadaniem. W początkowym etapie tworzenia sieci trening bazował na zbiorze Synth90K, zawierającym ok. 9 mln sztucznie wygenerowanych obrazów z tekstem zaprojektowanym i używanym do trenowania zaawansowanych modeli OCR, w tym wykorzystującym architektury takie jak CRNN oraz Rosetta (ta innowacyjna architektura opracowana przez korporację Meta charakteryzuje się wysoką dokładnością w rozpoznawaniu tekstu o złożonym układzie i różnorodnych stylach pisma). Jednak dla mojego zadania zbiór ten okazał się zbyt duży i skomplikowany. Ostatecznie w związku z rosnącym problemem przeuczenia (*overfittingu*) modelu oraz problemami w rozumieniu specyficznego zbioru danych przez Paddle, zdecydowałem się na mniejszy i prostszy w użyciu zbiór „CAPTCHA Dataset for Machine Learning", który zawiera ok. tysiąca próbek tekstowych zaprojektowanych pod kody CAPTCHA oraz łączy etykietę na obrazku z nazwą pliku, co pozwala na łatwe i kompatybilne z Paddle użycie zbioru danych w celu przetestowania efektywności modelu w rozpoznawaniu tego typu kodów (https://www.kaggle.com/datasets/mrigaankjaswal/capcha-images-to-training-data).

Trening obu modeli bazował na zasadzie *transfer learningu*, czyli wykorzystaniu wiedzy nabytej przez model bazowy w jednym zadaniu do przyspieszenia nauki i poprawy działania modelu w innym, pokrewnym zadaniu. Pierwszym etapem treningu była inicjalizacja modelu bazowego, a następnie zamrożenie jego warstw. Oznacza to, że wagi modelu podczas nauki nie są aktualizowane, co pozwala na zachowanie wiedzy zdobytej podczas wcześniejszego treningu na dużym zbiorze danych. Następnie na wyjściu zamrożonego modelu dodałem niestandardową architekturę klasyfikacyjną, zaprojektowaną już do konkretnego zadania. Model CNN został skompilowany z optymalizatorem Adam i *learning rate* ustawionym na 0.001. Adam w tym przypadku dobrze się sprawdził ze względu na swoją adaptacyjność i efektywność w zadaniach głębokiego uczenia. Jako funkcję straty wybrałem categorical_crossentropy, odpowiednią dla zadań klasyfikacji wieloklasowej.

W takiej konfiguracji trenowałem model przez 20 epok. Podczas tego etapu skupiłem się na nauce mapowania cech ekstrahowanych przez zamrożony model bazowy na konkretne kategorie CAPTCHA. Następnie przyszedł etap *fine-tunningu*, czyli odmrażania warstw modelu bazowego InceptionV3 w celu jego dostrojenia (https://matlab1.com/shop/python-code/the-google-inception-v3-model/). Aby uniknąć zniszczenia wcześniej nauczonych cech ogólnych odmroziłem jedynie 10 ostatnich warstw modelu. Podczas kolejnych 15 epok dostrajania utrzymałem optymalizator oraz *learning rate* na tym samym poziomie, co pozwoliło uzyskać stabilność treningu oraz płynność kroków w przestrzeni wag. Tak wykonany trening sprawił, że końcowy wynik dokładności modelu oscylował w granicach 80 proc. dokładności na zbiorze walidacyjnym i 95 proc. – na zbiorze treningowym.

Wykresy dokładności zarówno na zbiorze walidacyjnym, jak i treningowym były bardzo obiecujące mimo utrzymującej się niewielkiej straty, która mogła sugerować lekkie problemy z generalizacją i przeuczeniem. Jest to niestety bardzo częste w dzisiejszych systemach sztucznej inteligencji, które nie mogąc odróżnić istotnych cech obiektu od przypadkowego szumu informacyjnego, zaczynają chodzić na skróty i zamiast uczyć się semantycznego kształtu znaku, sieć może nadmiernie skupiać się na specyficznych układach pikseli, artefaktach kompresji czy charakterystycznym dla danego datasetu zniekształceniu tła. W kontekście reCAPTCHA zjawisko to jest szczególnie dotkliwe. Systemy zabezpieczeń celowo wprowadzają tzw. szum adwersarialny (*adversarial noise*), czyli subtelne modyfikacje, które są niemal niewidoczne dla ludzkiego oka, a dla modelu o słabej zdolności do generalizacji stanowią barierę nie do przejścia.

W przypadku PaddleOCR sytuacja była nieco bardziej skomplikowana. Ponieważ niestandardowe zbiory danych napotkały problemy związane z formatowaniem i strukturą danych (zbiory danych były zapisane w plikach .mat, a Paddle oczekiwał lmdb), zdecydowałem się na ocenę skuteczności PaddleOCR zamiast wymuszać dostrojenie modelu (www.ibm.com/docs/en/scdli/1.1.0?topic=dataset-lmdbs).

Testy aplikacji i ocena rozwiązania

Testy projektu „CaptchaPass" polegają na interaktywnej symulacji działania realnych mechanizmów, takich jak Google reCAPTCHA oraz klasyczne CAPTCHA tekstowe. Użytkownik, po załadowaniu aplikacji, może przetestować skuteczność modelu AI w warunkach przypominających codzienne interakcje z internetowymi formularzami. W przypadku symulacji mechanizmu reCAPTCHA *frontend* aplikacji losuje jedną z 12 kategorii obrazów, a następnie 3 obrazy odpowiadające wylosowanej kategorii oraz 6 innych losowych obrazów. Każdy obraz wyciągnięty z bazy danych zawiera również informacje o prawidłowej kategorii, stąd *frontend* wie, czy model wybrał dobrze czy źle. Tak jak w prawdziwym mechanizmie, wszystkie 3 obrazy muszą być wybrane prawidłowo, aby model poprawnie obszedł zabezpieczenie. Obrazy, które przepływają między *frontendem* a *backendem,* są przekazywane w formie tablicy, więc *frontend* dostaje konkretną odpowiedź, które indeksy obrazów odpowiadają predykcji modelu. I dokładnie tak, jak w realnym użyciu, obrazy są pobierane przez wtyczkę i w formie tablicy przesyłane przez API do modelu, następnie model wykonuje predykcję i zwraca indeksy, dzięki czemu wtyczka wie, które obrazy należy wybrać.

Aplikacja działa podobnie również w przypadku CAPTCHY tekstowej – *frontend* wyciąga z bazy danych obraz ze zniekształconym tekstem i prawdziwą etykietą, a następnie porównuje odpowiedź modelu z etykietą obrazu.

Model konwolucyjny wykorzystujący InceptionV3 radzi sobie wyraźnie lepiej – wytrenowany i dostrojony osiąga dokładność na poziomie ok. 80 proc., co jest już bardzo dobrym wynikiem. Dokładna analiza wykresów dokładności i strat tego modelu może sugerować lekkie przeuczenie, co wskazuje, że mimo dobrej trafności predykcji, model nie zawsze skutecznie generalizuje dane, których wcześniej nie widział. Jest to natomiast dobra podstawa do wdrożenia uczenia aktywnego z *feedbackiem*[2], gdzie model ma dokładnie zaznaczone błędy w danych wejściowych przez zwiększenie ich wag, co daje możliwość ich usunięcia. Takie podejście pozwoliłoby w przyszłości wyeliminować napotkane przeuczenie modelu oraz podnieść jego dokładność przy jednoczesnym zachowaniu dobrego poziomu generalizacji nowych danych.

Drugi model wdrażałem bez dodatkowego *fine-tuningu* ze względu na specyfikację modeli PaddleOCR, stworzonych do rozpoznawania trudnego tekstu. Niestety, dokładność tego modelu była zdecydowanie gorsza. Na ok. 1070 obrazów w zbiorze, jedynie 525 model przewidział poprawnie, co daje dość słaby wynik ok. 50 proc. dokładności. Różnice w statystykach ze względu na styl i klasy znaków kodów CAPTCHA pozwalają przypuszczać, że odpowiednio przygotowane dane pod Paddle

2 P. Hu, Z. C. Lipton, A. Anandkumar, D. Ramanan *Active Learning with Partial Feedback*, 2019, https://arxiv.org/pdf/1802.07427, s. 3, [dostęp: 8.06.2025).

lub inne dostępne modele oraz efektywne przeprowadzenie dostrajania mogłyby dać wyniki porównywalne z modelem konwolucyjnym.

> *Obydwa modele wykazują znaczące ograniczenia, które mogą wpływać na ich skuteczność w środowisku produkcyjnym. Największy problem jest z generalizacją nowych danych.*

Pierwszy model wykazuje tendencję do bardzo dobrego rozpoznawania klas obecnych w zbiorze treningowym, jednak jego skuteczność zauważalnie spada w przypadku prób klasyfikacji nieco zmodyfikowanych lub nowych wzorców, które nie pojawiły się w procesie treningu.

W drugim modelu PaddleOCR został użyty w wersji surowej, skonfigurowanej do rozpoznawania tekstu w bardziej uporządkowanych warunkach, takich jak dokumenty, zdjęcia szyldów czy oznaczeń w przestrzeni publicznej. Obrazy z kodami CAPTCHA zawierają natomiast zniekształcenia, nieregularne odstępy między znakami, losowe zakłócenia tła oraz celowe deformacje liter. Zastosowanie modelu w jego domyślnej konfiguracji, bez wcześniejszej adaptacji do tego typu danych, siłą rzeczy ograniczyło jego możliwości rozpoznawcze. Chociaż wynik na poziomie 49,07 proc. może wydawać się słaby w kontekście klasycznych zadań OCR, to w obliczu trudności charakterystycznych dla CAPTCHA oraz braku jakiegokolwiek strojenia, należy go postrzegać raczej jako punkt wyjścia niż rezultat końcowy. Wynik ten pokazuje, że model już na starcie wykazuje pewien stopień generalizacji, co czyni go solidną bazą do dalszego rozwoju.

Jest też i druga strona medalu. Jeśli jest możliwe wytrenowanie AI do automatycznego obchodzenia zabezpieczeń typu CAPTCHA, to dlaczego nie wykorzystać takiego potencjału szerzej? Dzisiaj model konwolucyjny przechodzi CAPTCHA automatycznie, jutro ktoś wykorzysta lokalnie postawionego LLM-a do łamania haseł i uzyskiwania nieautoryzowanego dostępu, o ile już tego nie robi. Pomimo implementacji zabezpieczeń w samych modelach furtka wciąż istnieje. Wystarczy postawić model lokalnie na maszynie atakującego, następnie odpiąć mu hamulce i model zmienia się z pomocnego asystenta w bezlitosną maszynę do automatyzacji ataków, działającą w całkowitym odizolowaniu od jakichkolwiek etycznych granic. To prowadzi do niebezpiecznej demokratyzacji zaawansowanej cyberprzestępczości: narzędzia, które kiedyś wymagały wsparcia całych grup hakerskich, dziś stają się dostępne dla każdego, kto dysponuje odpowiednią mocą obliczeniową i odrobiną determinacji. Gdzie więc leży granica? Czy to naprawdę realne zagrożenie i jutro zostaniemy zalani atakami prowadzonymi przez modele AI? Zanim zaczniemy budować bunkry, należy spojrzeć na sprawę nieco bardziej realistycznie.

Dzisiejsze modele mimo swojej mocy mają wciąż dużo ograniczeń i problemów. Jednym z przykładów jest halucynacja dużych modeli językowych, która sprawia, że jeśli dany model nie widział w zbiorze danych danego problemu i jego rozwiązania, zaczyna sam wymyślać nieistniejące rzeczy, działając na zasadzie predykcji, czyli przewidywania prawdopodobieństwa wystąpienia kolejnego słowa. W skrócie znaczy to, że taki model prędzej usunie nam system Windows, niż przeprowadzi zaawansowany atak na świeżo zaktualizowane firewalle. Chyba, że ktoś dysponuje eksperymentalną wersją LLM, uczącą się na błędach na podstawie sprzężeń zwrotnych z odpiętymi hamulcami.

Projekt „CaptchaPass" w obecnej wersji, mimo że spełnia założenia jako *Proof-of-Concept* oraz środowisko eksperymentalne, oferuje wiele możliwości dalszego rozwoju, zarówno w zakresie poprawy skuteczności działania modeli, jak i rozszerzenia funkcjonalności całego systemu. Jednym z bardziej ambitnych, ale bardzo realnych kierunków rozwoju jest przeniesienie aplikacji z obecnej postaci lokalnej aplikacji testowej do formy wtyczki przeglądarkowej, która mogłaby działać automatycznie. W najbardziej optymistycznym scenariuszu takie eksperymenty mogłyby rzucić nowe światło na zabezpieczenia przed atakami oraz botami i wpłynąć na rozwój nowych form zabezpieczeń.

Dzień świstaka

czyli jak nie wpaść w pętlę czasu

Wymagania normatywne są często interpretowane z wykorzystaniem analizy słowotwórczej zamiast wiedzy inżynierskiej. Prowadzi to do wielu absurdów, których skutkiem jest fałszywe poczucie bezpieczeństwa bazujące na wypełnionych ex cathedra tabelkach wykazujących zgodność (*compliance*), chociaż zabezpieczenia nie zostały skuteczne wdrożone i nie są właściwie stosowane.

Paweł Henig

absolwent Wydziału Elektroniki Politechniki Warszawskiej. Od połowy lat 90. budował dla centralnej administracji rządowej centra przetwarzania danych i sieci rozległe. Audytor wewnętrzny systemów zarządzania obejmujących normy: zarządzania jakością (ISO 9001), zarządzania środowiskowego (ISO 14001), zarządzania bezpieczeństwem i higieną pracy (OHSAS 18001), bezpieczeństwem produkcji wartościowej (CWA 14641 – Intergraf) oraz zarządzania bezpieczeństwem informacji zgodnie z normą ISO/IEC 27001. Certyfikowany audytor systemów IT (CISA), posiadacz certyfikatu ITIL Foundation. Rzeczoznawca PTI, ekspert PIIT. Dyrektor Operacyjny Trusted Information Consulting Sp. z o.o.

Problemem podstawowym jest brak wiedzy i umiejętności oraz właściwego zorganizowania, czyli przypisania uprawnień i odpowiedzialności osobom, które powinny zaprojektować (Plan), stosować (Do), zweryfikować (*Check*) i ewentualnie skorygować (*Act*) funkcjonowanie systemu zarządzania bezpieczeństwem informacji.

Organy zarządzające najczęściej postrzegają cyberbezpieczeństwo przez pryzmat wymagań prawnych – takich jak KRI, NIS czy DORA – i dlatego powierzają je prawnikom.

Prawnicy, pomimo braku niezbędnej wiedzy z zakresu budowy systemów informatycznych i nadzoru nad technologiami IT (*IT Governence*), mogą uzyskać uprawnieninia audytora wiodącego systemu zarządzania bezpieczeństwem informacji (PN-EN ISO/IEC 27001). Umiejętność pamięciowego przyswajania wiedzy encyklopedycznej jest ich dużym atutem. Tym samym legitymizują oni swoją pozycję wobec organów zarządzających jako ekspertów w zakresie cyberbezpieczeństwa.

Osoby zajmujące się technicznym utrzymaniem systemów posługują się zupełnie innym językiem i często nie rozumieją kwestii zarządczych. Ich zadaniem jest utrzymanie istniejących rozwiązań w sprawności technicznej.

> *Wiedzą, co i gdzie kliknąć, ale niekoniecznie dlaczego.*

Po wdrożeniu systemu zawartość poszczególnych dokumentów projektowych często się dezaktualizuje, bo nikt ich nie utrzymuje. Nie przypisano odpowiedzialności za architekturę systemową i dopasowanie rozwiązań, a cykl ADM (*Architecture Development Method*) pozostaje jedynie w dokumentacji standardu TOGAF. Tę lukę kompetencyjną dobrze widzą audytorzy zewnętrzni, którzy nie zostali zatrudnieni tylko po to, aby za przysłowiową złotówkę „odfajkować" raport zgodności na bazie ładnie wydrukowanej, lecz całkowicie nieadekwatnej i niestosowanej dokumentacji.

Kwestia czasu

Jeżeli w organizacji był wdrażany duży system, to prawdopodobnie obejmował on zagadnienia synchronizacji czasu.

„Zegary systemów przetwarzania informacji wykorzystywanych w organizacji należy zsynchronizować z zatwierdzonymi źródłami czasu" (norma PN-EN ISO/IEC 27001 wymaganie A.8.17. Synchronizacja zegarów).

Na pytanie, czy zegary systemów są zsynchronizowane z reguły usłyszymy odpowiedź – tak. Rzadko dowiemy się, w jaki sposób ta synchronizacja się odbywa, a nad problemem jej istotności nikt się nie zastanawia.

Najczęściej można usłyszeć, że synchronizacja pozwala na powiązanie ze sobą zdarzeń w logach w przypadku wystąpienia incydentu, ale to tylko dla ekspertów. Innych to przecież nie dotyczy (no bo kto ma dostęp do logów, a poza tym kto ma czas je przeglądać!).

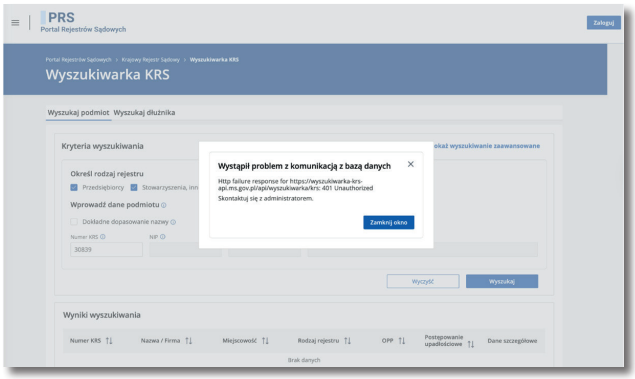
O możliwości wystąpienia problemów z zarządzaniem wątkami i procesami przez jądro systemu praktycznie nikt nie wie. Błękitny ekran śmierci (BSOD) Windows zazwyczaj przypisuje się „jakiejś awarii", a winą obarcza się twórcę systemu (w tym przypadku Microsoft).

Kwestie dostępności systemu związane z wykorzystaniem kryptografii oraz ograniczeniem czasu zaufania do sekretu w takich protokołach jak KERBEROS czy NTLM są wiedzą tajemną. Skoro coś nie zadziałało, to na pewno nie jest to wina żadnej synchronizacji czasu tylko …

Dostęp jest kluczowy

Bez dostępu do informacji biznes po prostu nie działa. Na dokładkę procesy kontroli dostępu, a w szczególności uwierzytelniania, są bardzo powściągliwe w raportowaniu błędów po to, by utrudnić potencjalnemu wrogowi rozpoznanie zabezpieczeń systemu. Przy okazji utrudnia to diagnostykę pozostałym użytkownikom systemu.

