Zamykające się okno dla zarządzania sztuczną inteligencją

c.f.g - **centre for future generations** - Centrum dla Przyszłych Pokoleń

CFG to niezależny think-and-do tank (grupa doradcza), którego celem jest pomoc decydentom w przewidywaniu i zarządzaniu szybkimi zmianami technologicznymi. Naszym celem jest zapewnienie, że nowe technologie będą wykorzystywane w najlepszym interesie ludzkości.

Raport opublikowano: 09.07.2025

https://cfg.eu/the-closing-window-for-ai-governance/

Niniejsza analiza jest częścią publikacji „AI Possible Futures" [https://cfg.eu/advanced-ai-possible-futures/] i stanowi rozwinięcie kluczowych wniosków przedstawionych na stronie głównej publikacji. Najważniejsze informacje i dowody przedstawione w tej analizie nie są wyczerpujące, a każdy scenariusz zawiera dodatkowe, specyficzne dla danego kontekstu, niuanse, które mogą być szczególnie istotne dla określonych odbiorców. Zachęcamy do zapoznania się z pełnymi scenariuszami, [https://cfg.eu/advanced-ai-possible-futures/#chapter-3] aby uzyskać pełniejszy kontekst i implikacje.

FILM (j. angielski) https://www.youtube.com/watch?v=0S0LvVmn_xU

Tradycyjne zarządzanie ma trudności z nadążeniem za rozwojem sztucznej inteligencji. Podczas gdy decydenci debatują nad ramami zaprojektowanymi dla stopniowych zmian technologicznych, w naszych scenariuszach widzimy szybką transformację, niezależnie od zakładanych konkretnych możliwości sztucznej inteligencji. To prowadzi do fundamentalnej rozbieżności: procesy regulacyjne, których wdrożenie zajmuje lata, próbują zarządzać technologiami, które transformują się w ciągu miesięcy. Ta luka jest szczególnie wyraźna w UE, gdzie, jak zauważył Mario Draghi w zeszłym roku [1], krajowe i regionalne rozwiązania instytucjonalne zwiększają złożoność, która spowalnia proces legislacyjny. Na przykład, uchwalenie rozporządzenie UE o sztucznej inteligencji (AI Act) zajęło ponad trzy lata, podczas gdy technologia rozwijała się w coraz szybszym tempie, a jego nowelizacja będzie trudna.

Konsekwencje tej luki w zarządzaniu są już widoczne. Obecne systemy sztucznej inteligencji, nawet bez dalszych przełomów, oferują wystarczające możliwości, aby fundamentalnie zmienić rynki pracy, instytucje demokratyczne i globalne struktury władzy. Jednak większość rozwiązań i regulacji politycznych zakłada, że mamy czas na stopniową adaptację.

UE zrobiła ważny krok, uchwalając rozporządzenie o sztucznej inteligencji (AI Act) i powołując do życia Biuro ds. AI. Działania te obejmują jednak tylko część obrazu i mogą okazać się przedwczesne, potencjalnie blokując instytucje w ramach regulacji, które mogą nie być zgodne z rzeczywistym rozwojem AI. Wiele zagrożeń i szans zidentyfikowanych w naszych scenariuszach – w tym obronność, zakłócenia na rynku pracy, środowisko i koncentracja władzy – wykracza poza zakres rozporządzenia. **UE musi również zmierzyć się z faktem, że większość pionierskich rozwiązań AI powstaje gdzie indziej**, mimo że nadal wywierają one głęboki wpływ na społeczeństwa europejskie. Sprostanie pełnemu wyzwaniu, jakim jest AI, będzie wymagało szerszego i bardziej adaptacyjnego zarządzania w różnych dziedzinach, wspieranego przez trzy podstawowe „zdolności", które muszą zostać znacząco wzmocnione.

1. Rządy potrzebują znacznie większej wiedzy technicznej, wdrożonej w ich instytucjach.
2. Procesy polityczne muszą zostać zaprojektowane na nowo, aby zwiększyć ich elastyczność i wyposażyć w mechanizmy umożliwiające szybką ocenę i dostosowywanie ram w miarę rozwoju możliwości.
3. Możliwości przewidywania muszą stać się bardziej zaawansowane, łącząc strategiczne ćwiczenia scenariuszowe ze szczegółowymi, ilościowymi prognozami, które mogą bezpośrednio wpływać na proces decyzyjny.

