Fides, Ratio et Patria. Studia Toruńskie 19/2023
ISSN 2392-1463; e-ISSN 2956-3267
s. 297-324

https://doi.org/10.56583/frp.2562

Krzysztof Michalski[*]

Politechnika Rzeszowska
https://orcid.org/0000-0002-2089-2160

Rafał Lizut[**]

Katolicki Uniwersytet Lubelski
Akademia Kultury Społecznej i Medialnej w Toruniu
https://orcid.org/0000-0002-6067-1469

Generatywna sztuczna inteligencja (GenAI/AGI) – zagrożenia, wymagania i wyzwania dla bezpieczeństwa

Streszczenie

W artykule przeglądowym omówiono obawy związane z żywiołowym rozwojem i upowszechnianiem systemów bazujących na generatywnej sztucznej inteligencji (GenAI) oraz ich wpływem na bezpieczeństwo, a także dotychczasowe wysiłki regulacyjne zmierzające do narzucenia ograniczeń w rozwijaniu i stosowaniu takich rozwiązań w trosce o poszanowanie podmiotowości człowieka oraz ważnych ludzkich dóbr oraz ich rezultaty. Rozbieżności w percepcji, ocenie i akceptacji zagrożeń wynikające z odmiennych wizji bezpieczeństwa są czynnikiem utrudniającym polityczny konsensus w kwestiach nieodzowności i zasięgu interwencji regulacyjnych oraz pilności ich podjęcia, co przejawia się w analizowanych próbach regulacyjnych. Autorzy stawiają tezę, że redukowanie bezpieczeństwa związanego z GenAI tylko do cyberbezpieczeństwa samo w sobie stanowi zagrożenie. W konkluzjach autorzy postulują zainicjowanie edukacji dotyczącej przedmiotu tak, aby było możliwe zintensyfikowanie społecznego i politycznego dialogu oraz zacieśnienie współpracy decydentów, przemysłów do-

[*] Krzysztof Michalski – doktor habilitowany, profesor na Wydziale Zarządzania Politechniki Rzeszowskiej, naukoznawca i metodolog specjalizujący się w postnormalnych formatach działalności naukowej (m.in. ocena technologii, badania bezpieczeństwa, foresight&forecasting, inter- i transdyscyplinarne przedsięwzięcia badawcze, problemy zawiłe) oraz metodyce planowania, e-mail: michals@prz.edu.pl

[**] Rafał Lizut – Katedra Informatyki Stosowanej KUL; wykładowca w AKSiM, email: lizut@kul.pl

starczających rozwiązania z zakresu GenAI i użytkujących takie rozwiązania, a także przedstawicieli społeczeństwa obywatelskiego w celu wyznaczenia priorytetów dla dalszego postępu na polu GenAI gwarantujących, że technologie te będą rozwijane i użytkowane wyłącznie w imię (integralnie pojętego) dobra człowieka.

Słowa kluczowe: *algorytmy samouczące, przemysł IT, polityka technologiczna, ocena technologii, prawa człowieka, bezpieczeństwo danych i ochrona prywatności*

GENERATIVE ARTIFICIAL INTELLIGENCE (GenAI) – THREATS, SECURITY REQUIREMENTS AND CHALLENGES

Abstract

The review article discusses concerns related to the rapid development and widespread adoption of systems based on generative artificial intelligence (GenAI) and their impact on security, as well as ongoing regulatory efforts aimed at imposing limitations on the development and use of such solutions in order to respect human dignity and important human goods and their outcomes. Discrepancies in the perception, assessment, and acceptance of threats arising from different visions of security are factors complicating political consensus on the necessity and scope of regulatory interventions and the urgency of their implementation, as evidenced by the analyzed regulatory attempts. The authors claim that reducing GenAI-related security only to cybersecurity in itself poses a threat. In the conclusions, the authors advocate for initiating education on the subject to intensify social and political dialogue and strengthen cooperation among decision-makers, industries providing GenAI solutions and using such solutions, as well as representatives of civil society, to set priorities for further progress in the field of GenAI, ensuring that these technologies will be developed and used solely for the sake of (integrally understood) human good.