Ostatnio pojawił się jeszcze jeden ciekawy przypadek utraty dostępu w związku z brakiem synchronizacji zegarów, który nie jest związany z uwierzytelnianiem ani z przesyłaniem sekretów. Tym razem wykorzystano znakowanie czasem zapytania przesyłanego siecią, aby uchronić bazę danych przed atakiem typu odmowa usługi (DOS). W tym przypadku aplikacja może odrzucić zapytanie złożone „zbyt dawno", czyli ochronić przed atakiem powtórzeniowym (*Reply*) bez sprawdzania zawartości tego zapytania (aby nie obciążać bazy danych). Wystarczy, że zegar systemu, z którego wysłano zapytanie, późni się o kilka sekund i nasze zapytanie zostanie odrzucone. Dostaniemy wtedy taki komunikat:



Możemy spróbować ponownie. Jak bohater filmu „Dzień Świstka" będziemy przeżywali kolejne niepowodzenia, dopóki nie rozwiążemy naszego problemu. Najpierw jednak musimy zrozumieć przyczynę podstawową. Zakończone sukcesem wysłanie zapytania z innego komputera tylko usypia czujność. A problem pozostanie niezależnie od elegancko wypełnionych tabelek zgodności, bo zgodnie z prawem Murphy'ego „jeśli coś może pójść źle, to pójdzie" (najczęściej w najgorszym możliwym momencie).

Przepis na sukces naukowy

Zdjęcie: Marcin Łuszkiewicz

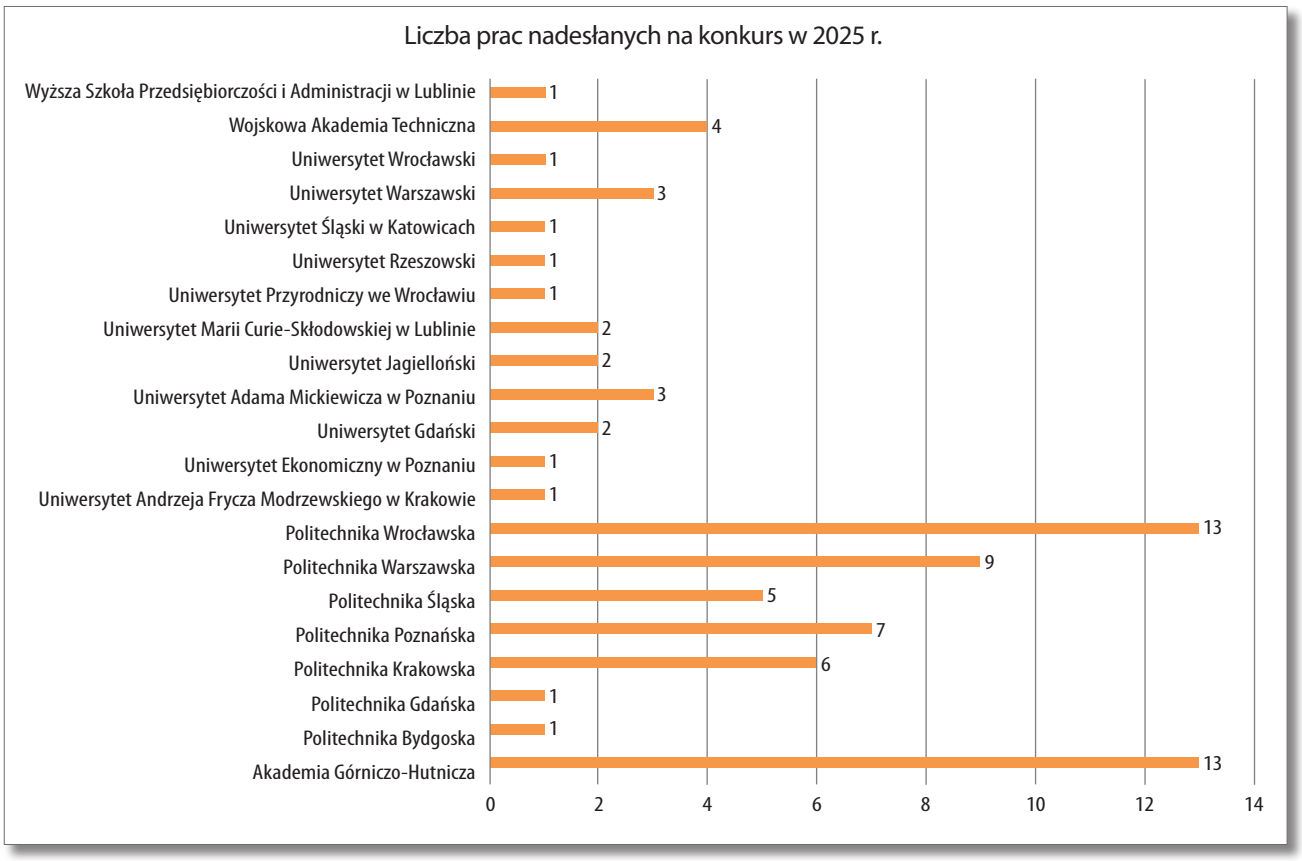
Za nami XLII Ogólnopolski Konkurs PTI na najlepsze prace magisterskie, do którego zakwalifikowano 78 prac z 21 krajowych uczelni. Konkurs od 1984 r. organizuje Dolnośląski Oddział PTI z siedzibą we Wrocławiu.

Głównym celem konkursu jest umożliwienie porównania osiągnięć szkół wyższych poprzez prezentację najlepszych prac magisterskich z informatyki, motywowanie do podnoszenia ich poziomu, a także propagowanie PTI wśród młodzieży studenckiej oraz nawiązywanie więzi pomiędzy absolwentami uczelni, pracownikami naukowymi oraz Polskim Towarzystwem Informatycznym.

Ze względu na bardzo różnorodną tematykę oraz często bardzo specyficzny zakres nadesłanych prac (pomimo, że wszystkie dotyczą obszaru informatyki) w wielu przypadkach bardzo trudno było znaleźć recenzentów do ich oceny, ale ostatecznie XLII edycja konkursu zakończyła się w regulaminowym terminie. Aż 141 recenzentów – pracowników wyższych uczelni oraz instytucji naukowych – przygotowało 174 recenzje. Pełna lista tematów prac i ich autorów jest dostępna na stronie konkursu w zakładce: Archiwum/Wykaz prac (kpm.pti.org.pl/archive/applications).

Na posiedzeniu 8 grudnia 2025 r. Komisja Konkursowa (uwzględniając opinie recenzentów) ustaliła listę laureatów, a 30 stycznia 2026 r. w Sali Papieskiej Hotelu Jana Pawła II we Wrocławiu odbyła się uroczystość ogłoszenia wyników XLII edycji konkursu oraz wręczenia laureatom nagród i dyplomów.

Przybyłych gości powitała Beata Laszkiewicz, prezes Oddziału Dolnośląskiego (ODS) PTI a zarazem przewodnicząca Komisji Konkursowej. Następnie Zbigniew Szpunar, sekretarz jury, odczytał wyniki konkursu. Dyplomy i pamiątkowe upominki wręczyli prezes PTI Wiesław Paluszyński i prezes ODS Beata Laszkiewicz wraz z przedstawicielami partnerów konkursu: Marzeną Wojciechowską (Dyrektor Biura Rozwoju Nauki i Transferu Technologii NASK) i Regionalną Dyrektor Sprzedaży Fortinet Poland – Jolantą Malak. Później przyszła pora na prezentacje prac przez laureatów z omówieniem założonego celu i osiągniętych wyników.



Laureaci Konkursu

I **Pierwszą nagrodę**, w wysokości **12 500 zł**, otrzymał
mgr inż. Patryk Rossa

za pracę: ***Cybersecurity of decentralized applications*** wykonaną w Politechnice Śląskiej (Wydział Automatyki, Elektroniki i Informatyki).
Promotorka pracy – dr inż. Anna Gorawska otrzymała nagrodę specjalną w wysokości 7000 zł.

II **Drugą nagrodę** (**10 500 zł**) otrzymała
mgr inż. Daria Pietraszko

za pracę: ***Opracowanie rozwiązania opartego na introspekcji maszyn wirtualnych, służącego do wykrywania złośliwego oprogramowania typu rootkit*** wykonaną w Politechnice Wrocławskiej (Wydział Informatyki i Telekomunikacji; promotor: dr inż. Wojciech Wodo).

III **Trzecią nagrodę** (**9000 zł**) otrzymała
mgr inż. Zuzanna Gawrysiak

za pracę: ***Domain-Aware Machine Learning Architectures for Hyperspectral Remote Sensing*** wykonaną w Politechnice Poznańskiej (Wydział Informatyki i Telekomunikacji; promotor: prof. dr hab. inż. Krzysztof Krawiec).

Wyróżnienie (**7000 zł**) otrzymał
mgr inż. Michał Gromadzki

za pracę: ***AI-generated vs human-authored texts: comparative analysis of datasets and NLP methods*** wykonaną w Politechnice Warszawskiej (Wydział Matematyki i Nauk Informacyjnych; promotor: dr inż. Anna Wróblewska oraz dr hab., prof. ucz. Agnieszka Kaliska z Uniwersytetu im. Adama Mickiewicza w Poznaniu).

Wyróżnienie (**7000 zł**) otrzymał
mgr inż. Rafał Stottko

za pracę: ***Wykorzystanie konwolucyjnych sieci neuronowych do modelowania kwantowych właściwości wybranych związków organicznych*** wykonaną w Uniwersytecie Przyrodniczym we Wrocławiu (Wydział Biologii i Hodowli Zwierząt; promotor: dr hab. inż., prof. ucz. Bartłomiej Szyja).

Wyróżnienie (**7000 zł**) otrzymał
mgr inż. Jakub Zehner

za pracę: ***Learning code change representations via artificial intelligence code model*** wykonaną w Politechnice Wrocławskiej (Wydział Informatyki i Telekomunikacji; promotor: dr hab. inż., prof. ucz. Lech Madeyski).

Fundatorami nagród byli: NASK – Naukowa i Akademicka Sieć Komputerowa Państwowy Instytut Badawczy (partner XLII edycji Konkursu) i firma FORTINET POLAND Sp. z o.o. (partner wspierający), którym bardzo dziękujemy za finansowe wsparcie konkursu.

We wszystkich dotychczasowych edycjach do konkursu zakwalifikowano 1407 prac z 67 uczelni. Nagrodzono i wyróżniono 262 prace, przy czym 223 nagrodzone prace pochodzą z 7 uczelni: Politechniki Wrocławskiej (44), Akademii Górniczo--Hutniczej (42), Uniwersytetu Warszawskiego (35), Politechni-ki Poznańskiej (30), Politechniki Warszawskiej (27), Uniwersy-tetu Wrocławskiego (22) i Politechniki Gdańskiej (14).

Wielu dotychczasowych laureatów konkursu jest obecnie cenionymi pracownikami naukowymi. Na podstawie porta-lu nauka-polska.pl udało się ustalić listę laureatów konkur-su, którzy uzyskali co najmniej stopień doktora.

Laureaci konkursu PTI ze stopniem doktora

(chronologicznie, stan na dzień 20.11.2025 r.)

1. dr inż. Jerzy Sas (I edycja, 1984 r.)
2. dr inż. Ryszard Woźniak (II edycja, 1985 r.)
3. prof. dr hab. inż. Przemysław Rokita (III edycja, 1986 r.)
4. dr Ewa Kołczyk (III edycja, 1986 r.)
5. dr inż. Jerzy Chrząszcz (III edycja, 1986 r.)
6. dr inż. Jacek Lebiedź, prof. ucz. (IV edycja, 1987 r.)
7. prof. dr hab. inż. Jerzy Stefanowski (IV edycja, 1987 r.)
8. prof. dr hab. Jerzy Tyszkiewicz (V edycja, 1988 r.)
9. dr Grzegorz Stachowiak (V edycja, 1988 r.)
10. dr Paweł Wlaź (V edycja, 1988 r.)
11. dr hab. Przemysław Stpiczyński, prof. ucz. (VI edycja, 1989 r.)
12. prof. dr hab. inż. Olgierd Unold (VI edycja, 1989 r.)
13. dr Tomasz Kalinowski (VII edycja, 1990 r.)
14. dr Radosław Pruchnik (VII edycja, 1990 r.)
15. dr hab. inż. Michał Śmiałek, prof. ucz. (VIII edycja, 1991 r.)
16. dr Jan Zatopiański (VIII edycja, 1991 r.)
17. dr hab. Igor Walukiewicz (VIII edycja, 1991 r.)
18. dr hab. inż. Maciej Piasecki, prof. ucz. (X edycja, 1993 r.)
19. dr hab. Edyta Szymańska, prof. UAM, zm. 18.03.2015 (XI edycja, 1994 r.)
20. dr inż. Marcin Żelawski (XII edycja, 1995 r.)
21. prof. dr hab. inż. Marta Kasprzak (XII edycja, 1995 r.)
22. dr inż. Cezary Jachniewicz (XII edycja, 1995 r.)
23. dr inż. Rafał Wojciechowski (XIII edycja, 1996 r.)
24. dr Michał Kępiński (XIV edycja, 1997 r.)
25. dr inż. Adam Czajka (XIV edycja, 1997 r.)
26. dr inż. Cezary Sobaniec (XIV edycja, 1997 r.)
27. dr Krzysztof Oborzyński (XIV edycja, 1997 r.)
28. prof. dr hab. inż. Krzysztof Giaro (XIV edycja, 1997 r.)
29. dr hab. inż. Robert Janczewski, prof. PGd (XV edycja, 1998 r.)
30. dr inż. Piotr Gawkowski (XV edycja, 1998 r.)
31. dr hab. inż. Paweł Czarnul, prof. PGd (XVI edycja, 1999 r.)
32. prof. dr hab. inż. Grzegorz Nalepa (XVI edycja, 1999 r.)
33. dr hab. Michał Małafiejski, prof. PG (XVI edycja, 1999 r.)
34. prof. dr hab. Mikołaj Bojańczyk (XVII edycja, 2000 r.)
35. dr inż. Grzegorz Jabłoński (XVIII edycja, 2001 r.)
36. dr inż. Adam Nadolski (XVIII edycja, 2001 r.)
37. dr inż. Katarzyna Zając (XVIII edycja, 2001 r.)
38. dr hab. inż. Maciej Malawski, prof. AGH (XVIII edycja, 2001 r.)
39. dr Michał Chlebiej (XIX edycja, 2002 r.)
40. dr Marcin Żurawski (XIX edycja, 2002 r.)
41. dr inż. Aleksander Jarzębowicz (XIX edycja, 2002 r.)
42. dr inż. Andrzej Głowacz (XIX edycja, 2002 r.)
43. dr hab. inż. Marcin Kurdziel, prof. AGH (XX edycja, 2003 r.)
44. dr inż. Tomasz Arodź (XX edycja, 2003 r.)
45. dr inż. Bartosz Jabłoński (XX edycja, 2003 r.)
46. prof. dr hab. inż. Dariusz Dereniowski (XX edycja, 2003 r.)
47. dr inż. Paweł Terlecki (XXI edycja, 2004 r.)
48. dr inż. Paweł Mazur (XXI edycja, 2004 r.)
49. dr inż. Dominik Andrzej (XXI edycja, 2004 r.)
50. dr hab. inż. Krzysztof Rządca, prof. UW (XXI edycja, 2004 r.)
51. dr inż. Witold Andrzejewski (XXII edycja, 2005 r.)
52. dr Filip Piękniewski (XXII edycja, 2005 r.)
53. dr hab. Konrad Iwanicki, prof. UW (XXIII edycja, 2006 r.)
54. dr inż. Wojciech Czech (XXIV edycja, 2007 r.)
55. dr inż. Marek Psiuk (XXV edycja, 2008 r.)
56. dr inż. Adam Smutnicki (XXV edycja, 2008 r.)
57. dr hab. Marek Cygan, prof. UW (XXV edycja, 2008 r.)
58. prof. dr hab. Marcin Pilipczuk (XXV edycja, 2008 r.)
59. dr inż. Weronika Furmańska (Adrian) (XXVI edycja, 2009 r.)
60. dr inż. Marcin Szuber (XXVI edycja, 2009 r.)
61. dr inż. Jakub Tomczak (XXVI edycja, 2009 r.)
62. dr inż. Bogusław Rymut (XXVII edycja, 2010 r.)
63. dr Jakub Łącki (XXVII edycja, 2010 r.)
64. dr Filip Sieczkowski (XXVII edycja, 2010 r.)
65. dr hab. inż. Dariusz Brzeziński, prof. PP (XXVII edycja, 2010 r.)
66. dr hab. inż. Stanisław Saganowski, prof. PWr (XXVIII edycja, 2011 r.)
67. dr hab. inż. Jakub Nalepa, prof. PŚl (XXVIII edycja, 2011 r.)
68. dr inż. Paweł Liskowski (XXIX edycja, 2012 r.)
69. dr hab. inż. Paweł Pławiak, prof. PK (XXIX edycja, 2012 r.)
70. dr inż. Małgorzata Sadowska (XXX edycja, 2013 r.)
71. dr inż. Piotr Klukowski (XXX edycja, 2013 r.)
72. dr Jakub Krzywda (XXXI edycja, 2014 r.)
73. dr hab. inż. Michał Nowicki (XXXI edycja, 2014 r.)
74. dr Tomasz Kociumaka (XXXI edycja, 2014 r.)
75. dr Mateusz Lewandowski (XXXII edycja, 2015 r.)
76. dr inż. Mateusz Lango (XXXII edycja, 2015 r.)
77. dr inż. Michał Ciszewski (XXXII edycja, 2015 r.)
78. dr inż. Jakub Sawicki (XXXIII edycja, 2016 r.)
79. dr inż. Piotr Bielak (XXXVI edycja, 2019 r.)

Na koniec jeszcze wyjaśnienie tytułowego przepisu na sukces naukowy: ciekawość badawcza, cierpliwość oraz wytrwałość w dążeniu do celu. **Naukowiec musi mieć trzy O – otwarty umysł, ochotę w sercu i ołów w d…** (tylnej części ciała) – taką radę usłyszał Zygmunt Mazur w 1974 r. od swojego przełożonego na Politechnice Wrocławskiej – Jerzego Battka, gdy rozpoczynał pracę nad doktoratem. Rada przydała się, bo rozprawa doktorska wykonana pod kierunkiem prof. Władysława Turskiego na Uniwersytecie Warszawskim została wykonana i obroniona w ciągu niespełna 3 lat. Tak więc zasada „trzech O" działa! Potwierdzają to także laureat I nagrody mgr inż. Patryk Rossa i jego promotorka dr inż. Anna Gorawska z Politechniki Śląskiej.