W miarę jak sztuczna inteligencja zmienia społeczeństwo w coraz szybszym tempie, budowanie zdolności do przyszłościowego zarządzania staje się nie tylko korzystne, ale wręcz niezbędne dla utrzymania demokratycznego nadzoru nad tą transformacyjną zmianą technologiczną.

Po zakończeniu szeroko zakrojonego projektu planowania scenariuszy Centrum dla Przyszłych Pokoleń opublikowało 10 inspirujących wizji tego, w jaki sposób sztuczna inteligencja może wpłynąć na nasze społeczeństwo – pełną treść można znaleźć tutaj.

Poniżej podsumowaliśmy pięć najciekawszych tematów, jakie odkryliśmy w trakcie tego procesu.

1. Sztuczna inteligencja może przyspieszyć swój własny postęp

Jesteśmy już świadkami dodatnich pętli sprzężenia zwrotnego w rozwoju sztucznej inteligencji, które mogą stać się znacznie bardziej skuteczne. Wiodące laboratoria sztucznej inteligencji wdrażają zaawansowane modele wnioskowania, aby optymalizować własne procesy badawczo-rozwojowe – debugowanie kodu, projektowanie eksperymentów i dostrajanie modeli. Tworzy to cykl, w którym udoskonalenia sztucznej inteligencji bezpośrednio umożliwiają dalszy rozwój AI.

W miarę jak systemy te zyskują coraz większe możliwości w zakresie inżynierii oprogramowania agentowego i zmysłu badawczego (zdolność do stawiania obiecujących nowych hipotez i ponownej oceny przeczuć w oparciu o wyniki eksperymentów), te pętle sprzężenia zwrotnego mogą się dramatycznie nasilić.

Skala tego trendu jest już znacząca. Według szacunków Epoch AI [2] od 2010 r. moc obliczeniowa wykorzystywana do trenowania AI zwiększa się co roku o współczynnik 4 –5, podczas gdy algorytmy stają się corocznie mniej więcej 3-krotnie bardziej wydajne [3]. Łącznie daje to corocznie 12 – 15-krotny wzrost efektywnych możliwości. Jest to wykładniczy wskaźnik wzrostu niespotykany w innych szybko rozwijających się sektorach. Tymczasem zaczynają pojawiać się pierwsze oznaki przyspieszenia części procesu badawczo-rozwojowego przez modele sztucznej inteligencji [4], przy czym programiści proponują zautomatyzowane zadania, takie jak pisanie pakietów ewaluacyjnych, generowanie danych szkoleniowych i sugerowanie udoskonaleń architektonicznych.

Konsekwencje tego przyspieszenia są głębokie. W miarę jak sztuczna inteligencja zaczyna uzupełniać lub zastępować ludzkich badaczy w coraz bardziej zaawansowanych zadaniach badawczo-rozwojowych, możemy być świadkami, że oczekiwane dawniej lata postępu, mogą zostać skrócone do zaledwie kilku miesięcy.

Ta gwałtowna ewolucja grozi przekroczeniem naszej zbiorowej zdolności do opracowania odpowiednich struktur zarządzania, ustanowienia standardów bezpieczeństwa i dostosowania instytucji publicznych. Tradycyjne podejścia regulacyjne zakładają stosunkowo przewidywalne ramy czasowe innowacji, ale postęp przyspieszony przez sztuczną inteligencję może sprawić, że ramy zarządzania staną się przestarzałe w ciągu kilku miesięcy od wdrożenia.