Keywords: *self-learning algorithms, IT industry, technology policy, technology assessment, human rights, data security and privacy protection*

~ • ~

Wstęp

Określenie „generatywna sztuczna inteligencja" (w skrócie: GenAI albo AGI), używane często zamiennie z określeniem „sztuczna inteligencja ogólnego przeznaczenia"[1], odnosi się do kategorii algorytmów, które generują nowe wy-

[1] Projekt unijnej ustawy o sztucznej inteligencji definiuje „system sztucznej inteligencji ogólnego przeznaczenia" jako taki, który w zamierzeniu dostawcy ma wykonywać powszechnie stosowane funkcje, takie jak rozpoznawanie obrazu i mowy, generowanie dźwięku i obrazu, wykrywanie wzorców, odpowiadanie na pytania, tłumaczenie z jednego języka na inny

niki w oparciu o dane wejściowe, na których zostały przeszkolone. Pierwsze lata doświadczeń z generatywną sztuczną inteligencją uświadomiły ludzkości, jak wielki potencjał rozwojowy, użytkowy i ekspansyjny drzemie w algorytmach samouczących się, zdolnych do generowania nowych danych (tekstowych, graficznych, dźwiękowych, programistycznych i in.) o cechach wzorowanych na strukturach wejściowych danych szkoleniowych[2]. Do systemów generatywnych sztucznej inteligencji należą np. chatboty oparte na dużych modelach językowych (LLM) takie jak ChatGPT, Copilot, Gemini czy LLaMA, systemy generowania obrazu przekształcające tekst na obraz, takie jak Stable Diffusion, Midjourney i DALL-E oraz generatory przetwarzające tekst na wideo, takie jak Sora[3]. Możliwości generatywnego systemu AI zależą od modalności lub rodzaju użytego zbioru danych uczenia maszynowego. Generatywna sztuczna inteligencja może być jednomodalna (takie systemy przyjmują tylko jeden rodzaj danych wejściowych, np. tylko dane tekstowe) albo multimodalna[4] (na przykład wersja GPT-4 OpenAI akceptuje zarówno tekst, jak i obraz czy wideo).

Generatywna sztuczna inteligencja ma szerokie spektrum zastosowań – od tworzenia oprogramowania[5] i generowania syntetycznych danych[6], poprzez autorstwo

język itp. Definicja ta ma zastosowanie niezależnie od tego, w jaki sposób system jest wprowadzany na rynek lub oddawany do użytku (np. jako oprogramowanie typu *open source*). System GenAI może być również systemem uniwersalnym, używanym w wielu kontekstach i zintegrowanym z wieloma innymi systemami AI, por. B. McElligott, *ChatGPT and the EU AI Act*, https://www.mhc.ie/latest/insights/chatgpt-and-the-eu-ai-act [dostęp: 29.10.2023].

2 A. Pasick, *Artificial Intelligence Glossary: Neural Ne-tworks and Other Terms Explained*, „The New York Times", March 27, 2023; A. Karpathy, P. Abbeel, G. Brockman, P. Chen, V. Cheung, Y. Duan, I. Goodfellow, D. Kingma, J. Ho, R. Houthooft, T. Salimans, J. Schulman, I. Sutskever, W. Zaremba, *Generative models*. "OpenAI", June 16, 2016, https://openai.com/research/generative-models [dostęp: 28.11.2023].

3 Zob. C. Metz, *OpenAI Plans to Up the Ante in Tech's A.I. Race*, „The New York Times" March 14, 2023; K. Roose, *A Coming-Out Party for Generative A.I., Silicon Valley's New Craze*, „The New York Times" October 21, 2022.