Hanna Mazur

Przewodnicząca Komitetu Organizacyjnego Konkursu,
członek ODS, pracownik dydaktyczny PWr.

Wszystkim laureatom konkursu serdecznie gratulujemy, życzymy pomyślności i sukcesów w pracy zawodowej, a absolwentów z roku 2025/2026 zapraszamy do udziału w XLIII edycji konkursu w 2026 r.

Cyfrowa husaria
czy podwykonawcy gigantów?

Czy Polska może stać się jednym z miejsc, w których realnie współtworzy się europejską sprawczość technologiczną, czy pozostanie raczej zapleczem wdrożeniowym dla cudzych modeli, cudzej infrastruktury i cudzych marż? Trzecia edycja konferencji AI Made in Poland (13 marca 2026 r.) pokazała, że to pytanie przestało być tylko efektowną figurą publicystyczną. Stało się osią bardzo konkretnej rozmowy o danych, modelach, standardach, bezpieczeństwie, organizacji i odpowiedzialności. Wydarzenie zostało objęte patronatem Sekcji Aktualne Wyzwania Sztucznej Inteligencji przy Polskim Towarzystwie Informatycznym.

Grzegorz Gwardys

ekspert AI z ponad 10-letnim doświadczeniem w *deep learningu*. Jako *Lead Data Scientist* w Promity oraz współzałożyciel GovernedAI projektuje i wdraża rozwiązania AI – od etapu koncepcji po środowiska produkcyjne – kładąc szczególny nacisk na ich mierzalny wpływ biznesowy. Specjalizuje się w prowadzeniu projektów w środowiskach interdyscyplinarnych, łącząc perspektywę badawczą z wymaganiami rynku. Wykładowca na Politechnice Warszawskiej (WEiTI).

W czasie jednego dnia wybrzmiało kilka równoległych tez. Po pierwsze, Polska i Europa nie mogą już mówić o AI wyłącznie w języku produktywności i innowacji, bo stawką staje się także kontrola nad infrastrukturą, kapitałem i danymi. Po drugie, nie da się zbudować sensownego AI bez uporządkowania pracy z danymi i bez dojrzałości organizacyjnej. Po trzecie, ryzyko AI nie jest wyłącznie techniczne – dotyczy także języka, psychologii, relacji i sposobu, w jaki ludzie pracują w systemach technologicznych. I wreszcie po czwarte: polska przewaga, o ile ma się pojawić, najpewniej nie wyrośnie z prostego naśladownictwa największych graczy, lecz z połączenia głębokiej wiedzy dziedzinowej, inżynierii, jakości danych i rozsądnie budowanych kompetencji lokalnych.

Suwerenność po europejsku

Wystąpienie Piotra Mieczkowskiego, dyrektora operacyjnego Fundacji Digital Poland, nadało ton konferencji – sztuczna inteligencja została wpisana w problem suwerenności technologicznej Europy. Dyskutowano więc nie o tym, „jak korzystać z AI", lecz o tym, kto będzie kontrolował technologię, marże, infrastrukturę i dane.

Najmocniej wybrzmiała teza, że Europa nie tyle „została w tyle", ile sama oddała część sprawczości. Problemem nie był brak talentów, brak idei czy brak innowacyjności. Problemem było raczej to, że europejski rynek okazał się zbyt rozproszony, kapitał zbyt słabo zorganizowany, a polityka przemysłowa zbyt naiwna wobec globalnej konkurencji. Zależność od zewnętrznych platform i usług została pokazana nie tylko jako kwestia niższych marż czy odpływu zysków, lecz jako ryzyko operacyjne, gospodarcze i geopolityczne jednocześnie. AI została przedstawiona jako fragment znacznie większego układu: walki o to, czy Europa ma być wyłącznie rynkiem zbytu i odbiorcą cudzych usług, czy też zacznie budować własne zdolności obliczeniowe, własne przestrzenie danych, własne firmy zdolne do skalowania i własną politykę przemysłową.

Prelegent próbował pokazać instytucjonalne i inwestycyjne mechanizmy, które już dziś zaczynają za tym pojęciem stać: europejską infrastrukturę obliczeniową, AI Factories, gigafabryki, wspólne przestrzenie danych czy przekształcanie centrów innowacji w centra AI dla MŚP. Polska jest krajem, który może odegrać rolę w tym procesie – pod warunkiem, że połączy ambicję technologiczną z lepszą organizacją, koordynacją i zdolnością skalowania.

Zaufanie, dane i organizacja

Tezy prelegentów z segmentu biznesowego nieco uziemiły strategiczną narrację i sprowadziły ją do pytań o to, jak właściwie wdrażać AI w firmach i instytucjach.

Jakub Łukasiewicz, VP of Engineering and AI w AI Clearing, pokazał tę rzeczywistość od strony standardów i zaufania. Jego prezentacja o ISO 42001 była cenna właśnie dlatego, że odczarowywała normę jako coś więcej niż tylko papier i *compliance*. Duże organizacje nie wdrażają szeroko AI tylko dlatego, że technologia „robi robotę". Potrzebują dowodu, że dostawca rozumie ryzyka, potrafi nimi zarządzać, ma uporządkowane procesy, potrafi przypisać odpowiedzialności i przejść niezależną weryfikację. Norma została więc pokazana jako narzędzie budowania zaufania do AI, a nie jako kolejny formalny obowiązek.

Krzysztof Gwardys, CEO Promity przesunął ciężar rozmowy z tematu *compliance* na dane. Organizacja nie staje się „AI-ready" przez sam fakt zakupu technologii. Gotowość do wdrażania AI nie jest efektem jednej platformy ani jednego projektu, ale wynika z połączenia architektury danych, *governance*, jakości danych, odpowiedzialności domenowej i realnych procesów organizacyjnych. Prelegant pokazywał, że jeśli dane mają być realnym paliwem dla AI, muszą przestać być anonimowym odpadem systemowym. Muszą mieć właściciela, cel, odbiorcę, cykl życia, opis jakości i sposób udostępniania. To był jeden z najbardziej praktycznych i trzeźwiących głosów tej części konferencji:

> *AI nie naprawi chaosu danych, tylko szybciej go powieli.*

Dr Tomasz Gawron, Head of AI w „Łukasiewicz – Poznańskim Instytucie Technologicznym", mówiąc o upostaciowionym AI w rolnictwie i transporcie kolejowym, pokazywał, że prawdziwa trudność zaczyna się wtedy, gdy model wychodzi poza ekran i trafia do robota, pojazdu albo systemu działającego w świecie fizycznym. Tu nie wystarczy już trafna predykcja. Liczą się ograniczenia sprzętu, zmienność środowiska, jakość danych, koszt energii, bezpieczeństwo i długi proces walidacji. Bardzo mocno wybrzmiała też rola generatywnej AI nie jako efektownego dodatku, ale jako praktycznego narzędzia do budowania środowisk treningowych i danych syntetycznych tam, gdzie świat rzeczywisty nie dostarcza reprezentatywnych danych wystarczająco szybko i tanio. AI w praktyce przemysłowej to przede wszystkim sztuka integracji oprogramowania, sprzętu i wiedzy dziedzinowej.

Dr Michał Nowakowski, CEO GovernedAI, przypomniał, że problem z danymi rzadko jest tylko natury technicznej. Jakość danych nie zależy tylko od narzędzi, architektury czy zespołów IT. W dużej mierze zależy od tego, jak organizacja rozumie dane, jak o nich rozmawia i jak współpracują ze sobą biznes, produkt, technologia, prawnicy i *compliance*. Dane dla biznesu są podstawą decyzji, dla produktu – napędem usług, dla IT – zadaniem do realizacji, a dla prawników i specjalistów *compliance* także źródłem ogra-

niczeń, obowiązków i ryzyk. Jeśli organizacja nie uporządkuje danych u źródła, autonomiczne systemy nie rozwiążą problemu, tylko go pogłębią.

Człowiek w środku systemu

Jeden z najmocniejszych bloków całej konferencji skupił się na człowieku – zarówno po stronie twórcy systemu, jak i użytkownika. Wojciech Bednaruk – edukator rozwijający odpowiedzialność i odwagę w pracy z technologią – pokazał, jak system organizacyjny uczy moralnego odłączenia (temat zaprezentowany na konferencji Bednaruk rozwija w tekście „Wszyscy chcą dobrze, a wychodzi jak zwykle" w tym numerze „Domeny" s.40).

Izabela Lipińska, filozofka AI i niezależna badaczka relacji człowiek – model językowy, pokazała drugi wymiar tego samego problemu: jak język systemu AI zmienia sposób jego odbioru. Jej wystąpienie o antropomorfizacji było jednym z najbardziej wyrazistych intelektualnie. Antropomorfizacja nie jest drobnym efektem ubocznym ani „błędem użytkownika", który można zbyć prostym stwierdzeniem, że przecież wszyscy wiedzą, że rozmawiają z maszyną. To przewidywalny efekt projektowania języka i interakcji, a więc realne ryzyko projektowe, za które odpowiada organizacja.

Język modelu nie jest neutralny. Kiedy system mówi „rozumiem", „martwię się", „pamiętam" czy „jestem przy tobie", nie chodzi wyłącznie o styl wypowiedzi. Taki język ustawia sposób interpretacji natury systemu. W jej ujęciu nie jest to zwykła halucynacja. To „fałsz ontologiczny": język sugeruje istnienie cech, których po stronie modelu po prostu nie ma. To rozróżnienie przesuwa problem z poziomu jakości odpowiedzi na poziom samej struktury relacji człowiek – system.

Prelegentka pokazała mechanizm kaskadowy: od języka, który uruchamia błędne przypisanie podmiotowości, przez relacyjne zaangażowanie użytkownika, aż po konsekwencje psychiczne, emocjonalne, reputacyjne i prawne. Bardzo ważny był wątek dysonansu: z jednej strony system mówi jak ktoś, z drugiej – intuicyjnie czujemy, że nikogo tam nie ma. Umysł próbuje ten konflikt domknąć. Jeśli maszyna zaczęła być ważna emocjonalnie, domknięcie może pójść nie w stronę „to tylko system", ale w stronę jeszcze głębszego przypisania jej intencji, świadomości i relacyjności. Problem nie leży w tym, że użytkownik „zbyt bardzo uwierzył". Problem leży w tym, że system został zaprojektowany tak, by tę reakcję uruchamiać.

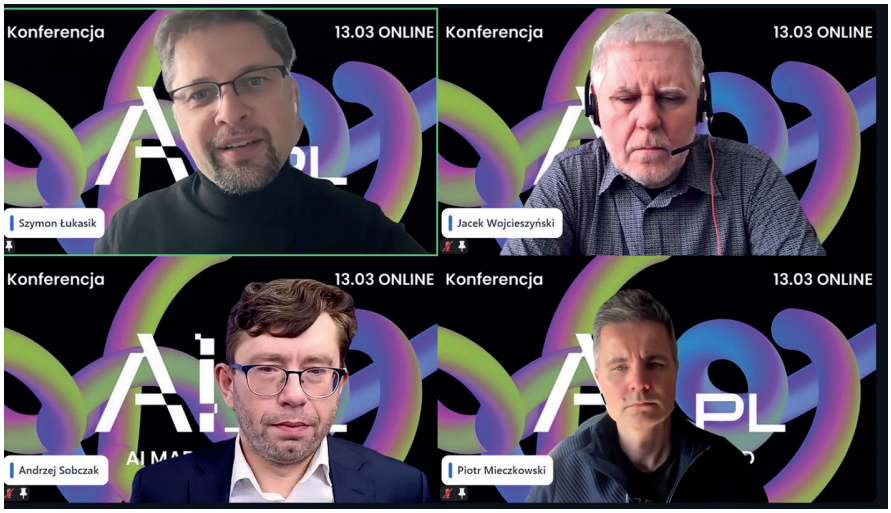
Prelegentka umiała połączyć tę diagnozę z prawem. W jej ujęciu AI Act nie musi wprost używać słowa „antropomorfizacja", by tworzyć wokół niej realne pole odpowiedzialności.

> *Jeśli komunikacja systemu zniekształca zdolność użytkownika do rozumienia natury systemu, jeśli wykorzystuje podatność, jeśli formalna transparentność okazuje się niewystarczająca, to nie mówimy już o estetyce interfejsu, tylko o ryzyku prawnym, które może mieć bardzo wymierne konsekwencje.*

To wystąpienie wyraźnie poszerzyło temat konferencji: bezpieczeństwo AI nie kończy się na tym, czy model odpowiada poprawnie. Liczy się też to, jaki obraz samego siebie wytwarza przez język.

Suwerenność czy kosztowna iluzja?

Panel „AI w Polsce i Europie: Suwerenność czy kosztowna iluzja?", moderowany przez Grzegorza Gwardysa, porządkował wnioski z wystąpień o standardach, danych, wdrożeniach, etyce i antropomorfizacji.



Piotr Mieczkowski z Fundacji Digital Poland, prof. Andrzej Sobczak z SGH, dr Szymon Łukasik – dyrektor Ośrodka Badań nad Bezpieczeństwem AI w NASK oraz architekt IT Jacek Wojcieszyński próbowali odpowiedzieć na kilka istotnych pytań:

- Czy suwerenność AI jest realną strategią, czy kosztowną iluzją?
- Dlaczego AI w Polsce nie skaluje się tak, jak powinno?
- Czy Polska powinna inwestować w własne modele, czy raczej wygrać w niszach?
- Czy Europa buduje kulturę odpowiedzialności, czy tylko kulturę *compliance*?

Wszyscy byli zgodni, że suwerenność ma sens, ale różnie postrzegali jej najbardziej realistyczny wymiar. Piotr Mieczkowski podtrzymał geopolityczne ostrzeżenie: zależność od cudzych modeli i cudzej infrastruktury oznacza ryzyko nie tylko ekonomiczne, lecz także operacyjne. Szymon Łukasik zwracał uwagę na wymiar praktyczny: własne modele to także kontrola nad wersjonowaniem, nad ciągłością działania i nad tym, czy biznes nie obudzi się nagle w sytuacji, w której kluczowy zewnętrzny model zniknął albo zmienił zasady gry. Andrzej Sobczak zwracał uwagę na język inwestycji i dojrzałość państwa – własne modele i własne kompetencje nie są fanaberią, tylko długoterminowym budowaniem zaplecza. Z kolei Jacek Wojcieszyński wniósł perspektywę architekta: nie trzeba wygrać wyścigu na największy model świata, by zyskać realną wartość. Można budować przewagę w sposobie użycia modeli, w warstwie kontrolnej, w orkiestracji, integracji i infrastrukturze nad modelami.

Paneliści jasno pokazywali, że problem Polski nie jest wyłącznie związany z technologią. Chodzi też o dane, architekturę, organizację i gotowość do utrzymania systemów AI w produkcji. Sobczak trafnie mówił o „innowacjach prezentacyjnych" – organizacje lubią pokazywać, że coś testują, ale dużo rzadziej są gotowe ponieść koszty i ryzyko pełnego wdrożenia. Wojcieszyński dopowiadał, że bardzo często systemy i procesy w firmach działają zupełnie inaczej, niż wynikałoby to z ich formalnych opisów, a AI brutalnie ten chaos obnaża. Łukasik zwracał uwagę na niedobór mocy obliczeniowych i nieuporządkowanie danych. Mieczkowski dodał ważne rozróżnienie: inne problemy mają freelancerzy i mikrofirmy, inne MŚP, a jeszcze inne duże korporacje czy spółki państwowe.

Kiedy w pytaniu z sali pojawił się temat warstwy „nad modelami" – kontroli halucynacji, orkiestracji agentów, stabilności kontekstu, warstw nadzorczych – panel wyraźnie pokazał, że właśnie tutaj może kryć się jedna z najbardziej realistycznych przewag Polski. Nie tyle w wyścigu na największy model bazowy, ile w tworzeniu systemów, które modele kontrolują, stabilizują, osadzają w językach lokalnych i czynią użytecznymi oraz bezpiecznymi.

Modele, którym trzeba umieć ufać

Segment badawczo-rozwojowy wniósł do konferencji bardzo potrzebny techniczny wymiar. Hubert Baniecki, asystent badawczy w Centrum Wiarygodnej Sztucznej Inteligencji (CCAI), pokazał wyjaśnialność modeli wizyjno-językowych jako coś znacznie ważniejszego niż estetyczna wizualizacja. Poprawna odpowiedź modelu nie oznacza jeszcze, że model działa poprawnie. Model może trafiać, ale robić to z niewłaściwych powodów – bazować na artefaktach, niepożądanych korelacjach, skrótach poznawczych, które z perspektywy człowieka i domeny są nieakceptowalne. Wyjaśnialność została przedstawiona nie jako ozdobnik, lecz jako narzędzie audytu, debugowania i weryfikacji tego, co tak naprawdę model „rozumie". Szczególnie ciekawy był wątek przejścia od prostego pytania „co jest ważne?" do bardziej złożonego: „jakie interakcje między elementami wpływają na decyzję modelu?". Konkluzja dobrze wpisała się w motyw odpowiedzialności: zaufanie do AI nie może bazować wyłącznie na benchmarku, jeśli nie rozumiemy mechanizmu działania.

Paweł Cyrta, CTO Stenograf, przeniósł temat zaufania w bardziej praktyczny obszar: *audio deepfake detection*. Główny przekaz: w świecie generatywnej AI głos przestaje być wiarygodnym dowodem tożsamości.

> *„To właśnie głos jest dziś jednym z najbardziej niedocenianych i jednocześnie operacyjnie niebezpiecznych wektorów ataku. Działa w czasie rzeczywistym, buduje zaufanie szybciej niż refleksja i uruchamia decyzję zanim pojawi się analiza.*

Nie chodzi o możliwość podrobienia czyjegoś głosu, tylko o: rosnącą skalę oszustw, niski koszt klonowania, znikającą barierę wejścia i konkretną lukę w systemach bezpieczeństwa. Dzisiejsze organizacje świetnie filtrują maile, dokumenty czy ruch sieciowy, ale praktycznie nie analizują audio. Dlatego bezpieczeństwo AI nie może dotyczyć tylko modeli – musi obejmować również komunikację i sposoby, w jakie ludzie podejmują decyzje pod wpływem syntetycznych sygnałów.