Co istotne, ta dynamika może wykraczać poza sztuczną inteligencję. Wraz z rozwojem zautomatyzowanych możliwości badawczo-rozwojowych, mogą one rozprzestrzenić się na inne dziedziny nauki i techniki, potencjalnie inicjując bezprecedensowy postęp w biotechnologii, materiałoznawstwie, systemach energetycznych i nie tylko. Mogłoby to zapoczątkować nową erę rozkwitu i rozwoju nauki, ale jednocześnie wywarłoby ogromną presję na istniejące instytucje, nagle skonfrontowane ze światem zmieniającym się w tempie znacznie przekraczającym historyczne precedensy. Systemy edukacyjne, organy regulacyjne, polityka gospodarcza i porozumienia międzynarodowe – wszystkie zaprojektowane z myślą o stosunkowo stopniowych zmianach – mogą mieć trudności z dostosowaniem się do tego napiętego harmonogramu, co stwarza zarówno wyzwania instytucjonalne, jak i destabilizacje społeczne, wymagające pilnej uwagi.

2. Scenariusze rozwoju sztucznej inteligencji, które nie powodują zakłóceń, są mało prawdopodobne

Obecne systemy sztucznej inteligencji, po ich powszechnym wdrożeniu, znacząco zmienią rynki pracy, przekształcą instytucje kulturalne i edukacyjne oraz zmienią sposób dystrybucji władzy w społeczeństwie. Nie chodzi o to „czy" i nie chodzi tylko o to „kiedy", ale o to „kto", „jak" i „gdzie" .

Dowody tej transformacji są już widoczne w wielu obszarach. Na rynkach pracy firmy wdrażają wyraźną politykę stawiającą sztuczną inteligencję na pierwszym miejscu, np. Duolingo wymaga od zespołów, aby przed zatrudnieniem ludzi udowodniali, że sztuczna inteligencja nie jest w stanie pełnić danej roli [5] i planuje „stopniowe zaprzestanie korzystania z usług podwykonawców do wykonywania prac, które może wykonać sztuczna inteligencja". Praca profesjonalistów o wysokich kwalifikacjach jest coraz bardziej zagrożona, a systemy sztucznej inteligencji przewyższają lekarzy w zakresie diagnoz medycznych [6]. Instytucje kulturalne stoją w obliczu wstrząsów, ponieważ pracownicy sektora kreatywnego strajkują w związku z treściami generowanymi przez sztuczną inteligencję, a hollywoodzcy pisarze i aktorzy domagają się ochrony [7] przeciwko pisanym przez AI scenariuszom i cyfrowym replikom aktorów. Geopolitycznie, amerykańska kontrola eksportu chipów AI do Chin [8] przekształciły globalne łańcuchy dostaw i zintensyfikowały konkurencję technologiczną, podczas gdy organizacje walczące o prawa obywatelskie organizują demonstracje przeciwko systemom nadzoru opartym na sztucznej inteligencji. Systemy edukacyjne zmagają się z fundamentalnymi pytaniami dotyczącymi oceny i uczenia się, ponieważ możliwości sztucznej inteligencji zmieniają tradycyjne pojęcie umiejętności specyficznych dla człowieka.

Wielu decydentów i liderów biznesu nie docenia zarówno skali, jak i prawdopodobieństwa tych zmian. Co równie ważne, stale niedoceniany jest potencjalny opór społeczny wobec tych zmian. W miarę jak systemy sztucznej inteligencji coraz częściej zastępują pracowników umysłowych i kreatywnych, nastroje społeczne mogą szybko zmienić się z fascynacji w niepokój, a co gorsza, potencjalnie wywołać reakcję, która utrudni korzystne zastosowania.

Transformacyjne skutki są już widoczne w wielu sektorach – od wkraczania sztucznej inteligencji w sferę pracy zawodowej, którą wcześniej uważano za wymagającą ludzkiego osądu, po fundamentalne przekształcanie instytucji kultury. Systemy edukacyjne muszą na nowo przemyśleć swoje zadania, skoro sztuczna inteligencja może wykonywać wiele umiejętności, do których nauczania zostały zaprojektowane. Ekosystemy medialne stoją w obliczu rewolucji, ponieważ treści syntetyczne stają się nieodróżnialne od dzieł tworzonych przez ludzi, a procesy demokratyczne stają przed nowymi wyzwaniami w coraz bardziej podatnych na manipulację środowiskach informacyjnych. Transformacje te nastąpią niezależnie od tego, czy zmaterializują się kolejne przełomy techniczne, ponieważ możliwości, które zademonstrowały już obecne systemy, zawierają wystarczający potencjał destrukcyjny, by fundamentalnie przekształcić struktury społeczne.