4 A. Islam, *A History of Generative AI: From GAN to GPT-4*, „Marktechpost" March 21, 2023; https://www.marktechpost.com/2023/03/21/a-history-of-generative-ai-from-gan-to-gpt-4/ [dostęp: 18.11.2023].

5 Oprócz tekstu w języku naturalnym duże modele językowe można trenować na tekście w języku programowania, co umożliwia generowanie kodu źródłowego dla nowych programów komputerowych. Zob. np. I. Paz, *16 najlepszych generatorów kodów AI*, https://www.morningdough.com/pl/ai-tools/best-ai-code-generators/, [dostęp: 28.11.2023].

6 Generowanie danych syntetycznych jako alternatywy dla danych rzeczywistych znajduje ostatnio coraz szersze zastosowania w walidacji i kalibracji modeli matematycznych albo w trenowaniu algorytmów uczenia maszynowego przede wszystkim ze względu na rosnące ograniczenia prawne w korzystaniu z tych drugich (np. wymagania ochrony danych spersonalizowanych/ochrony prywatności albo wynikające z niejawności informacji), ale także w procesach wymagających danych o odpowiedniej jakości, zob. S. Owen, *Synthetic Data for Better Machine Learning*, data-bricks.com, April 12, 2023 [dostęp: 14.12.2023]; H. Sharma, *Synthetic Data Platforms: Unlocking the Power of Generative AI for Structured Data*, kdnuggets.com, July 11, 2023 [dostęp: 14.12.2023].

tekstów[7] (również naukowych), sztukę (m.in. muzykę[8], film[9] czy inne sztuki wizualne[10]), finanse, sprzedaż i marketing[11], obsługę klienta[12], aż po planowanie generatywne[13]

[7] Najczęściej stosowane generacyjne systemy AI szkolone na słowach lub tokenach słów, takie jak GPT-3, LaMDA, LLaMA, BLOOM, GPT-4, Gemini, są zdolne do przetwarzania języka naturalnego (np. parafrazowania, streszczania, kompilacji itp.), tłumaczenia maszynowego oraz syntetyzowania nowych treści języka naturalnego na podstawie wielojęzycznych zbiorów danych, takich jak BookCorpus, Wikipedia i in. Mogą być wykorzystywane samodzielnie albo jako modele bazowe do innych zadań. Zob. J. Coyle, *In Hollywood writers' battle against AI, humans win (for now)*, „Associated Press News", September 27, 2023; https://apnews.com/article/hollywood-ai-strike-wga-artificial-intelligence-39ab72582c3a15f77510c9c30a45ffc8 [dostęp: 28.11.2023].

[8] Generatywną sztuczną inteligencję można również intensywnie trenować na klipach audio, aby uzyskać naturalnie brzmiącą syntezę mowy i możliwości zamiany tekstu na mowę, czego przykładem są kontekstowe syntezatory ElevenLabs lub Voicebox Meta Platform. Aplikacje takie jak MusicLM i MusicGen można również szkolić na gotowych ścieżkach audio z opisami tekstowymi do generowania nowych próbek muzyki według instrukcji tekstowej. Można w ten sposób wygenerować nowe aranżacje dźwiękowe do tekstów - na przykład imitujące głos ulubionego wokalisty. Kompozycje, teksty i nagrania artystów są chronione prawami autorskimi, ale brzmienia głosu artystów, brzmienia instrumentów oraz style wykonań nie doczekały się dotąd ochrony prawnej przed nieautoryzowanym wykorzystaniem z użyciem generatywnych systemów AI. Z tą kwestią wiąże się wiele kontrowersji dotyczących prawa autorskiego, należnych tantiem etc. Por. A. Lamar, *Jay-Z's Delaware producer sparks debate over AI rights*, https://www.delawareonline.com/story/entertainment/2023/04/21/jay-zs-delaware-producer-young-guru-artificial-voice-generators-rights/70107389007/ [dostęp: 28.11.2023]..