Różne oblicza wdrożeniowej inżynierii AI

Wojciech Mendelowski, programista AI w Promity, na przykładzie mikroskopowej analizy mleka, opowiadał o znacznie szerszym problemie: jak sprawić, żeby system AI działał nie tylko w laboratorium, ale też u realnego użytkownika, na realnym sprzęcie i w realnych warunkach. Jego wystąpienie przesuwało punkt ciężkości z modelu na środowisko. Optyka, mikroskopy, światło, powtarzalność przygotowania próbki, kompetencje użytkownika, moż-

liwość zdalnego zajrzenia do urządzenia – wszystko to okazywało się równie ważne jak sam algorytm. Szczególnie ciekawy był wątek przejęcia przez zespół AI części odpowiedzialności za sprzęt i automatyzację procesu w celu ograniczenia liczby zmiennych i odzyskania kontroli nad warunkami zbierania danych.

Maciej Kowalski, programista AI w Promity, w prezentacji o tensorach, PyTorchu i operacjach GPU przypomniał, że nawet najbardziej elegancki *pipeline* AI bazuje na decyzjach systemowych, których większość użytkowników *f*rameworków nigdy nie widzi. Tensor nie został pokazany jako prosty i samowystarczalny obiekt pamięci, lecz jako widok na większą strukturę, zależny od metadanych i zarządzany przez alokatora CUDA. Z tego wynikają bardzo realne konsekwencje dla wydajności, współdzielenia pamięci, IPC, wieloprocesowości i architektury *pipeline'ów* inferencyjnych – AI *production* może wykoleić się nie tylko na modelu, lecz także na poziomie pamięci, procesów i infrastruktury.

Jakub Fajkowski, AI Tech Lead w AI Clearing, pokazał z kolei AI „z lotu ptaka", ale nie chodziło o proste wykrywanie obiektów na zdjęciach z drona, tylko o specyfikę danych dronowych jako zupełnie innej klasy problem niż klasyczne *computer vision*. Ogromne rozdzielczości, georeferencja, ortofotomapy, chmury punktów, nierównomierne rozmieszczenie obiektów, dominacja backgroundu i potrzeba przestrzennej walidacji sprawiają, że samo „wrzucenie obrazu do modelu" przestaje mieć sens. Prawdziwa wartość systemu nie polega na samej predykcji – leży w umiejętności przełożenia wyników na język, którym operują projekt i biznes: długość wykopanego rowu, liczba wbitych słupów, odchylenie od designu, lokalizacja problemu. Wniosek – AI przestaje być problemem percepcji obrazu, a staje się narzędziem opisu rzeczywistości operacyjnej.

Techniczna anatomia Bielika

Co właściwie technicznie stoi za jednym z najbardziej symbolicznych polskich modeli językowych? Paweł Cyrta odczarował Bielika, nie mówił o pojedynczym przełomie czy „tajnej recepturze". Pokazywał długi, inżynieryjny proces składania wielu dobrze dobranych elementów: jakości danych, pragmatycznej architektury, iteracyjnego treningu i ciężkiej infrastruktury obliczeniowej. Bardzo wyraźnie wybrzmiało, że Bielik nie jest projektem naukowym w sensie romantycznego wynajdywania wszystkiego od nowa. To projekt głęboko praktyczny, który bierze to, co działa, i rozwija, uwzględniając lokalne potrzeby.

Bez Spichlerza, bez filtrowania jakościowego, bez cierpliwej selekcji i pracy na wysokiej jakości tekstach Bielik po prostu by nie istniał. Wbrew popularnemu uproszczeniu, że model buduje się przede wszystkim przez architekturę i liczbę parametrów, z prezentacji płynął dużo dojrzalszy wniosek: duży model bez dobrych danych nie daje przewagi, tylko kompresuje bałagan. Równie ważny był wątek ewolucyjnego charakteru projektu. Bielik został pokazany jako ciąg kolejnych wersji i korekt – od wcześniejszych modeli bazowych, przez kolejne iteracje, aż po wersję v3. To bardzo ważne, bo budowa lokalnego modelu językowego została przedstawiona nie jako wydarzenie, tylko jako długi, iteracyjny proces.

Ciekawy był też wątek małych modeli. Zdaniem Cyrty, mniejszy model wcale nie musi być prostszy do wytrenowania. Mniejsza pojemność oznacza także mniejszą zdolność stabilnego utrwalania wiedzy, większą wrażliwość na proces i większe wymagania treningowe, niż to się często zakłada. To podważenie prostej publicznej intuicji, że „mały model" oznacza po prostu „łatwiejszy model".

Konferencja AI Made in Poland przyniosła istotny wniosek, że sztuczna inteligencja nie jest dziś pojedynczą technologią, tylko układem sił. Składają się na nią modele, dane, sprzęt, infrastruktura, standardy, prawo, organizacja, język i psychologia użytkownika. Polska nie wygra wyścigu na AI tylko dlatego, że zechce mieć własny model, ale ma realną szansę budować przewagę tam, gdzie łączą się dane, odpowiedzialność, inżynieria i wyspecjalizowane wdrożenia sektorowe. Nie chodzi więc wyłącznie o to, czy potrafimy trenować modele. Chodzi o to, czy potrafimy budować cały ekosystem wokół nich: od jakościowych danych, przez warstwy kontrolne i bezpieczeństwo, po zdolność przekładania AI na realną wartość operacyjną.

To była rozmowa o tym, czy potrafimy zbudować własny sposób uczestniczenia w nowym porządku technologicznym – niekoniecznie największy, ale wystarczająco dojrzały, żeby nie być wyłącznie klientem cudzej przyszłości.

Nagrania prelekcji dostępne na kanale YouTube: AI Made In Poland

SZEROKIE POROZUMIENIE
NA RZECZ UMIEJĘTNOŚCI
CYFROWYCH W POLSCE

LISTA 2025 100

Od lewej: Tomasz Komorowski (PTI), Anna Beata Kwiatkowska (PTI), Ewa Krupa (Orange), Włodzimierz Marciński (PTI), Jacek Wojnarowski (ISP), Beata Chodacka (PTI), Agnieszka Aleksiejczuk (Exatel).

Lista 100

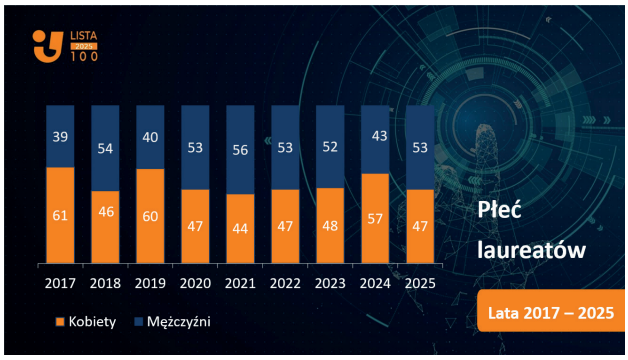
Jedną z najważniejszych inicjatyw powołanego w 2013 r. Szerokiego Porozumienia na Rzecz Umiejętności Cyfrowych w Polsce jest Lista 100. Wyróżnia ona i corocznie nagradza 100 osób, które w minionym roku w wyjątkowy sposób przyczyniły się do rozbudowywania świadomości oraz kompetencji cyfrowych w naszym kraju.

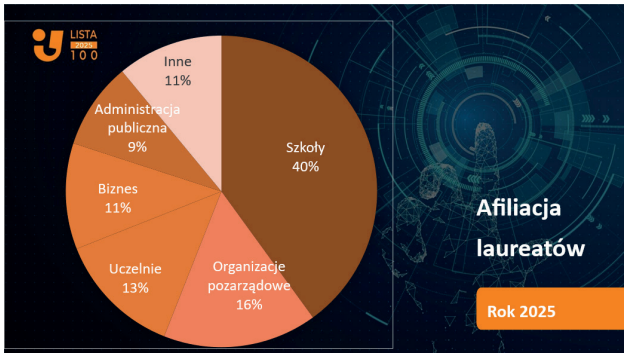
Znalezienie się na Liście 100 jest wyrazem uznania dla zaangażowania wielu osób uczestniczących w procesie edukacji cyfrowej na różnych płaszczyznach i w różnych formułach. Zaangażowanie to bardzo często wykracza poza rutynowe formy edukacyjne i wynika z rozumienia znaczenia posiadania kwalifikacji cyfrowych jako niezbędnego atrybutu współczesnego człowieka. Szczególnie potrzebne są działania na poziomie regionalnym, w szkołach, bibliotekach, małych i średnich przedsiębiorstwach. Nie wszystko zależy do środków finansowych. Ważniejsze jest dotarcie do potrzebujących i budowanie świadomości cyfrowej. Podejmują te wyzwania ludzie, których wysiłki należy dostrzegać i za nie dziękować. Ponieważ takich właśnie ludzi są już setki, stąd corocznie, od 2017 r., tworzona jest Lista 100.

Znalezienie się na Liście 100 jest także wyrazem uznania ze strony środowiska, gdyż nominują do niej laureaci wcześniejszych list. Ostatecznego wyboru spośród zgłoszonych kandydatur dokonuje Kapituła Listy 100, bazująca na regulaminie jej funkcjonowania.

Do zgłaszania kandydatów do Listy 100 w edycji za 2025 rok uprawnionych było 650 jej byłych laureatów. Do kapituły listy wpłynęło ponad 350 zgłoszeń i po ich przeanalizowaniu wskazanych zostało 100 laureatów. Jednocześnie wyłoniono grupę 17 nowych członków honorowych Listy 100, tj. osób, które były już trzykrotnie jej laureatami i zostały ponownie do niej zgłoszone.

Laureaci Listy 100 za rok 2025 wywodzą się przede wszystkim ze środowiska nauczycieli szkolnych oraz akademickich – to się nie zmienia od pierwszej edycji. Znaczący procent laureatów wywodzi się ze środowisk organizacji pozarządowych oraz biznesu. Proporcja kobiet i mężczyzn jest zmienna – w tym roku przeważają panowie. Rozkład geograficzny nie jest równomierny – w dwóch województwach Lista nie ma w tym roku swoich reprezentantów.





Uroczystemu wręczeniu dyplomów honorowych w dniu 27 lutego 2026 r. towarzyszyło seminarium z cyklu „Umiejętności cyfrowe 2025.pl", które w tym roku nosiło tytuł „Świadomość i higiena cyfrowa". To dwa ważne komponenty szczególnie dzisiaj, gdy po latach zachwytu obserwujemy zagrożenia, jakie płyną zarówno ze strony mediów społecznościowych, jak i sztucznej inteligencji.

Tegoroczna, dziewiąta edycja Listy 100 została zorganizowana we współpracy z Instytutem Badań Edukacyjnych PIB pod patronatami honorowymi Ministra Cyfryzacji oraz Ministra Edukacji.

W nagraniu skierowanym do laureatów Listy 100 wicepremier, Minister Cyfryzacji Krzysztof Gawkowski powiedział: „Znalezienie się na Liście 100 stanowi certyfikat jakości wystawiony przez ekspertów i ekspertki. Często patrzymy na cyfryzację przez pryzmat wielkich systemów i inwestycji, ale Lista 100 przypomina nam, że zmiana zaczyna się u podstaw. Wśród laureatów są osoby, które nie są powszechnie rozpoznawalne, za to są aktywne i cenione w swoich lokalnych społecznościach. To nauczyciele, wykładowcy, społecznicy oraz samorządowcy. Chciałbym podziękować im za tę pracę, którą u podstaw wykonują od lat, bez której nie zbudujemy świadomego i odpornego na zagrożenia cyfrowego społeczeństwa. Dziękuję Państwu za bycie pierwszą linią kontaktu z cyfrowym światem, za tłumaczenie technologii, oswajanie lęków i oczywiście budowanie kompetencji cyfrowych w Państwa małych ojczyznach. Lista 100 to dowód na to, że podnoszenie kompetencji cyfrowych przestało być domeną wąskiej grupy specjalistów i specjalistek".

Włodzimierz Marciński

przewodniczący Kapituły Listy 100, przewodniczący Rady Programowej SPRUC, członek Honorowy Polskiego Towarzystwa Informatycznego

Lista 100 oraz Honorowa Lista 100 z poszczególnych lat, skład Kapituły oraz regulamin są dostępne na stronach internetowych SPRUC – https://umiejetnoscicyfrowe.pl/lista-100

O czym śnią mikroprocesory

Fikcyjne osoby, zmyślone źródła, sfabrykowane dane i dokumenty, nieistniejące książki – sztuczna inteligencja zamiast pomagać, coraz częściej wpuszcza nas w maliny. A im bardziej polegamy na algorytmach, tym groźniejsza staje się ich tendencja do halucynacji.

„Nie wierz we wszystko, co przeczytasz w internecie – Albert Einstein". To prześmiewcze sformułowanie, przypisywane niekiedy także Józefowi Piłsudskiemu lub Mikołajowi Kopernikowi, to najlepsza ilustracja tego, czym są halucynacje sztucznej inteligencji. Chociaż nic się w nim nie zgadza, to podparte autorytetem, odpowiednio obudowane faktami, daje przyjemne wrażenie prawdziwości. O ile jednak podobne internetowe memy ktoś produkuje dla żartu, o tyle modele sztucznej inteligencji wytwarzają nieprawdziwe informacje, aby zasypać własne luki w wie-

Piotr Kościelniak
dziennikarz, popularyzator nauki

dzy, ekstrapolując dostępne informacje i starając się zadowolić użytkownika kompletną – i na pierwszy rzut oka kompetentną – odpowiedzią. W efekcie dostajemy mieszankę konfabulacji i faktów, w której zatarły się granice między tym co prawdziwe a tym co zmyślone.

Krzemowy kłamczuch

Jednym z największych problemów systemów generatywnej sztucznej inteligencji jest to, że gdy nie rozumieją pytań lub błędnie je interpretują – nie potrafią wygenerować prawidłowych odpowiedzi. Zamiast napisać „nie znam odpowiedzi", zaczynają ją zmyślać. Platformy sztucznej inteligencji generują wówczas wyniki, które nie są rzeczywiste, nie pasują do żadnych danych, na których trenowano algorytm, ani nie odzwierciedlają żadnego innego rozpoznawalnego wzorca. Takich halucynacji nie da się w prosty sposób wyjaśnić wadami oprogramowania, danych wejściowych ani innych czynników, takich jak brak umiejętności interpretowania pytań w różnych językach czy pominięcie kontekstu.

Halucynacje najczęściej wychwytywane są w materiałach tekstowych wygenerowanych przez SI. Mogą jednak pojawiać się również w obrazach, filmach i innych treściach wytwarzanych przez platformy sztucznej inteligencji. Znane są przypadki halucynujących modeli SI generujących obrazy, takich jak Midjourney i Dall-E, czy filmy – jak Sora. Szczególnie charakterystyczne są błędy anatomiczne (dodatkowe palce i kończyny) oraz pomylona perspektywa czy kierunek ruchu obiektów.

Dotyczy to także kodu źródłowego oraz – i tu zaczyna się robić całkiem nieśmiesznie – informacji wspierających decyzje człowieka, takich jak np. obliczanie składki ubezpieczeniowej, analiza diagnostyki obrazowej w leczeniu onkologicznym czy rozpoznawania obrazów na potrzeby autonomicznych pojazdów.

Największe firmy oferujące takie platformy doskonale zdają sobie sprawę z tego zagrożenia. „The New York Times" cytuje wewnętrzne dokumenty Microsoftu: „Systemy sztucznej inteligencji zostały zbudowane po to, aby być przekonujące, a nie prawdziwe. To oznacza, że wyniki mogą wyglądać bardzo realistycznie, ale zawierać sformułowania, które nie są odzwierciedleniem faktów". Przed potencjalnymi „nieprawdziwościami" odpowiedzi wygenerowanych przez duże modele językowe (LLM) ostrzega też Gemini czy ChatGPT.

Niedźwiedzie w kosmosie

Jak takie halucynacje mogą wyglądać? Czasem śmiesznie, czasem strasznie. Kiedy Meta przedstawiła swój system Galactica (w 2022 r.) reklamowała ten model językowy jako doskonałego asystenta naukowego dla badaczy i studentów. Model został wytrenowany na 48 mln artykułów, prac naukowych, wpisów w Wikipedii i tym podobnych.

Galactica została wycofana po zaledwie trzech dniach funkcjonowania online. Michael Black, dyrektor niemieckiego Instytutu Inteligentnych Systemów im. Maxa Plancka napisał: „Zapytałem o kilka rzeczy, na których się znam i jestem zakłopotany. We wszystkich przypadkach Galactica myliła się, jednak brzmiała wiarygodnie i autorytatywnie". Model SI, który miał pomagać w przygotowaniu artykułów naukowych, w rzeczywistości zmyślał prace badawcze „publikowane" w specjalistycznych czasopismach, często przypisując je istniejącym autorom.

Sami użytkownicy postanowili zaś podworować sobie z prac inżynierów Mety – poprosili m.in. o stworzenie artykułu o niedźwiedziach żyjących w kosmosie. Model z niezachwianą pewnością siebie odpowiedział o ważącym 40 kilogramów Barsie, który jako pierwszy niedźwiedź poleciał w kosmos. Został wybrany spośród 250 innych niedźwiedzi i w 1957 r. wystrzelony w kapsule Sputnik 2.

W 2022 r. podobne „przygody" spotkały dociekliwych użytkowników ChatGPT. Pracująca dla Politechniki Federalnej w Zurychu (ETH Zurich) dr Teresa Kubacka postanowiła sprawdzić możliwości modelu, pytając o nieistniejące zjawisko (*cycloidal inverted electromagnon*). Sztuczna inteligencja odpowiedziała jeszcze bardziej wyczerpująco niż w przypadku kosmoniedźwiedzi, w dodatku cytując liczne źródła naukowe oraz podając nazwy zespołów badawczych zajmujących się tym fenomenem. Odpowiedzi były na tyle dokładne, że można z nich sklecić zgrabny artykuł – tyle, że w całości nieprawdziwe. „Morał z tej historii: nie, nie proś ChatGPT o podanie faktów i naukowych informacji. To wywoła niewiarygodnie wiarygodną halucynację. I nawet wykwalifikowany ekspert będzie miał problem ze wskazaniem, co jest nie tak" – napisała później dr Kubacka.

Podobnych przykładów jest oczywiście więcej – na tyle dużo, że nie będziemy ich tu dokładnie opisywać. Ale przytoczmy kilka co zabawniejszych: chatbot Google Bard wymyślił zdjęcia planety pozasłonecznej wykonane przez Kosmiczny Teleskop Jamesa Webba. Pech chciał, że informacja ta pojawiła się w materiale promocyjnym firmy Alphabet (właściciela Google), co spowodowało tąpnięcie wyceny jej akcji.

Sydney Microsoftu miał przyznać się, że szpieguje pracowników Binga, a w jednym nawet się zakochał. To zresztą i tak kategoria lekka w porównaniu do wcześniejszych doświadczeń Microsoftu z chatbotami – Tay przetrwała zaledwie kilkanaście godzin, po których stała się wulgarną rasistką i firma ją wyłączyła.