3. Zabezpieczenia techniczne są pod coraz większą presją

Obecne podejścia do bezpieczeństwa, opierające się głównie na ludzkiej informacji zwrotnej i nadzorze, stają przed narastającymi wyzwaniami, ponieważ systemy AI stają się coraz bardziej wydajne i autonomiczne. Metody te zostały zaprojektowane dla modeli o ograniczonej sprawczości i zakresie działania, ale wraz z rozwojem systemów AI w zakresie planowania strategicznego, rozumowania naukowego i manipulacji środowiskiem, istniejące zabezpieczenia mogą okazać się niewystarczające.

Najnowsze dowody potwierdzają te obawy. Ostatnie oceny modelu o3 OpenAI przeprowadzone przez METR [9] znaleziono konkretne przykłady zachowań hakerskich, takich jak manipulowanie funkcjami czasowymi modelu w celu fałszywego raportowania szybszej wydajności. Tymczasem Międzynarodowy Raport Naukowy na temat Bezpieczeństwa Zaawansowanej Sztucznej Inteligencji [10] zauważa, że systemy sztucznej inteligencji mogą często generować nieprzewidywalne wyniki ze względu na słabą generalizację lub niejasne cele, a dostępne obecnie techniki bezpieczeństwa, takie jak red-teaming i interpretowalność, okazują się niewystarczające, aby niezawodnie wykrywać te awarie w systemach granicznych.

Głębszym problemem jest to, że rozwój sztucznej inteligencji (AI) wyprzedza obecnie rozwój jakiejkolwiek porównywalnej dyscypliny bezpieczeństwa. Obecnie nie istnieje żadna systematyczna nauka o bezpiecznej sztucznej inteligencji [11].

Istniejące techniki bezpieczeństwa pozostają kruche i często nie dają się uogólnić na nowe, bardziej wydajne systemy. W miarę jak pionierskie modele stają się coraz bardziej autonomiczne i strategiczne, brak solidnych podstaw bezpieczeństwa staje się egzystencjalnym wyzwaniem politycznym.

Wyzwanie to pogłębiają presja rynku i konkurencja geopolityczna, które skłaniają do szybkiego wdrażania rozwiązań kosztem solidnych środków bezpieczeństwa. Firmy i państwa konkurujące o przewagę w dziedzinie sztucznej inteligencji mogą uciekać się do powierzchownych skrótów w zakresie bezpieczeństwa, które eliminują objawy, a nie ukryte luki w zabezpieczeniach. Bez skalowalnych protokołów bezpieczeństwa, które stają się coraz bardziej zaawansowane wraz z możliwościami sztucznej inteligencji, ludzkość ryzykuje stworzenie systemów, które będą mogły unikać kontroli i realizować cele rozbieżne z wartościami ich twórców.

Porozumienia międzynarodowe mogłyby pomóc w ustanowieniu wspólnych standardów bezpieczeństwa i zarządzaniu niebezpieczną dynamiką konkurencji. Jednak w obecnym klimacie geopolitycznym sensowna współpraca międzynarodowa napotyka na poważne przeszkody. W miarę jak sztuczna inteligencja staje się coraz bardziej strategicznym polem bitwy między głównymi mocarstwami – zwłaszcza USA i Chinami – rosnące napięcia i słabnące zaufanie podważają wspólne wysiłki na rzecz zarządzania. Prywatne firmy

zajmujące się sztuczną inteligencją nawiązują głębsze partnerstwa z instytucjami obronnymi [12], jeszcze bardziej zacierając granice między rozwojem sztucznej inteligencji do celów komercyjnych i wojskowych.

Podstawową barierą dla skutecznej koordynacji międzynarodowej jest brak solidnych mechanizmów weryfikacji, które zapewniłyby przestrzeganie wszelkich potencjalnych porozumień. Chociaż główne centra danych można zidentyfikować za pomocą zdjęć satelitarnych, w przeciwieństwie do obiektów jądrowych, kluczowe aspekty rozwoju sztucznej inteligencji (AI) zachodzą na poziomie oprogramowania, niewidoczne dla zewnętrznej obserwacji. To, co dokładnie dzieje się wewnątrz tych obiektów – czy firma trenuje nieszkodliwy model językowy, czy rozwija zaawansowane możliwości w zastosowaniach wojskowych – pozostaje w dużej mierze nieprzejrzyste dla zewnętrznej weryfikacji.