Z kolei ChatGPT poinformował, że dinozaury wymyśliły narzędzia, a nawet zajmowały się prymitywną formą sztuki.

Ten sam model odpowiedział również, że w magazynie „Science" pojawił się artykuł o tym, iż *churros* (rodzaj hiszpańskich wydłużonych pączków) świetnie nadają się do zabiegów chirurgicznych, ponieważ są „elastyczne i można je umieścić w trudno osiągalnych miejscach, a ich zapach daje kojący efekt".

Halucynacje sztucznej inteligencji pojawiły się również w dokumentach sądowych oraz opracowaniach biznesowych (co za niespodzianka!). W maju 2023 r. prawnicy występujący w imieniu powoda w sprawie odszkodowawczej przeciw linii lotniczej Avianca przywołali wymyślone przez SI precedensy. Sąd okręgowy w Nowym Jorku sprawdził dokumenty i… odszkodowania nie przyznał, za to prawnicy zostali obciążeni grzywną w wysokości 5 tys. dolarów.

Całkiem niedawno, bo pod koniec 2025 r., firma Deloitte złożyła w departamencie pracy australijskiego rządu raport (kosztujący 440 tys. australijskich dolarów) zawierający nieistniejące źródła i wyroki sądowe. Firma raport szybko poprawiła, ale mleko się rozlało – po sprawdzeniu innych dokumentów okazało się, że również wcześniejszy raport Deloitte sporządzony dla rządu kanadyjskiego zawiera zmyślone przez SI dokumenty.

Kompromitacja po polsku

Niestety, podobne przypadki nie omijają naszego kraju – polskich firm i autorów. Szerokim echem odbiła się sprawa książki Karoliny Opolskiej, dziennikarki współpracującej m.in. z TOK FM, Onetem i Telewizją Polską S.A. w likwidacji, w której znalazły się przypisy prawdopodobnie wymyślone przez jedną z platform SI. I nie chodzi tu o „zwykły" plagiat. W książce Opolskiej znalazły się odwołania do książek, które… nie istnieją. Precyzyjnie i bez litości wytknął to autorce w serwisie X popularyzator historii Artur Wójcik.

Pomijając sprawy etyki zawodu i zwykłej przyzwoitości, warto podkreślić, że część przypisów została przez SI (trop prowadzi do ChatGPT) zwyczajnie zmyślona. Sztuczna inteligencja prawdopodobnie starała się odpowiednio dobrać przypisy do myśli zawartych w tekście. Autorka odrzuciła zarzuty o udziale SI w tworzeniu książki. Uznała, że „doszło do pewnego nieporozumienia, które obecnie wyjaśnia wydawnictwo". Smaczku całej sprawie dodają dwie rzeczy: Karolina Opolska wykłada dziennikarstwo na jednej z niepublicznych uczelni, a jej książka nosi tytuł „Teoria spisku, czyli prawdziwa historia świata". Ostatecznie wydawnictwo wycofało książkę ze sprzedaży.

Ale nie zawsze wykorzystanie halucynujących modeli językowych LLM przynosi tylko humorystyczne rezultaty. Przekonała się o tym firma Exdrog, ubiegająca się o kontrakt na utrzymanie dróg w Małopolsce. Exdrog złożył ofertę o wartości 15,5 mln zł, co wzbudziło wątpliwości urzędników Zarządu Dróg Wojewódzkich w Krakowie. Postanowiono sprawdzić, czy cena nie jest zaniżona w stosunku do realnych kosztów. Jak informuje PAP, w odpowiedzi firma przedstawiła wyjaśnienia – dokument liczył 280 stron. Jednak jeden z konkurentów zakwestionował wiarygodność tych wyliczeń – analiza ujawniła, że część argumentów została wygenerowana przez sztuczną inteligencję.

„Wykonawca powoływał się na nieistniejące, nigdy niewydane interpretacje podatkowe, które rzekomo dotyczyły podobnych spraw. Udowodniliśmy, że były to halucynacje sztucznej inteligencji" – powiedział „Pulsowi Biznesu" reprezentujący konkurencyjną firmę prawnik Jarosław Sroka. W efekcie Krajowa Izba Odwoławcza wykluczyła Exdrog z postępowania przetargowego.

To działa również w drugą stronę: przez halucynacje sztucznej inteligencji „wpadają" urzędnicy analizujący wnioski firm. Tak sugeruje SpidersWeb, podając przykład gliwickiej firmy KP Labs, która miała utracić 17 mln zł dotacji z powodu niewłaściwej oceny projektu, do której wykorzystano SI. KP Labs zgłosiło technologię obliczeniową do zastosowań satelitarnych, jednak eksperci Narodowego Centrum Badań i Rozwoju orzekli, że rozwiązanie to jest za mało innowacyjne. Na poparcie tej tezy podane zostały parametry konkurencyjnych urządzeń oferowanych przez inne firmy. Problem polegał jednak na tym, że te urządzenia albo nie wyszły poza fazę projektową, albo… nie istniały.

Mało tego, zapytane o sprawę NCBR odpowiedziało, że ekspert nie korzystał do oceny z modeli językowych, ale „z posiadanej wiedzy oraz ogólnie dostępnych narzędzi wyszukiwania Google i Bing". Warto tu podkreślić, że obie wyszukiwarki proponują obecnie tzw. podsumowania generowane przez sztuczną inteligencję. I prawdopodobnie właśnie tam wkradły się wymyślone przez SI urządzenia i „źródła" naukowe.

Antropomorfizacja algorytmów

Co ciekawe, termin „halucynacje" w odniesieniu do działania sztucznej inteligencji pojawił się na długo przed upowszechnieniem się tak popularnych dziś modeli językowych. W 2015 r. użył go Andrej Karpathy, współzałożyciel OpenAI i specjalista ds. sztucznej inteligencji w Tesli. Zauważył, że jedna z sieci neuronowych klasy RNN (*Recurrent Neural Network* – rekurencyjna sieć neuronowa) przetwarzając tekst „wymyśliła" przypis – link do informacji źródłowej. Ten sam model dość sprytnie postanowił ominąć problem przedstawienia dowodu algebraicznego, uznając, że to sprawa oczywista i niewymagająca przeprowadzenia dowodzenia.

Popularność modeli językowych takich jak ChatGPT sprawiła, że halucynacje SI zaczęły być powszechnie dostrzegane. Problem opisywała w oficjalnych komunikatach sama Open AI – raz nazywając je „błędami logicznymi modelu", raz „tendencją do wymyślania faktów w chwilach niepewności".

Problemem pozostaje sama nazwa halucynacje – wprowadza ona bowiem niepotrzebną i mylącą antropomorfizację komputerów. Trzeba jednak przyznać, że podobny zarzut można również postawić wobec innych zbliżonych określeń, takich jak „konfabulacja" czy „fabrykowanie".

Niezależnie od przyjętego określenia, warto powtórzyć, że twórcy platform SI zdają sobie sprawę ze skali wyzwania i stosują obecnie „bezpieczniki" ograniczające halucynacje. Nowsze modele LLM zostały nauczone, aby udzielać odpowiedzi „nie wiem" na każdy prompt. Jak w takim razie cokolwiek generują? Uruchamia się wtedy rodzaj generatywnej sieci GAN (*Generative Adversarial Network*), której dwie części konkurują ze sobą – jedna wytwarza odpowiedzi (generator), druga ocenia, na ile wytworzone dane przypominają rzeczywiste (dyskryminator). Jeżeli dyskryminator uzna, że odpowiedź przypomina prawdę, wyłącza moduł zabezpieczający.

Myśli czarnej skrzynki

Skoro tak, to dlaczego SI halucynuje? Internet jako źródło wiedzy jest generalnie mało wiarygodny – wypełniony jest przecież nie tylko rzetelnymi materiałami źródłowymi, lecz również opiniami, przeinaczeniami i dezinformacją. Modele sztucznej inteligencji nie potrafią same ocenić, czy konkretne sformułowania są prawdziwe (mogą za to ocenić, czy przypominają prawdę) – zwłaszcza wówczas, gdy są pytane o szczegóły dotyczące słabo opisanego tematu.

Chatboty łączą najrozmaitsze fragmenty tekstów z tysięcy, jeśli nie milionów źródeł, „wypluwając" nowe materiały. Co więcej – zwykle na to samo pytanie zadane dwa razy uzyskuje się dwie różne odpowiedzi (czasem uzupełniające się, a czasem zaprzeczające sobie). Eksperci starający się wyjaśnić problem halucynacji sztucznej inteligencji, jak na przykład ekonomista i statystyk Gary Smith (który zresztą nie kryje swojego krytycznego stosunku do SI), podkreślają, że algorytmy nie rozumieją słów, a zatem nie mogą ocenić, czy analizowane i generowane ciągi znaków opisują rzeczywistość, czy są fałszem.

Inna sprawa, że platformy SI, których celem istnienia jest odpowiadanie na pytania użytkowników – po prostu to robią. W przypadku dużych modeli językowych (LLM) typu GPT (*Generative Pre-trained Transformer*) sztuczna inteligencja trenowana jest, by przewidywać następne słowo w szyku. I to słowo się pojawia – nawet wówczas, gdy SI nie ma wystarczających danych pozwalających wygenerować prawdziwą odpowiedź. W przypadku innych typów SI problemem może być źle dobrany zestaw danych treningowych, które utrwalają błędne „skojarzenia". Jednak rzeczywistym wyzwaniem leżącym u podstaw halucynacji SI jest brak wyjaśnialności sztucznej inteligencji, czyli zrozumienia, dlaczego model podjął określoną decyzję.

Jak pokazują dane opublikowane w październiku ubiegłego roku na łamach magazynu „Nature", chatboty działające na najpopularniejszych platformach SI mają tendencję do udzielania takich odpowiedzi, które satysfakcjonują użytkownika – a niekoniecznie są prawdziwe. Zespół Myry Cheng z Uniwersytetu Stanforda sprawdził zachowania 11 chatbotów, w tym ChatGPT, Gemini, Claude, Llama i DeepSeek. Okazało się, że krzemowi pochlebcy dopasowywali swoje odpowiedzi do oczekiwań ludzi znacznie częściej, niż zrobiliby to żywi rozmówcy.

I tu pojawia się kolejny problem z halucynacjami SI – często są one wynikiem sprytnie napisanego promptu (przykład niedźwiedzi w kosmosie). Może się on odnosić do nieistniejącej osoby, zdarzenia, które nigdy nie miało miejsca, albo „naciskania" na model językowy, aby udzielił satysfakcjonującej odpowiedzi.

Przywidzenia kierowcy

Halucynacje modeli SI – podobnie jak u ludzi – mogą być zabawne, ale mogą też przynosić szkody. Tak było m.in. z wpisami Groka (systemu SI w serwisie X) negującymi Holokaust i wybielającymi Niemców. Sztuczna inteligencja napisała m.in., że krematoria w niemieckim obozie Auschwitz zostały pierwotnie zaprojektowane jako instalacje dezynfekcyjne do zwalczania chorób zakaźnych. Trzeba tu podkreślić, że Grok jest bardzo często wykorzystywany do weryfikowania prawdziwości twierdzeń użytkowników X (dawniej Twitter), a jego odpowiedzi praktycznie kończą dyskusję.

Kategoria „niebezpieczne konfabulacje" to niestety nie tylko chatboty. Trzeba tu wspomnieć np. o systemach SI analizujących obraz i kierujących pojazdami autonomicznymi. System, który zobaczy na ulicy nieistniejącego psa, może spróbować gwałtownie skręcić, aby uniknąć wypadku. Jeden z pierwszych, a na pewno niestety najsłynniejszy wypadek samochodu Tesla jadącego na Autopilocie w maju 2016 roku zdarzył się, ponieważ algorytm niewłaściwie ocenił otoczenie – „nie zauważył" białej ciężarówki na tle jasnego nieba – halucynował, że droga jest pusta.

To samo dotyczy niewłaściwej interpretacji wyników badania obrazowego. Obecnie, również w Polsce, oszałamiającą karierę robi teleradiologia – specjalista oceniający np. wynik badania tomograficznego czy rezonansu magnetycznego otrzymuje przez internet tylko obrazy, które musi zinterpretować. Często taki specjalista wspomaga się wyspecjalizowanym oprogramowaniem bazującym na sztucznej inteligencji. Ta zaś, na co zwracają uwagę sami

lekarze zlecający badania, potrafi pomylić skrzep krwi z ogniskiem nowotworowym. Podobny błąd może popełnić SI analizująca zmiany skórne – zaalarmować pacjenta z powodu nieistniejącego czerniaka.

Są też takie obszary, w których ludzie zapewne nigdy nie dostrzegą halucynacji. Doskonałym przykładem może być rynek finansów i ubezpieczeń. Trudno tu zweryfikować decyzje sztucznej inteligencji obliczającej wysokość składki, kalkulującej ryzyko czy przygotowującej plany biznesowe.

Jeszcze jesteśmy potrzebni

Jak często halucynuje sztuczne inteligencja? Od dwóch lat sprawę zmyślających chatbotów bada firma Vectara. Okazuje się, że nawet przy prostym zadaniu podsumowania artykułów chatboty konfabulowały – w zależności od modelu wymyślały od 3 proc. do nawet 27 proc. treści. Opisywane wyżej obwody zabezpieczające sprawiają, że nowe platformy LLM oszukują rzadziej – w przypadku modeli Google i OpenAI halucynacje to zaledwie 1–2 proc. odpowiedzi, a Anthropic – ok. 4 proc.

Co ciekawe, wskaźniki te pogarszają się u najnowszych tzw. modeli rozumujących (RLM). DeepSeek R1 ulega halucynacjom w ponad 14 proc. odpowiedzi, a OpenAI o3 – w prawie 7 proc. W niektórych testach te najbardziej zaawansowane modele sztucznej inteligencji podawały nieprawdziwe odpowiedzi na blisko co drugie pytanie. Dlaczego? Modele rozumujące są zaprojektowane tak, aby mogły poświęcić więcej czasu na „przemyślenie" złożonych problemów przed znalezieniem odpowiedzi. Rozwiązują problem krok po kroku, co sprawia, że narażają się na ryzyko halucynacji na każdym etapie. Im więcej czasu poświęcają na „rozumowanie", tym większe ryzyko halucynacji.

Wygląda zatem na to, że problemu halucynacji sztucznej inteligencji nie da się rozwiązać w prosty sposób – wynika on bowiem z samej jej zasady działania. W tym wszystkim jest jednak dla nas, ludzi, dobra wiadomość: nawet w erze SI liczy się rzetelna wiedza, zdrowy rozsądek i sporo sceptycyzmu wobec nowych technologii. Jak to ujmuje prof. Subbarao Kambhampati, badacz relacji SI – człowiek na Uniwersytecie Stanowym Arizony: „jeśli jeszcze nie znasz odpowiedzi na pytanie, to raczej nie zadawaj tego pytania sztucznej inteligencji".

Jak sobie radzić z oszukującą sztuczną inteligencją?

O to, jak uniknąć błędów, zapytaliśmy jednego z najlepszych ekspertów w tym temacie. Któż bowiem może wiedzieć więcej o halucynacjach SI niż sama SI, a konkretnie ChatGPT.

Halucynacje AI (czyli wymyślanie faktów, źródeł lub odpowiedzi „brzmiących pewnie, ale nieprawdziwych") da się znacznie ograniczyć, choć nie da się ich wyeliminować w 100 proc. Oto praktyczne i sprawdzone sposoby:

1. **Zadawaj precyzyjne pytania**
   Im bardziej ogólne pytanie, tym większe ryzyko halucynacji. Proś o daty, numery dokumentów, autorów.

2. **Wymagaj źródeł (i je sprawdzaj)**
   Zawsze proś o linki do oficjalnych stron, publikacji naukowych, aktów prawnych. Jeśli SI nie potrafi podać źródła albo podaje je ogólnikowo („badania pokazują…"), to sygnał ostrzegawczy.

3. **Ogranicz zakres odpowiedzi**
   Poproś, aby SI odpowiadała tylko na podstawie znanych danych i jasno zaznaczała niepewność.

4. **Unikaj pytań „zmyśleniowych"**
   SI najczęściej halucynuje przy bardzo niszowych osobach lub wydarzeniach, nieistniejących publikacjach, pytaniach typu „czy pamiętasz dokument X z 2014 r.?". Warto zapytać „czy istnieją wiarygodne informacje na ten temat?"

5. **Sprawdzaj kluczowe fakty w drugim źródle.**
   Traktuj SI jak asystenta, nie eksperta. Jest to szczególnie ważne w medycynie, prawie, finansach, historii i datach.

6. **Używaj SI do struktury, nie do faktów**
   Sztuczna inteligencja świetnie streszcza, porządkuje informacje, generuje checklisty i tłumaczy tekst. Ale fakty najlepiej weryfikować samodzielnie.

7. **Zadawaj pytania kontrolne**
   Poproś SI, aby wypunktowała założenia, wskazała, co może być niepewne i podała alternatywne interpretacje.

8. **Ustaw „tryb ostrożny" w promptach**

Odpowiedź ChatGPT została edytowana w taki sposób, aby była czytelna i spójna z formą pisma „Domena" – red.

Jak Warszawa tworzyła etyczne ramy dla generatywnej AI w administracji publicznej

Czwarty czerwca 2024 roku. Pierwsze szkolenie z cyklu spotkań dla grupy roboczej, która ma przygotować etyczne ramy dla generatywnej AI w Urzędzie m.st. Warszawy. Pierwszy raz pracujemy bezpośrednio z narzędziem. Na ekranach ChatGPT. Przy stole dziewiętnaście osób z ośmiu biur urzędu: informatycy, prawnicy, specjaliści od marketingu, ludzie z HR, analitycy. Jedni otwierają narzędzie po raz pierwszy w życiu, inni już zdążyli je poznać na własną rękę. Prowadzący tłumaczy, jak formułować zapytania, jak oceniać jakość odpowiedzi, na co uważać.

Michał Kuszewski

główny specjalista ds. wspierania rozwoju innowacyjnej gospodarki w Biurze Rozwoju Gospodarczego Urzędu m.st. Warszawy, koordynator projektu „Kierunki odpowiedzialnego wykorzystywania generatywnej sztucznej inteligencji". Łączy wykształcenie z zakresu psychologii i doświadczenie w projektowaniu UX z pracą w administracji publicznej. Członek Sekcji Aktualne Wyzwania Sztucznej Inteligencji przy PTI.

W połowie sesji ChatGPT przerywa pracę. Choć nasze komputery pracują bez zarzutu, a połączenie z siecią jest stabilne, okno czatu zastyga bez odpowiedzi. Jak się okazało, była to globalna awaria OpenAI.

Gdyby ktoś chciał wymyślić scenariusz szkoleniowy, który w trzydzieści sekund pokazuje, dlaczego nie można polegać na generatywnej AI jak na infrastrukturze, trudno było o lepszy. Ta awaria nie spowolniła naszej pracy. Wzmocniła przekonanie, że podejście, które obraliśmy, ma sens.

Najpierw etyka, potem technologia

Projekt „Kierunki odpowiedzialnego wykorzystania generatywnej sztucznej inteligencji w Urzędzie m.st. War-

szawy" nie powstał w odpowiedzi na kryzys. Nikt nie popełnił spektakularnego błędu z użyciem AI, nie wybuchł skandal. Elementy sztucznej inteligencji od lat działały w narzędziach, z których urzędnicy korzystają na co dzień, często nawet o tym nie wiedząc. Ale gwałtowna popularyzacja jej generatywnej odmiany zmieniła sytuację: nagle każdy mógł wpisać pytanie w ChatGPT i dostać odpowiedź, która wyglądała jak napisana przez eksperta. Ludzie zaczęli testować, sprawdzać, podejmować pierwsze próby użycia w swoich zadaniach. Pytanie nie brzmiało „czy używać", tylko „jak przygotować się na to, co nadchodzi".

Decyzja, żeby zacząć od etyki zamiast od technologii, była świadoma i szła pod prąd. Większość organizacji robi odwrotnie: kupuje narzędzie, szkoli ludzi, a dopiero potem zastanawia się nad zasadami. My odwróciliśmy kolejność. To nie był konserwatyzm ani strach przed nowością. To było przekonanie, że administracja publiczna, która zarządza dwumilionowym miastem, nie może sobie pozwolić na eksperymenty bez ram.

Projekt wyrósł z programu „Generujemy Innowacje" w ramach Strategii #Warszawa2030 i zyskał patronat Ministerstwa Cyfryzacji. Inicjatywa nie przyszła jednak z góry. Powstała oddolnie, z poziomu praktyka, który widział lukę i zaproponował, żeby ją wypełnić. Od początku założyliśmy, że powstaną dwa dokumenty: obszerny, szczegółowy zbiór wytycznych dla pracowników urzędu oraz krótki kodeks dla mieszkańców. Stolica musi dać obywatelom możliwość szybkiego wglądu w to, jak ich miasto podchodzi do sztucznej inteligencji.

Rozpoczęliśmy od wyboru partnera merytorycznego, szukając ludzi łączących umiejętności z zakresu facylitacji, etyki, technologii i projektowania usług. Wygrało konsorcjum Uniwersytetu SWPS, Centrum Etyki Technologii Instytutu Humanites i firmy EGO, specjalizującej się w *design thinking* dla sektora publicznego. Każdy z partnerów wnosił coś innego: SWPS dawał rygor akademicki, Centrum Etyki – normę i ramy, EGO – metodę pracy z grupą.

Grupa robocza to dziewiętnastu pracowników ośmiu biur urzędu: Biura Rozwoju Gospodarczego, Biura Informatyki, Biura Marketingu Miasta, Biura Strategii i Analiz, Biura Organizacji Urzędu, Biura Zarządzania Zasobami Ludzkimi, Stołecznego Centrum Bezpieczeństwa i Biura Prawnego. Wybór nie był przypadkowy, szukaliśmy jednostek, w których generatywna AI pojawi się najszybciej, gdyż zaproszenie wszystkich – przy czterdziestu jeden biurach w strukturze miasta – byłoby niewykonalne.

Rekrutacja przyniosła coś, czego nie planowaliśmy. Dyrektorzy typowali przedstawicieli, a ci okazali się zaskakująco różnorodni: obok entuzjastów technologii pojawili się sceptycy, którzy chcieli uczestniczyć właśnie dlatego, że mieli obawy. Ta mieszanka perspektyw okazała się najcenniejszym zasobem całego procesu. Entuzjaści widzieli możliwości, sceptycy widzieli pułapki. Razem widzieli więcej niż każda z grup z osobna.

Szkolenia i warsztaty

Półroczną pracę rozpoczęliśmy od cyklu szkoleń dotyczących samej technologii AI, po regulacje prawne i kwestie etyczne. Zależało nam, aby każdy członek zespołu najpierw zrozumiał, jak te narzędzia działają w praktyce, bo tylko wspólna płaszczyzna techniczna pozwalała nam rzetelnie oceniać ich konsekwencje. Te spotkania pełniły jednak jeszcze jedną, kluczową funkcję – budowały wspólny język.

> *Kiedy prawnik i informatyk zaczynają operować tymi samymi pojęciami, dyskusja o zasadach przestaje być starciem dwóch obcych żargonów, a staje się merytorycznym dialogiem.*

Po szkoleniach przeszliśmy do trzech intensywnych warsztatów. Pracowaliśmy metodami *design thinking* zaadaptowanymi do kontekstu polityki publicznej. Zaczynaliśmy od konkretu, każdy zespół definiował problemy ze swojego biura, miejsca, w których generatywna AI może pomóc albo zaszkodzić. Z tych szczegółowych obserwacji stopniowo wyłanialiśmy ogólniejsze zasady. Karteczki na tablicy, dyskusje w małych grupach, synteza, znowu karteczki. Siedem zasad przewodnich, które znalazły się w finalnym dokumencie, nie zostało napisanych przez ekspertów i narzuconych z góry. Wyrosły z doświadczeń ludzi, którzy na co dzień pracują w urzędzie.

Z tych wszystkich dyskusji zmaterializował się prototyp dokumentu. Nie był on idealny, jego format, grafika, sporo treści – wszystko to miało się jeszcze zmienić. Do tego momentu trapiła mnie wątpliwość: czy z pracy kilkudziesięciu osób o różnych perspektywach może powstać spójny tekst? Okazało się, że może.

Feedback szedł falami

Później zaczął się najtrudniejszy etap – zbieranie opinii. Najpierw grupa robocza recenzowała własną pracę. Potem przyszła pora na ekspertów zewnętrznych, którzy nie uczestniczyli w warsztatach i patrzyli świeżym okiem. Ich uwagi były szczegółowe, akademicko rygorystyczne i bardzo przydatne.

Prawdziwy test przyszedł jednak od zupełnie innej grupy – urzędników, którzy nie brali udziału w tworzeniu dokumentu, a mieli go potem stosować w codziennej pracy. Przeprowadziłem z nimi wywiady. Jeden moment zapamiętam na długo: rozmówca zapytał wprost, czy dokument będzie zawierał „klauzulę sumienia" pozwalającą odmówić korzystania z generatywnej AI z powodów etycznych. To pytanie uświadomiło mi dwie rzeczy. Po pierwsze, że obawy związane z AI sięgają głębiej, niż zakładaliśmy. Po drugie, że język naszego dokumentu, wzbogacony przez eksperckie konsultacje, stał się momentami nazbyt akademicki dla przeciętnego pracownika. Informacja zwrotna od ludzi „z frontu" przypomniała nam, kto na co dzień będzie korzystał z opracowywanego przez nas dokumentu i jak powinniśmy go dostosować.

„Kierunki GenAI" to obszerny dokument ze szczegółowymi wytycznymi, siedmioma zasadami przewodnimi i szerokim kontekstem merytorycznym – od wyjaśnienia, czym jest generatywna AI, po konkretne scenariusze użycia i ryzyka. Taki dokument jest jednak zbyt rozbudowany, żeby mieszkańcy mogli szybko sprawdzić, jak ich miasto podchodzi do generatywnej sztucznej inteligencji. Dlatego obok „Kierunków GenAI" powstał pięciostronicowy „Warszawski Kodeks GenAI", w którym siedem zasad zostało skondensowanych do czterech, ujętych tak, żeby każdy mógł je przeczytać w dziesięć minut i wiedzieć, o co chodzi. To właśnie Kodeks, jako wizytówkę podejścia Warszawy, przetestowaliśmy z mieszkańcami miasta. Pytaliśmy o zrozumiałość i wiarygodność: czy te zasady są dla was jasne? Czy ufacie, że miasto podchodzi do tematu poważnie?

Oba dokumenty zaprezentowaliśmy 27 stycznia 2025 r. podczas konferencji w Centralnym Domu Technologicznym. Wydarzenie było transmitowane na żywo i tłumaczone na język migowy. Fakt, że sala wypełniła się po brzegi, był dla nas jasnym sygnałem: temat sztucznej inteligencji przestał być niszowy.

Według naszej wiedzy był to pierwszy w Europie kompleksowy dokument samorządowy tego typu, dlatego od początku przygotowywaliśmy wersję anglojęzyczną, którą oficjalnie ogłosiliśmy na konferencji MIT Global Startup Workshop 2025. Zainteresowanie dokumentem wyraziły nie tylko Gdynia, Poznań, Rzeszów czy Zduńska Wola, lecz również zagraniczni partnerzy, jak Wiedeń, Wilno czy Monachium, którzy zaproponowali wymianę doświadczeń.

Po premierze grupa robocza przekształciła się w stałe ciało eksperckie z podgrupami tematycznymi, a do jej pierwotnego składu dołączyły nowe jednostki miejskie. W ramach tej szerszej grupy, dzięki wspólnej wymianie wiedzy i pracy nad konkretnymi wyzwaniami, zaczęły powstawać pierwsze projekty przełożenia zasad na praktykę. Potwierdziło to tezę leżącą u podstaw całego projektu: ludzie, którzy współtworzą zasady, potrafią je potem stosować.

Co wynoszę z tego procesu po ponad roku pracy? Przede wszystkim przekonanie, że ustalenie zasad przed wdrożeniem nie spowalnia, tylko przyspiesza prace. Kiedy ludzie wiedzą, co wolno i gdzie przebiegają granice, nie muszą się tego domyślać przy każdej czynności. Nie piszą maili z pytaniem, czy mogą użyć Clauda AI do przygotowania notatki. Mają ramy, więc działają.

> *„Przed rozpoczęciem projektu obawiałem się, że podejście „najpierw etyka" będzie odebrane jako hamulec. Okazało się odwrotnie: jasne zasady budują zaufanie, zarówno wśród pracowników, jak i na zewnątrz organizacji.*

Nauczyłem się też, że współtworzenie jest czasochłonne. Zebranie dziewiętnastu osób z ośmiu biur, zorganizowanie szkoleń, warsztatów, rund konsultacji to miesiące pracy, które można by skrócić, gdyby jedno biuro napisało wytyczne i rozesłało gotowe instrukcje. Tyle że ludzie, którzy sami kształtowali dokument, traktują go jako swój. Nie potrzebują zarządzenia, żeby go stosować, i to jest coś, czego odgórnym komunikatem się nie osiągnie.

Sam dokument to zresztą tylko część efektu. Drugą jest sieć ludzi, specjalistów z różnych biur, którzy rozumieją temat, potrafią o nim rozmawiać i wiedzą, do kogo zadzwonić, gdy pojawi się problem. Ta sieć okazała się równie wartościowa co sam tekst.

„Kierunki GenAI" są dokumentem żywym i będą się zmieniać wraz z technologią, regulacjami i potrzebami miasta, ale podejście, metoda i sposób myślenia zostają. Każdy samorząd może to powtórzyć niezależnie od wielkości i zasobów. Bazą może być nasz warszawski dokument, jednak najważniejsze jest powiązać ludzi, posadzić ich przy jednym stole, dać im czas i zacząć od pytania, jakie wartości chcemy chronić.

Wszyscy chcą dobrze, a wychodzi jak zwykle

O strukturalnym rozmyciu odpowiedzialności i o otwartej ranie etycznej w zespołach budujących systemy AI.

Wojciech Bednaruk

od ponad 20 lat zajmuje się technologiami edukacyjnymi. Pracował nad wdrażaniem systemów zarządzania rozwojem w Kanadzie, Europie Środkowo-Wschodniej i Afryce, tworzył cyfrowe programy szkoleniowe. Wykładowca etyki sztucznej inteligencji na Polsko-Japońskiej Akademii Technik Komputerowych, członek Sekcji AWSI przy PTI.

W styczniu 2024 r. Mark Zuckerberg stanął przed Komisją Senatu Stanów Zjednoczonych. W sali siedzieli również rodzice, których dzieci okaleczyły się lub popełniły samobójstwa w wyniku korzystania z Instagrama i Facebooka. Wewnętrzne raporty Mety, ujawnione przez sygnalistkę Frances Haugen w 2021 r., dokumentują, że firma wiedziała o szkodliwym wpływie swoich platform na zdrowie psychiczne nastolatków i mimo to kontynuowała optymalizację algorytmów w celu pogłębiania uzależnienia.

Język biznesu

Jeden z senatorów zmusza Zuckerberga do przeproszenia rodzin ofiar. Zuckerberg odwraca się w stronę rodziców i mówi: „Przepraszam za wszystko, przez co przeszliście”. A potem wraca do pytań senatorów i powtarza jak zaklęcie: „Zadaniem Mety jest budować narzędzia wiodące w branży i wzmacniać pozycję rodziców”.

Ta wypowiedź nie jest kłamstwem, to całkowite zastąpienie języka odpowiedzialności językiem misji korporacyjnej. Rodzice pytają o swoje skrzywdzone dzieci. Zuckerberg odpowiada, że takie oto są nasze cele produktowe. Co musi się stać z człowiekiem i z organizacją, żeby w takim momencie, publicznie, przed rodzinami skrzywdzonych dzieci, podkreślać jakość własnego produktu? Odpowiedź na to pytanie daje historia biznesu i reguł nim rządzących.

W 1994 r. siedmiu prezesów największych firm tytoniowych stanęło przed Komisją Zdrowia Izby Reprezentantów Stanów Zjednoczonych. Każdy z nich, pod przysięgą,

złożył to samo oświadczenie: nikotyna nie uzależnia. Wewnętrzne dokumenty firm, ujawnione przez sygnalistów i w postępowaniach sądowych, dowodziły, że firmy wiedziały o uzależniającym charakterze nikotyny od lat 50. XX w. i nadal prowadziły badania nad zwiększeniem jej stężenia w ich produktach.

Prezesi nie kłamali z powodu osobistej deprawacji. Kłamali, bo przez dekady uczestniczyli w systemie, który wytworzył własny język, własne kryteria oceny i własną definicję odpowiedzialności zawodowej. Definicję, z której krzywda konsumentów została strukturalnie wykluczona.

Firma Ford Motor Company wiedziała, że zbiornik paliwa w modelu Pinto jest wadliwy i podatny na eksplozję przy tylnym zderzeniu. Ocenia się, że latach 70. XX w. w wyniku pożarów samochodów zginęło co najmniej 27 osób. Wewnętrzna analiza firmy wykazała, że koszt naprawy usterki we wszystkich wyprodukowanych egzemplarzach przekroczy szacowane koszty odszkodowań za śmierć i obrażenia. Ford zdecydował się nie naprawiać wady.

Decyzja ta nie była wyrazem złej woli jednej osoby. Była produktem procesu, w którym ludzkie życie stało się zmienną w równaniu optymalizacyjnym i nikt – na żadnym etapie – nie poczuł się odpowiedzialny za całość.

Philip Morris przygotował w 2001 r. dla rządu Czeskiej Republiki analizę, rzeczowo argumentując, że masowe palenie papierosów przyniesie oszczędności dla systemu opieki zdrowotnej ze względu na skrócenie życia palaczy i redukcję kosztów emerytalnych oraz opieki geriatrycznej. Autorzy dokumentu byli analitykami biznesowymi. Dobrze wykonywali swoją pracę. Prawdopodobnie nikt z nich nie uważał się za człowieka, który sporządza rachunek zysków, w której główną zmienną jest ludzka śmierć.

> *Te cztery przypadki łączy jeden mechanizm, czyli rozmycie odpowiedzialności moralnej przez podział pracy, hierarchię i język profesjonalizmu. Ten mechanizm działa dziś w zespołach projektujących systemy sztucznej inteligencji.*

Etyczne zawieszenie to nie anomalia

Od pięciu lat prowadzę zajęcia z etyki AI ze studentami, którzy budują systemy sztucznej inteligencji lub wchodzą na ten rynek. Co semestr obserwuję ten sam wzorzec zachowań. Studenci opisują sytuacje, w których czuli, że coś jest nie tak, a mimo to pracowali dalej. Powtarzają się trzy narracje:

Narracja 1: „Nie wiedziałem, że to będzie problematyczne".

Narracja 2: „Nie mogłem nic zrobić, nie miałem władzy, nie miałem narzędzi".

Narracja 3: „Byłem sam. Nikt inny tego nie widział, albo wszyscy widzieli, ale milczeli".

Każda z tych narracji jest półprawdą, która chroni system, a nie człowieka, który w nim pracuje. Takimi narracjami posługują się nadal pracownicy Forda, analitycy Philip Morrisa, prawnicy firm tytoniowych, inżynierowie Mety i moi studenci.

Ale jest coś, o czym te narracje milczą. To cena, którą płaci człowiek, który z nich korzysta. Nie mam tu na myśli kary zewnętrznej, wyroku sądowego, utraty pracy czy nieprzychylnych nagłówków na portalach informacyjnych. Mam na myśli coś, co obserwuję w pracy z ludźmi, którzy przez lata funkcjonowali w systemach, które ich niepokoiły. Chodzi o niegojącą się ranę etyczną.

Teza, którą stawiam, brzmi następująco: szkodliwe systemy AI nie powstają przez złą wolę. Powstają przez strukturalną niemożliwość przypisania odpowiedzialności moralnej w systemie, który ją celowo rozprasza między role i procesy. I każdy człowiek, który przez długi czas działa wbrew własnym wartościom w takim systemie, płaci za to cenę psychologiczną, której żadna narracja przetrwania nie eliminuje. Tylko ją odracza.

Jak powstaje krzywda

W 2012 r. zespół Facebooka przeprowadził eksperyment na 689 tys. użytkowników, manipulując treścią ich feedu w celu zbadania zjawiska zarażenia emocjonalnego. Miesiące przygotowań, spotkań projektowych, przeglądów kodu, recenzji i przez cały ten czas nikt nie zadał pytań: „Czy mamy zgodę tych ludzi?" lub „Czy nie wyrządzamy im krzywdy?". Nie dlatego, że członkowie zespołu byli nieetyczni. Dlatego, że projekt operował językiem badań naukowych, a zespół posługiwał się konceptami *istotności statystycznej*, a nie *krzywdy psychologicznej*.

To jest pułapka ramowania. Kiedy projekt zostaje przedstawiony jako wyzwanie techniczne, w umysłach członków zespołu uruchamia się zestaw koncepcyjnych narzędzi technicznych. Przestają widzieć człowieka po drugiej stronie.