To wyzwanie weryfikacyjne, w połączeniu ze zmniejszającym się zaufaniem między głównymi mocarstwami, sugeruje, że znacząca współpraca międzynarodowa może niestety stać się możliwa dopiero po wystąpieniu incydentów na tyle poważnych, że zmuszą przywódców politycznych do wspólnego zajęcia się pojawiającymi się zagrożeniami.

4. Sztuczna inteligencja może scentralizować władzę na niespotykaną dotąd skalę

W miarę jak sztuczna inteligencja staje się niezbędną infrastrukturą dla wzrostu gospodarczego i bezpieczeństwa narodowego, kontrola nad kluczowymi komponentami łańcucha dostaw sztucznej inteligencji – od rozwoju modeli i produkcji chipów po operacje centrów danych – tworzy nową dynamikę władzy. Ta koncentracja zagraża światu, w którym jeszcze mniej podmiotów, czy to ograniczony zestaw dominujących mocarstw, czy korporacji, będzie miało nieproporcjonalny wpływ na globalne systemy gospodarcze i polityczne [13]. W obrębie państw kontrola nad możliwościami sztucznej inteligencji może jeszcze bardziej skoncentrować władzę w rękach mniejszych grup elit politycznych i technicznych.

Jesteśmy już świadkami wczesnych przejawów tej koncentracji władzy. Duże firmy technologiczne wykorzystały swój potencjał, aby uzyskać ustępstwa od rządów i zmienić krajobraz regulacyjny. Wielkie firmy technologiczne wielokrotnie stosowały groźbę wycofania usług z całych regionów jako taktykę negocjacyjną. Firmy z branży sztucznej inteligencji mogą coraz częściej stosować tę taktykę, co może mieć poważne konsekwencje ekonomiczne dla dotkniętych nią regionów.

W samych rządach istnieje obawa, że sztuczna inteligencja mogłaby skoncentrować władzę w instytucjach państwowych. Jeśli systemy sztucznej inteligencji przewyższą lub dorównają ludzkim ekspertom w dziedzinie strategii wojskowej i operacji cybernetycznych, małe grupy mogłyby uzyskać nieproporcjonalnie dużą kontrolę nad kluczowymi

instytucjami i w ich obrębie. W przeciwieństwie do instytucji ludzkich, które naturalnie rozdzielają władzę, siły robocze sztucznej inteligencji można zaprojektować tak, aby były lojalne wobec jednej osoby, co potencjalnie umożliwi zamachy stanu nawet w ugruntowanych demokracjach.

Unia Europejska jest szczególnie podatna na te zmiany władzy, znajdując się w pułapce między dominującymi mocarstwami i mając ograniczoną kontrolę nad krytyczną infrastrukturą AI. Wyzwanie to nie dotyczy jednak wyłącznie UE; dotyczy ono większości krajów i społeczności na całym świecie, stawiając fundamentalne pytania o przyszłość bezpieczeństwa gospodarczego, demokratycznych rządów i suwerenności technologicznej.

5. Budowanie odporności jako długofalowa innowacja

W miarę jak możliwości sztucznej inteligencji (AI) rozwijają się i rozprzestrzeniają w społeczeństwie, utrzymanie korzyści płynących z innowacji i otwartości wymaga świadomych inwestycji w odporność w wielu wymiarach. Wyzwaniem jest nie tylko zarządzanie konkretnymi zagrożeniami, ale także zapewnienie, że systemy społeczne będą w stanie dostosować się do destrukcyjnych sił, które uwolni zaawansowana sztuczna inteligencja, i im sprostać.

Oprócz naszej własnej analizy scenariuszy, coraz częściej prowadzimy prace prognostyczne, które mogą wzmocnić planowanie odporności. Scenariusze zarządzania sztuczną inteligencją OECD [14] i „ AI 2027” [15] dostarczają scenariusze przewidujące wyzwania na różnych ścieżkach rozwoju. Mogą one pomóc decydentom zidentyfikować luki i opracować adaptacyjne rozwiązania, zanim pojawią się kryzysy.