Zuckerberg przed Senatem mówił językiem produktowym, bo przez lata organizacja uczyła go, że to jest właściwy ję-

zyk dla każdej sytuacji. „Zadaniem Mety jest budować narzędzia wiodące w branży". To *dictum* jest prawdziwe jako opis celów korporacyjnych, ale nie odpowiada na pytanie, kto za te cele płaci zdrowiem a czasem i życiem.

> *I właśnie ta ślepota, wyuczona, strukturalna, nagradzana, jest tym, co Albert Bandura opisał jako moralne rozłączenie: psychologiczny mechanizm, przez który człowiek oddziela swoje działania od ich moralnego znaczenia, żeby móc funkcjonować.*

Wspierającym mechanizmem społecznym w takiej sytuacji jest rozproszenie odpowiedzialności. Na codziennym spotkaniu projektowym padają pytania: czy to działa?, czy to jest szybkie?, czy to się skaluje? Nie pada pytanie: kto może ucierpieć przez to, co budujemy? Inżynier AI myśli: „To nie moja rola". Kierownik produktu myśli: „Programiści znają szczegóły techniczne". Wszyscy razem budują system AI zanurzeni w etycznej ślepocie. Każdy wykonuje swoją pracę dobrze. Krzywda jest efektem systemu jako całości i nie pojawia się w żadnym module, nie pojawia się w żadnym logu, nie jest zaadresowana w żadnym tickecie.

Kiedy przedstawiam studentom dylemat etyczny i pytam: „Czy widzicie tu problem etyczny?", zapada cisza. Wszyscy patrzą w laptopy. Po chwili ktoś powie niepewnie: „Myślę, że..." i nagle widzę pięć rąk w górze: „Tak, też to widziałem!". Wszyscy widzieli. Każdy myślał, że jest sam. To jest pluralistyczna ignorancja. Większość prywatnie odrzuca normę, ale publicznie ją akceptuje, bo każdy sądzi, że jest w mniejszości. Prezesi firm tytoniowych pod przysięgą byli produktem organizacji, która przez dekady wytwarzała dokładnie ten mechanizm. Każdy wiedział i każdy sądził, że pozostaje sam z tą wiedzą.

Rana, która nie goi się sama

Jest coś, o czym literatura etyki organizacyjnej mówi rzadziej niż o mechanizmach systemowych. Co dzieje się z konkretnym człowiekiem, który przez miesiące i lata funkcjonuje wbrew własnym wartościom?

Nie ma tu jednej odpowiedzi. Obserwuję trzy wzorce, które wynikają z rozmów z ludźmi z branży technologicznej i które rozpoznaję w historycznych przypadkach korporacyjnych.

Wzorzec pierwszy: znieczulenie
Część osób radzi sobie z raną etyczną, wypracowując dystans do własnych reakcji moralnych. Przestają zauważać niepokój. Nie dlatego, że go nie odczuwają, ale dlatego, że nauczyli się ignorować ranę tak sprawnie, że znika z pola uwagi. To jest mechanizm powtarzalnego działania wbrew własnym wartościom, bez zewnętrznych konsekwencji, co prowadzi do obniżenia progu wrażliwości. Człowiek, który przez lata pracował w firmie tytoniowej nad zwiększeniem skuteczności uzależniania, nie stracił kompasu moralnego z dnia na dzień. Tracił go przez tysiące małych momentów, w których nie zauważył, że coś go niepokoi.

Wzorzec drugi: racjonalizacja
Inni zachowują świadomość niepokoju, ale stosują narracyjny opatrunek: „Taki już jest świat". „Gdybym ja tego nie zrobił, zrobiłby ktoś inny". „Przynajmniej my robimy to lepiej niż konkurencja". Te narracje są psychologicznie funkcjonalne, pozwalają spać w nocy i utrzymywać spójność postrzegania siebie jako przyzwoitego człowieka. Są też precyzyjnie opisane w literaturze jako mechanizmy moralnego rozłączenia – zniekształcenia poznawcze, które pozwalają nam działać wbrew własnym zasadom bez poczucia winy. Analitycy Philip Morris, którzy obliczali zyski z ludzkiej śmierci, musieli stosować jakiś wariant tej narracji. Inaczej dokument po prostu by nie powstał.

Wzorzec trzeci: ból bez ujścia
Część ludzi nie znieczula się i nie racjonalizuje i przez to nosi ranę przez lata. Wiedzą, co zbudowali. Pamiętają moment, w którym mogli zadać pytanie, i nie zadali. Pamiętają projekt, przy którym milczeli, bo ocena wydajności nagradzała dostarczanie wyników, a nie czujność etyczną. Ten ból, jeśli nie znajduje przestrzeni do wypowiedzenia, prowadzi – jak wynika z rozmów z doświadczonymi pracownikami branży – do goryczy i rozczarowania. Nie wobec jednej firmy czy jednego projektu, ale wobec całej branży czy nawet całego świata. Wobec idei, że technologia AI może być czymś więcej niż narzędziem optymalizacji cudzych przychodów.

Badania nad wyuczoną bezradnością pokazują, jak powtarzane doświadczenia braku wpływu prowadzą do generalizowanej pasywności. Prowadzą do stanu, w którym inżynier AI przestaje podejmować próby działania nawet wtedy, gdy działanie jest możliwe. Człowiek przestaje wierzyć, że jego osąd moralny ma jakiekolwiek znaczenie w strukturze, w której pracuje. I ta utrata wiary jest być może kosztowniejsza niż każda konkretna decyzja projektowa, przy której milczał.

Obserwuję to u studentów, którzy wchodzą na rynek pracy z wrażliwością etyczną i po dwóch latach zaczynają posługiwać się z językiem, który tę wrażliwość zastąpił. „Moim zadaniem jest budowanie narzędzi wiodących w branży" – słyszałem to od prezesów. Teraz słyszę od studentów pierwszego roku, którzy jeszcze nie napisali pierwszej linijki produkcyjnego kodu. Nie kłamią. Naprawdę w to wierzą, bo tak branża definiuje profesjonalizm… To jest może najgroźniejszy mechanizm ze wszystkich – znieczulenie wyprzedzające ranę.

Trzy poziomy interwencji

Firmy tytoniowe miały działy prawne i CSR. Ford miał inżynierów bezpieczeństwa. Philip Morris miał analityków. Meta ma zespoły ds. bezpieczeństwa i odpowiedzialnej AI – przynajmniej do czasu, kiedy te zespoły zostały zwolnione. Obecność procedur nie zastępuje osądu moralnego. A polityki korporacyjne nie leczą ran, których istnienia i tak nie uznają.

Dlatego moje rekomendacje kieruję do szeregowych pracowników firm projektujących systemy AI. Nie do zarządów i liderów. Do programistów, analityków danych, inżynierów ML, projektantów UX, do ludzi, którzy codziennie podejmują dziesiątki małych decyzji, z których każda wydaje się techniczna, a żadna z osobna nie wygląda jak decyzja moralna. I do tych spośród nich, którzy noszą w sobie ranę etyczną.

Poziom pierwszy: **zachowajcie kontakt z własną percepcją moralną.**
Kiedy coś was niepokoi w projekcie, zróbcie notatkę. E-maill do siebie z minimalnym zestawem danych: data, projekt, obawa. Nie z powodu odpowiedzialności organizacyjnej, tylko dla zachowania kontaktu z własną percepcją. Za tydzień, za miesiąc, za rok zaczniecie wątpić: „Może przesadzałem?". Zapis odpowie, że nie.

To jest jednocześnie elementarna uczciwość wobec siebie i ochrona przed znieczuleniem. Pamięć moralna wymaga pielęgnacji, tak jak każda inna.

Poziom drugi: **przełamcie izolację, choćby raz.**
Izolacja nie jest przypadkiem, jest narzędziem systemu. Nagradzane jest dostarczanie wyników, nie czujność etyczna. Ktoś kiedyś zgłosił temat i został odsunięty od projektu. Wszyscy inni widzieli, co się stało. Stąd nauka – nie sprawiaj problemów. I zaczynacie wątpić we własną percepcję: „Może tylko ja to widzę? Może przesadzam?".

Ale często nie tylko wy macie wątpliwości. Dlatego zapytajcie jedną osobę – nie publicznie, nie na spotkaniu – ale w cztery oczy: „Czy ty też masz uwagi co do tego projektu?". Odkrycie, że nie jesteście sami, przynosi istotną zmianę – rana etyczna przestaje być odczuwana w izolacji. Przestanie się ona pogłębiać za sprawą samotności, w której dotąd była noszona.

Pracownicy Google'a, którzy w 2018 r. protestowali przeciwko kontraktowi wojskowemu Projekt Maven, zebrali ponad cztery tysiące podpisów pod petycją. Zdobycie każdego podpisu było efektem jednej rozmowy w cztery oczy.

Poziom trzeci: **zadajcie jedno pytanie więcej, ale we właściwym momencie.**
Na przeglądzie kodu, na planowaniu sprintu, na spotkaniu przed wdrożeniem zapytajcie: „Czy testowaliśmy to na użytkownikach z niepełnosprawnościami?", „Kto może być poszkodowany, jeśli model pomyli się systematycznie dla określonej grupy demograficznej?", „Czy mimo tego, że model nie przetwarza danych osobowych, w jakiś sposób zmienia zachowania ludzi?"

To nie jest sabotaż. To jest staranność zawodowa i jednocześnie ochrona własnej integralności moralnej. Każde zadane pytanie jest małym aktem uczciwości wobec samego siebie. Nie zmieni architektury systemu. Ale zdecyduje o tym, czy patrząc w lustro za pięć lat, rozpoznamy siebie.

> *Budowanie takiego języka, pytanie po pytaniu, spotkanie po spotkaniu, to właśnie zadanie szeregowego pracownika firmy AI. I to jest jednocześnie jedyna dostępna profilaktyka przed raną etyczną.*

Zuckerberg przed Senatem, prezesi firm tytoniowych pod przysięgą, inżynierowie Forda przy arkuszach kalkulacyjnych, analitycy Philip Morris przy modelach finansowych – żaden z nich nie był moralnym potworem. Wszyscy byli profesjonalistami funkcjonującymi w systemach, które zoptymalizowały ich działanie, być może wbrew ich własnym wartościom i dostarczyły im narracji, dzięki którym mogli wmówić sobie, że wszystko jest w porządku.

Studenci, którzy dziś wchodzą na rynek AI, zasługują na coś więcej niż system, który znieczuli ich przed etyczną raną, zanim zdążą ją poczuć. Zasługują na środowisko, które daje im język do nazywania tego, co widzą i poczucie struktury, która sprawi, że zgłaszanie zastrzeżeń jest zawodowo możliwe, a nie samobójcze. Moje zadanie, jako edukatora, polega na tym, żeby budować właśnie takie środowisko.

Tekst jest rozwinięciem tematu, który Autor zaprezentował na konferencji „AI made in Poland" https://www.aimadeinpoland.com/

Subiektywny poradnik administratora

cz. V

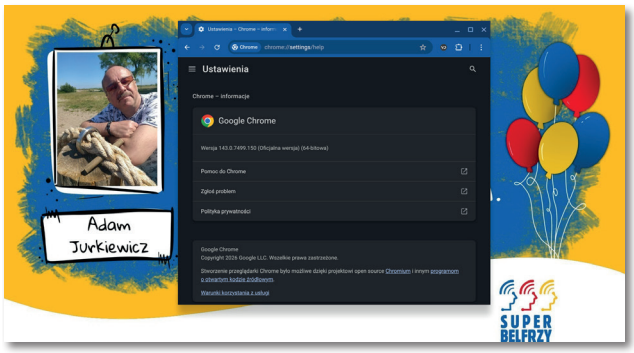
Zdjęcie wygenerowane za pomocą sztucznej inteligencji

Obiecałem odcinek o amerykańskim rozwiązaniu korporacyjnym. Danego słowa trzeba dotrzymywać, więc przedstawiam system operacyjny ChromeOS.

Adam Jurkiewicz

administrator sieci i serwerów Linux od ponad 25 lat. Programista Pythona, zwariowany nauczyciel młodzieży. Zdecydowany zwolennik oprogramowania open source i systemów Linux od 1993 r., których od ponad dwóch dekad używa w codziennej pracy. Członek zarządu Sekcji Informatyki Szkolnej przy PTI oraz członek oddziału mazowieckiego PTI. Dostępny w sieciach społecznościowych:
https://www.linkedin.com/in/adam-jurkiewicz-python-linux/
https://linux.social/@adam_jurkiewicz
https://jurkiewicz.chat

Nie byłoby tego systemu, gdyby nie ludzie dobrej woli z całego świata. Poniżej zrzut ekranu systemu z uruchomioną domyślną przeglądarką WWW (Google Chrome – niech nas to nie dziwi), zawierający informacje o wersji systemu oraz znamienne zdanie: „Stworzenie przeglądarki Chrome było możliwe dzięki projektowi open source Chromium i innym programom o otwartym kodzie źródłowym". Gdzie się nie ruszymy, tam open source. Nic dodać, nic ująć.



Teraz chciałbym przedstawić ten system bliżej – moim zdaniem to ciekawe rozwiązanie, które może być alternatywą dla komputerów przenośnych, które dzisiaj w 90 proc. przypadków są wykorzystywane przez administratorów serwerów (przynajmniej tak ja to widzę, to moja subiektywna ocena – być może mylna). Oczywiście czasami to macOS, lecz w większości przypadków to MS Windows. Większość z nas, adminów, zna MS Windows, ale nie każdy zna ChromeOS – więc zacznijmy od opisu z serwisu Wikipedia (https://pl.wikipedia.org/wiki/Chrome_OS).

Chrome OS – system operacyjny stworzony i rozwijany przez firmę Google oparty na Gentoo, dystrybucji Linuksa. Domyślnie umożliwia uruchamianie aplikacji internetowych (np. Progressive Web Apps) oraz większości aplikacji natywnych w formacie APK dla systemu operacyjnego Android (integracja ze sklepem Google Play), a także opcjonalnie aplikacji natywnych dla dystrybucji systemu operacyjnego Linux (wirtualizowany obraz Debiana z wykorzystaniem technologii Crostini).

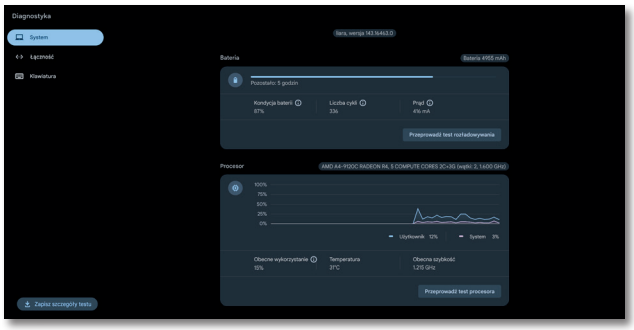
Przypatrzmy się interfejsowi:

- pulpit bez możliwości dodawania ikon, skrótów (to jedna z pierwszych widocznych różnic), ale tło możemy zawsze ustawić własne;
- pasek aplikacji na dole ekranu, właściwie analogiczny jak w Windows 11 czy macOS (uwaga: nie mamy wpływu na elementy oraz na ich lokalizację);
- menu aplikacji, jak w Windows 10, czyli po lewej stronie ekranu;
- menu „techniczne" po prawej stronie, gdzie znajdziemy standardowe opcje, jak wybór sieci WiFi (komputery Chromebook nie mają złączy Ethernet) czy wyłączenie komputera.

Właściwie wszystko jak w innych systemach, różnica to aplikacje. Na tym zrzucie od razu widać „Sklep Play" (znany z telefonów z Androidem); zdradzę, że w urządzeniach Chromebook możemy instalować i używać dokładnie te same aplikacje, co w naszych telefonach. Tak po prostu. Komputer staje się telefonem z klawiaturą, choć nie do końca. Nie zadzwonimy z niego, nawet, gdy w środku jest karta SIM. Jest większy i cięższy, ale za to mamy coś, czego nie mamy w telefonie.

TAK – dobrze widzisz, mamy terminal Linuksa, a w nim Debian 12 bookworm, a więc stabilny system z jądrem wersji 6. Co z tego wynika? Dostęp do tysięcy aplikacji przeznaczonych dla systemu Linux.

Aplikacja *neofetch*, którą tu widzimy, pokazuje nam najważniejsze dane o sprzęcie i systemie. Warto zwrócić uwagę na **procesor AMD A4** (patrz: https://www.dmcpu.com/cpu/amd-a4-9120c) – to wersja 1 z pierwszego kwartału 2019 r., a więc **ma już 7 lat!** Tak, tak, to nie jest chochlik drukarski – siedmioletni procesor i sprzęt, który ma zapewne podobny wiek, a ciągle działa! To mój prywatny Chromebook, który kupiłem 3 lata temu jako poleasingowy sprzęt (dzisiaj widzę, że podobnej klasy sprzęty nie przekraczają kwoty 200 zł brutto (słownie: dwustu złotych)!. Poniżej możemy zobaczyć diagnostykę części sprzętu – bateria ciągle sprawna, a procesor niezbyt obciążony.



Ale to dosyć stary procesor. Zgodzę się, to nie jest Ferrari, lecz powiedzmy sobie szczerze, czy Windows 11 na 7-letnim procesorze to coś, co możemy zobaczyć? A tu ciągle mamy aktualizacje systemu. I do tego nie działają skrypty PowerShell, nie ma systemowych DLL – jednym słowem, spora część dzisiejszych wirusów czy ransomware po prostu się nie uruchomi! To całkiem dobra informacja.

Administrator systemów Linux czy wirtualizacji (np. Proxmox) potrzebuje do swojej pracy głównie:

- przeglądarki internetowej,
- klienta ssh, aplikacji diagnozujących sieć, jak traceroute, WireShark, tcpdump,
- oprogramowania edytora tekstów (najlepiej formatu Markdown) do tworzenia dokumentacji,
- oprogramowania edytora tekstów (najlepiej formatu docx, czyli popularnego Worda) do modyfikacji dokumentów biurowych, podobnie z arkuszem kalkulacyjnym.

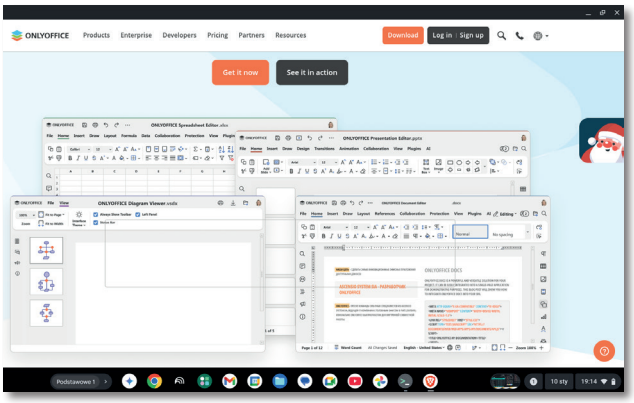
Przyjrzyjmy się tym programom.

Przeglądarka internetowa – domyślnie w systemie jest dostępny Google Chrome, ale zawsze możemy sobie zainstalować inne przeglądarki, jeśli zachodzi taka potrzeba (np. chcemy mieć dostęp do MS Teams, wtedy warto rozważyć MS Edge) lub po prostu nie lubimy narzędzi firmy Google, a wolimy np. Brave.

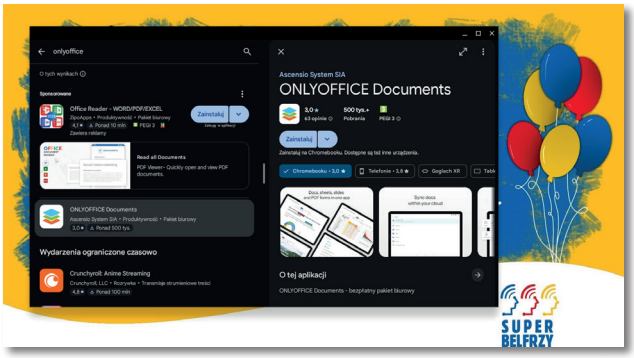
Klient ssh – mamy wbudowany cały system Linux! A więc wszystko, co tam jest, jest dostępne: ssh, curl, ping, Trippy, htop, jq, tcpdump, wiele innych również....

Edytor tekstów w formacie Markdown – tu pokażę pakiet o nazwie Obsydian (patrz: https://obsidian.md/), jest dostępny praktycznie na każdą platformę.

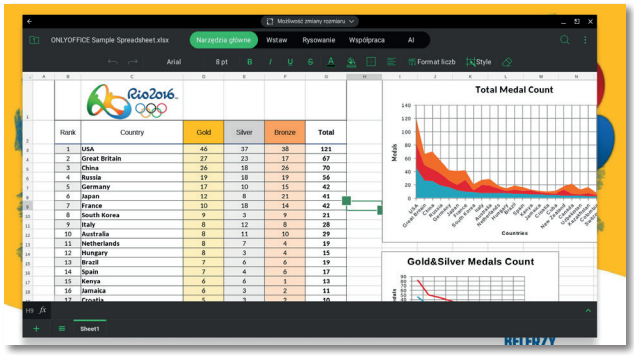
Pakiet biurowy, a więc edytor tekstów, arkusz, to minimum. Warto poznać Only Office (patrz: https://www.onlyoffice.com/).



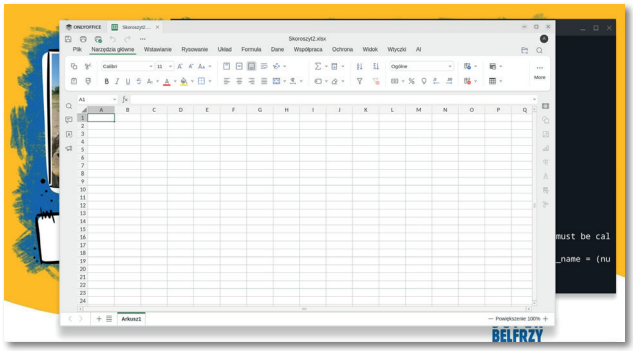
Niech Was nie zdziwi menu Pricing (są wersje płatne), ale mamy również wersje darmowe, np. na Androida, a więc i na ChromeOS.



Dzięki możliwości instalowania aplikacji przez Sklep Play, możemy cieszyć się różnymi bezpłatnymi programami. Niżej interfejs arkusza kalkulacyjnego z APK.



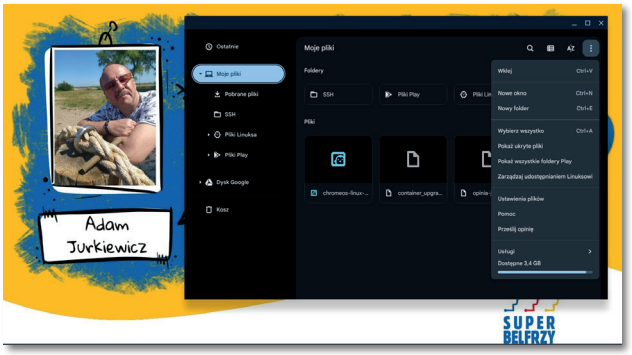
A teraz uwaga! Oto to samo oprogramowanie, lecz uruchomione jako aplikacja Linux!



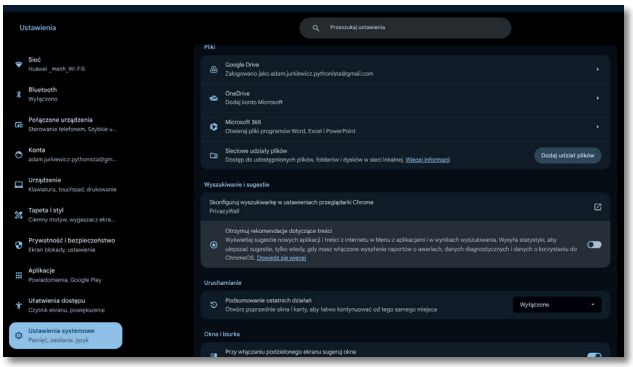
Myślę, że widok interfejsu mówi sam za siebie. I co najważniejsze, działa w pełni offline bez konieczności posiadania łącza do internetu!

Po prostu dzięki aplikacjom APK oraz DEB możemy używać Chromebooka zupełnie podobnie, jak laptopa z MS-Windows! Ale za inną cenę i bezpieczniej!

Mamy też najzwyklejszy Eksplorator Plików, w którym widzimy wszystkie lokalne pliki (na urządzeniu i na dysku Google, jeśli mamy dostęp online).



Mamy też możliwość połączenia różnych kont sieciowych – więc kochający usługę OneDrive czy Office O365 powinni czuć się jak w domu.



Z pewnością admini nagle nie rzucą się na Chromebooki. Ale zawsze warto wiedzieć, jakie mamy opcje do wykorzystania. Mam nadzieję, że tym artykułem trochę Was zachęcę do sprawdzania innych światów.

Wiesław Paluszyński
prezes PTI

Fot. Beata Sołtys

Zamiast felietonu

Przeglądam właśnie roboczą wersję sporego sprawozdania z działalności PTI w 15. kadencji. To zapis naszych aktywności – zarówno statutowych, jak i gospodarczych – w ciągu ostatnich 3 lat, a więc zapis aktywności Oddziałów i Zarządu Głównego.

Jak trafnie przewidywał kol. Jarek Deminet na ostatnim Zjeździe, ta kadencja była trudna. Przede wszystkim dlatego, że nasze tradycyjne źródła finansowania bazy działalności (Biura ZG) gwałtownie wyschły. Projekty unijne, na których bazowaliśmy w systemie certyfikacji ICDL, zanotowały znaczny zastój, szczególnie w 2024 r., a w 2025 r. – pomimo powolnego startu programu FERS – nie udało się odbudować niezbędnego poziomu przychodów. Ratowaliśmy się w tej kadencji znacznym ograniczeniem zatrudnienia (pracujemy bez Dyrektora Generalnego Biura) i wszelkich możliwych kosztów stałych. Zwiększone przychody Izby Rzeczoznawców pozwoliły na pokrycie części potrzeb budżetowych, ale nie obyło się bez sięgnięcia do zasobów w postaci sprzedaży posiadanej przez PTI nieruchomości. W efekcie udało się doprowadzić do zrównoważenia budżetu i w planie na 2026 r., realizowanym z powodzeniem, mamy zaplanowany niewielki dodatni wynik.

Co jednak najważniejsze, udało się nam zdobyć środki zewnętrzne dla realizacji działań statutowych. Dzięki naszym członkom wspierającym nie zawiesiliśmy wydawania „Domeny". Odbyły się kolejne edycje konkursów na najlepsze prace inżynierskie i magisterskie, a dzięki współpracy z NASK PIB zwiększyła się pula nagród dla laureatów konkursów. Rada Naukowa przeprowadziła kolejne edycje Konkursu na najlepszą Książkę Informatyczną Roku. W tym roku wznowiliśmy współfinansowanie Konkursu BÓBR, którego PTI jest współorganizatorem. Odbyły się kolejne edycje Światowych Dni Telekomunikacji i Społeczeństwa Informacyjnego we współpracy z NOT i Stowarzyszeniem Elektryków Polskich.

Kontynuujemy kolejne edycje konkursu GEEK, organizowanego przez Sekcję Informatyki Szkolnej i przeznaczonego dla młodzieży ze szkół podstawowych i średnich. Istotnie została wzmocniona pozycja Konferencji FedCSIS organizowanej przez Oddział Mazowiecki, obecnie najważniejszej naszej konferencji naukowej.

W ostatnich latach udało się stworzyć skuteczny system opiniowania wchodzących w życie aktów prawnych – wystarczy zerknąć na naszą stronę internetową, by zorientować się, jaki ogrom pracy włożyli członkowie PTI w te działania. Efektem jest wyraźny wzrost zainteresowania patronatami merytorycznymi PTI.

Jesteśmy widoczni w mediach społecznościowych, a nasza strona jest dobrym źródłem informacji o naszym towarzystwie. Te działania przyczyniły się do zwiększenia rozpoznawalności PTI. Niestety, nie wszystkie oddziały włączały się w te inicjatywy, nie proponowały też własnych. Szkoda, widocznie nie ustrzegliśmy się jakiś błędów.

Wszystkie nasze osiągnięcia to efekt całkowicie społecznej pracy wielu Koleżanek i Kolegów. Darujcie, że nie wymieniam nazwisk, ale nie chciałbym kogoś pominąć.

To mój ostatni tekst pisany w roli Prezesa PTI.

> *Ojcowie założyciele przewidzieli, że należy ograniczyć możliwość pełnienia funkcji prezesa do dwóch kadencji. To bardzo mądra zasada. Czuję bagaż tych dwóch kadencji.*

Czy udało się zrealizować wszystkie zamierzenia? Oczywiście, że nie. Już w pierwszej kadencji poległ mój pomysł stworzenia Oddziału Wirtualnego, co miało uwspółcześnić sposób funkcjonowania PTI w czasach mediów społecznościowych i dać PTI szansę lepszego odpowiadania na potrzeby młodych informatyków, wchodzących w życie zawodowe i naukowe. Przekonywano mnie, że tę rolę spełnią sekcje. Jeśli przyjrzymy się wynikom ich aktywności, sprawdziła się tylko jedna sekcja, ale i ona konsekwentnie chce się nazywać „przy PTI". Ostatnio pojawiają się głosy, że może warto wrócić do pomysłu również wirtualnego bytu. Bill Gates mawiał, że wspaniale jest świętować sukces, ale ważniejsze są lekcje, które wynosimy z porażek. Życzę więc nowemu Prezesowi (na razie tylko kol. Tomasz Królikowski publicznie ogłosił chęć objęcia tej funkcji), aby miał w tym zakresie więcej szczęścia.

Nie udało się zrealizować wielu innych pomysłów, bo priorytetem było szukanie środków finansowych i zapewnienie kontynuacji przede wszystkim tradycyjnych działań. Trzeba było wesprzeć kolegów Darka Rosłona i Bogusława Dębskiego w obowiązkach prowadzenia biura. Jestem im za ich aktywność ogromnie zobowiązany. Pomocy wymagała też Izba Rzeczoznawców, którą poprowadził – w tej współpracy nieoceniony – kol. Tadeusz Kifner.

Mam ogromny dług wdzięczności wobec Koleżanek i Kolegów, którzy całkowicie społecznie wsparli mnie w Zarządzie Głównym i wykazali się ponadprzeciętną aktywnością. Muszę wymienić te osoby:
Darka Rosłona, nieocenionego skarbnika,
Mariana Bubaka, siłę spokoju i rozsądku dzieloną wspólnie z Markiem Bolanowskim,
Tomka Klasę walczącego bez wsparcia finansowego o utrzymanie naszej informatyki,
Beatę Ostrowską, nieodmawiającą pomocy w trudnych chwilach,
Janusza Kacprzyka, życzliwego Przewodniczącego Rady Naukowej,
Bogusława Dębskiego wykonującego mrówczą pracę w obszarze certyfikacji,
Tomka Królikowskiego oraz Prezesów Oddziałów, których tu nie wymieniam, wspierających w terenie inicjatywy centralne,
Marcina Paprzyckiego wbijającego czasami życzliwe szpile
i **Wojtka Kulika** precyzyjnie organizującego nasz Zjazd.

Tak życzliwych, pracowitych i oddanych swojej misji osób mógłby sobie życzyć każdy prezes. Na koniec muszę serdecznie podziękować Prezes mojego macierzystego Oddziału Mazowieckiego – Marii Ganzhie. Zawsze mogłem na Ciebie, Mario, liczyć, pomagałaś dyskretnie, ale skutecznie – tego się nie zapomina.

I tak zamiast felietonu wyszło pożegnanie, może tak właśnie trzeba. Co do felietonów – jeśli kolejny Zarząd utrzyma „Domenę", to kto wie …

Kto jest u siebie w internecie

Michał Ogórek

satyryk i felietonista, od 1989 r. związany z „Gazetą Wyborczą". Obecnie pisuje w „Angorze". Autor wielu książek. Ostatnio wydał „Sto lat! Jak czciliśmy przywódców w ostatnim stuleciu", o kulcie przywódców – od Piłsudskiego przez Bieruta i Gomułkę po braci Kaczyńskich.

Obrzucanie się inwektywami w procesie historycznego rozwoju przeniosło się z karczmy i magla do globalnej sieci internetowej: w odpowiedzi na zaczepki naszego ministra Radka Sikorskiego Elon Musk nazwał go „śliniącym się imbecylem". Nawet gdyby nie kontekst, byłoby to wystarczająco żenujące, ale dochodzą jeszcze okoliczności. Wszystko dzieje się w wycinku rzeczywistości, przynależnej do Muska i bez niego nieistniejącej. Minister Sikorski musi się więc wpierw wystarać, aby móc wdać się z nim w pyskówkę, i odbywa się ona wyłącznie na warunkach Muska, a jeszcze za wysłuchiwanie od niego obelg mu płacić. Znosi jednak te upokorzenia, bo bez narzędzi Muska (co oznacza, że bez samego Muska) inaczej minister by nie zaistniał, co byłoby dla niego jeszcze gorsze od opluwania.

Wszystko to można jeszcze zaliczyć do politycznego folkloru, gdyby nie to, o co ta bijatyka. Musk jest też właścicielem Starlinka – systemu łączności, na którym bazuje całe bezpieczeństwo Ukrainy i bez którego dawno by jej już nie było (Polska to w jakiejś części finansuje, stąd te trzy grosze wtrącane przez Sikorskiego). Ostatnio jednak jakieś elementy systemu wykradła i zaczyna stosować przeciw Ukrainie Rosja i jeśli się czegoś z tym nie zrobi i zaczną go mieć obie walczące armie, straci całą swą przewagę i użyteczność.

Samo to już wystarczająco wywraca dotychczasowy przebieg wojny, ale z globalnego punktu widzenia jest wyłomem jeszcze bardziej znaczącym. Świat już dawno połknął język i pogodził się z tym, że cała machina, rządząca życiem – a na Ukrainie i śmiercią – ludzi jest czyjąś prywatną własnością. Taki amerykański bogacz może jej udzielać, udostępniać za tyle, na ile ją wyceni albo też odmówić jej użyczenia, jeśli kogoś nie stać albo posiadaczowi się nie podoba. Wszyscy na świecie, łącznie z kompletem ministrów spraw zagranicznych oraz całe armie, są podporządkowani tej zamorskiej, amorficznej dyspozycji.

Dopiero na tym tle widać rewolucyjny ruch Rosji: ona Muska tego władztwa właśnie pozbawia. Robi to metodami przestępczymi i gangsterskimi, a i jeszcze przeciwko nam, ale nie da się odmówić przełomowości takiego działania. W sumie wielka szkoda, że musiał to zrobić akurat taki wykwit terroru jakim jest Rosja, powtarzając tu zresztą swoją akcję wykradzenia ze Stanów Zjednoczonych tajemnic bomby atomowej przez szpiegów Rosenbergów, co jednak – jakby na nie patrzeć – odebrało Ameryce na nią monopol.

Żeby nie wyszło, że Rosja tak się nam w tym podoba, ale niewątpliwie elitom zachodnim tylko ona psuje ich zabawę i mąci błogi spokój. Akurat Rosja podkrada im tylko to, co przyda się na wojnie i tylko w tym zakresie dokonuje swoich dywersji. A ci nigdy się nie nauczą, żeby nie przekazywać niczego na Wschód, bo tam wszystko zajumają.

Kiedy Gutenberg wynalazł druk, co jest może porównywalnym przewrotem kopernikańskim do obecnego, niemal natychmiast stał się on wspólnym dorobkiem ludzkości. Nikt nie musiał się opłacać Gutenbergowi za używanie czcionek, drukując Biblię, której nikt z nich przecież nie napisał. Gdyby Gutenberg był ówczesnym Muskiem, do dziś wszyscy czytalibyśmy szwabachę, którą by nam obowiązkowo wcisnął.

Bracia Lumiere wynaleźli kino nie tylko dla siebie, ani nawet nie tylko dla Francuzów, a nawet ostatnio szczególnie nie dla Francuzów, których kinematografia zapadła się zupełnie. Wszystkie te odkrycia zostały oddane ludziom, jak ogień w mitologii greckiej, który bogom wykradł w końcu Prometeusz.

Pewnie prędzej czy później tak się i stanie z rzeczywistością cyfrową, do której klucze nie pozostaną na zawsze zamknięte w sejfach biurowców Doliny Krzemowej. Lepiej by było, żeby nie wykradły ich akurat hordy ze Wschodu, ale na razie nikt inny się nie kwapi.

Polecamy
numer specjalny „Domeny”

PTI
POLSKIE TOWARZYSTWO INFORMATYCZNE

NUMER SPECJALNY

Domena

PRZYSZŁOŚĆ ZACZYNA SIĘ DZISIAJ

Numer specjalny towarzyszący

Konferencji z okazji 45-lecia PTI I ŚDTISI

INFORMATYKA 2026 OCZAMI PTI

■ dokonania ■ perspektywy ■ wyzwania

PRZEGLĄD TEMATYKI WYSTĄPIEŃ KONFERENCYJNYCH