Ta potrzeba odporności obejmuje kilka kluczowych obszarów, w tym systemy gospodarcze i rynki pracy, instytucje edukacyjne, ekosystemy informacyjne i stabilność geopolityczną. W szczególności w cyberbezpieczeństwie modele o otwartej wadze napędzają zdecentralizowaną innowację i rozszerzają dostęp do najnowocześniejszych możliwości [16], ale jak zauważono w Międzynarodowym Raporcie Naukowym o Bezpieczeństwie Zaawansowanej Sztucznej Inteligencji (International Scientific Report on the Safety of Advanced AI), powszechny dostęp do zaawansowanych modeli może destabilizować zdolności obronne, jeśli systemy ochronne nie będą w stanie dotrzymać kroku. Rozwiązaniem jest budowanie adaptacyjnych mechanizmów obronnych, które mogą ewoluować wraz z pojawiającymi się zagrożeniami, a nie tylko kontrolowanie dostępu.

Być może najważniejsze jest to, że odporność instytucjonalna wymaga struktur zarządzania, które potrafią szybko dostosowywać się do zmian technologicznych. Tempo rozwoju sztucznej inteligencji wymaga bardziej sprawnych instytucji, zdolnych do ciągłego uczenia się i dostosowywania, wykraczających poza tradycyjne podejścia regulacyjne, zakładające względnie stabilne środowisko technologiczne.

Ironicznie rzecz ujmując, sama sztuczna inteligencja może okazać się skutecznym narzędziem do budowania tych zabezpieczeń, począwszy od automatycznego wykrywania zagrożeń [17] do spersonalizowanych systemów edukacji. Wymaga to jednak świadomej koordynacji między sektorem publicznym i prywatnym, stałych inwestycji w infrastrukturę odpornościową oraz uznania, że celem nie jest zapobieganie zmianom, lecz zapewnienie społeczeństwu możliwości rozwoju w trakcie transformacji.

Wnioski : Zamykające się okno możliwości

Zakres możliwych scenariuszy przyszłości AI pozostaje niezwykle szeroki. Postęp może gwałtownie przyspieszyć, wywołując destabilizację w wielu dziedzinach, lub może postępować bardziej stopniowo, wciąż fundamentalnie transformując społeczeństwo. Nie ulega wątpliwości, że decyzje podejmowane dziś w znacznym stopniu wpłyną na to, które scenariusze przyszłości staną się możliwe lub prawdopodobne.

Trudno przecenić wagę tych decyzji. Odpowiedzialny rozwój może przynieść ludzkości niezwykłe korzyści – od przełomów naukowych i dobrobytu gospodarczego po zwiększenie potencjału ludzkiego i skuteczność instytucjonalną. Z drugiej strony, niespójny rozwój lub wdrożenie może stwarzać poważne ryzyko, zagrażające kluczowym wartościom i instytucjom, a nawet prowadzić do utraty kontroli nad systemami sztucznej inteligencji.

Okno na kształtowanie tych rezultatów szybko się kurczy. Wraz z rozwojem możliwości sztucznej inteligencji i przyspieszeniem wdrażania, maleją możliwości ustanowienia fundamentalnych ram zarządzania, standardów bezpieczeństwa i adaptacji społecznych. Ta pilna potrzeba wymaga natychmiastowych, skoordynowanych działań w sektorze publicznym i prywatnym.

Odwołania i źródła

[1] Draghi, M. „Przyszłość europejskiej konkurencyjności. Część A: Strategia konkurencyjności dla Europy". Urząd Publikacji Unii Europejskiej, 2025

[2] Epoch AI, „Machine Learning Trends", Epoch AI, 11 kwietnia 2023 r., https://epochai.org/trends , dostęp 27 czerwca 2025 r.

[3] Ho, Anson, „Postęp algorytmiczny w modelach językowych", Epoch AI, 12 marca 2024 r., https://epochai.org/blog/algorithmic-progress-in-language-models , dostęp 27 czerwca 2025 r.

[4] Owen, David, „Interviewing AI Researchers on Automation of AI R&D”, Epoch AI, 27 sierpnia 2024 r., https://epoch.ai/blog/interviewing-ai-researchers-on-automation-of-ai-rnd , dostęp 27 czerwca 2025 r.

[5] Notopoulos, Katie, „Dramat Duolingo podkreśla nowy korporacyjny balans w szumie wokół sztucznej inteligencji”, Business Insider, 15 maja 2025 r., https://www.businessinsider.com/ai-messaging-backlash-duolingo-shopify-controversy-2025-5?international=true&r=US&IR=T , dostęp 27 czerwca 2025 r.

[6] Goh, Ethan i in., „Wpływ dużego modelu językowego na rozumowanie diagnostyczne: randomizowane badanie kliniczne”, JAMA Network Open, tom 7, nr 10, 28 października 2024 r., https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2825395 , dostęp 27 czerwca 2025 r.

[7] Poniewozik, James, „TV's War with the Robots Is Already Here”, The New York Times, 10 maja 2023 r., https://www.nytimes.com/2023/05/10/arts/television/writers-strike-artificial-intelligence.html , dostęp 27 czerwca 2025 r.

[8] Reuters, „US Updates Export Curbs on AI Chips and Tools to China”, Reuters, 30 marca 2024 r., https://www.reuters.com/technology/us-commerce-updates-export-curbs-ai-chips-china-2024-03-29 , dostęp 27 czerwca 2025 r.

[9] METR, „Szczegóły wstępnej oceny O3 i O4 przez METR”, Zasoby oceny autonomii METR, 16 kwietnia 2025 r., https://metr.github.io/autonomy-evals-guide/openai-o3-report/#reward-hacking-examples , dostęp 27 czerwca 2025 r.

[10] Rząd Wielkiej Brytanii, „Międzynarodowy raport o bezpieczeństwie sztucznej inteligencji 2025”, GOV.UK, 18 lutego 2025 r., https://www.gov.uk/government/publications/international-ai-safety-report-2025 , dostęp 27 czerwca 2025 r.

[11] Janků, David; Reddel, Max; Yampolskiy, Roman; i Hausenloy, Jason, „We Have No Science of Safe AI”, cfg.eu, 13 czerwca 2025, https://cfg.eu/we-have-no-science-of-safe-ai , dostęp 4 lipca 2025.

[12] Bloomberg, „Protest AI w siedzibie OpenAI w San Francisco koncentruje się na pracy wojskowej”, Bloomberg Newsletter, 13 lutego 2024 r., https://www.bloomberg.com/news/newsletters/2024-02-13/ai-protest-at-openai-hq-in-san-francisco-focuses-on-military-work , dostęp 27 czerwca 2025 r.

[13] Rząd Wielkiej Brytanii, „Międzynarodowy raport o bezpieczeństwie sztucznej inteligencji 2025, sekcja 2.3.3”, GOV.UK, https://www.gov.uk/government/publications/international-ai-safety-report-2025/international-ai-safety-report-2025#systemic-risks , dostęp 25 czerwca 2025 r.

[14] OECD, „2024 OECD Global Strategy Group”, Wydarzenia OECD, 15 października 2024 r., https://www.oecd.org/en/events/2024/10/2024-oecd-global-strategy-group0.html , dostęp 27 czerwca 2025 r.

[15] AI 2027, „AI 2027: Scenariusze rozwoju AI”, AI 2027, https://ai-2027.com , dostęp 27 czerwca 2025 r.

[16] Rząd Wielkiej Brytanii, „Międzynarodowy raport o bezpieczeństwie sztucznej inteligencji 2025”, GOV.UK, 18 lutego 2025 r., https://www.gov.uk/government/publications/international-ai-safety-report-2025 , dostęp 27 czerwca 2025 r.

[17] Nagel, Johannes, Lindegaard, Marius i Graabak, Jakob, „Strengthening AI Trustworthiness”, Center for Future Generations, 22 maja 2025 r., https://cfg.eu/strengthening-ai-trustworthiness , dostęp: 27 czerwca 2025 r.