[9] GenAI przeszkolona na klipach wideo z opisami tekstowymi może generować nowe wideoklipy spójne w czasie i wysoce realistyczne. Obecnie dostępny jest szeroki wachlarz generatorów wideo, np. Sora firmy OpenAI, Gen-1 i Gen-2 firmy Runway czy Make-A-Video firmy Meta Platforms.

[10] Podobnie tworzenie wysokiej jakości dzieł sztuki jest od kilku lat jedną z głównych domen generatywnej sztucznej inteligencji. Jest ona zresztą wykorzystywana do tworzenia dzieł sztuki praktycznie już od chwili powstania. Już na początku lat siedemdziesiątych XX w. Harold Cohen tworzył i wystawiał obrazy wygenerowane przez program komputerowy AARON jego autorstwa, zob. N. Bergen, A. Huang, *A Brief History of Generative AI*, Dichotomies, Generative AI, Navigating Towards a Better Future, Deloitte, 2023, Issue 2, s. 4. W nurcie sztuki generatywnej są powszechnie wykorzystywane systemy AI trenowane na zestawach obrazów z opisami tekstowymi oraz systemy zdolne do transformowania tekstu w obraz. W 2021 r. wraz z pojawieniem się na rynku DALL-E - modelu generującego piksele opartego na transformatorze - a zaraz potem Midjourney i Stable Diffusion, nastała nowa era wysokiej jakości sztuki wizualnej tworzonej przez AI na podstawie instrukcji w języku naturalnym.

[11] *Don't fear an AI-induced jobs apocalypse just yet*, „The Economist", March 6, 2023; https://www.economist.com/business/2023/03/06/dont-fear-an-ai-induced-jobs-apocalypse-just-yet [dostęp: 22.11.2023]

[12] E. Brynjolfsson, D. Li, L.R. Raymond, *Generative AI at Work* (Working Paper 31161), Working Paper Series, National Bureau of Economic Research, Cambridge MA, April 2023.

[13] Planowanie generatywne wspomagane AI wykorzystuje się w planowania procesów wspomaganych komputerowo do generowania sekwencji działań prowadzących do osiągnięcia określonego celu, np. do generowania planów zarządzania kryzysowego w sektorze cywilnym i wojskowym, planowania procesów produkcyjnych, czy wielowariantowego planowania ryzykownych przedsięwzięć technologicznych, takich jak np. załogowe misje kosmiczne.

i szeroko pojęte projektowanie[14], w tym również syntetyzowanie nowych cząsteczek, m.in. DNA[15]. Rosnącej fascynacji przełomowymi innowacjami na wielu polach działalności biznesowej czy technicznej możliwymi dzięki GenAI towarzyszą jednak obawy przed różnego typu zagrożeniami, których wachlarz sięga od nadużyć (np. do działalności cyberprzestępczej, dezinformacji, manipulowania ludźmi), poprzez masową likwidacją miejsc pracy w branżach, w których obserwuje się największą ekspansję AI, aż po spontaniczne pełzające zmiany zagrażające ludzkiej podmiotowości, autonomii, prawom i swobodom obywatelskim, obyczajom oraz innym cennym zdobyczom cywilizacyjnym.

Odkąd AI wyruszyła na podbój świata, badacze, użytkownicy i legislatorzy zgłaszali prawne, etyczne i inne zastrzeżenia dotyczące konsekwencji tworzenia sztucznych istot posiadających inteligencję podobną do ludzkiej, zdolnych do imitowania człowieka – a na wielu polach również przewyższających go pod względem sprawności działania, co grozi postępującym rugowaniem człowieka traktowanego jako źródło błędów[16] z domen o fundamentalnym znaczeniu cywilizacyjnym, takich jak choćby wymiar sprawiedliwości, diagnostyka medyczna, nadzór nad bezpieczeństwem czy działalność naukowo-techniczna. Eliminacja czynnika ludzkiego przez AI budzi obawy i zastrzeżenia zwłaszcza w obszarach regulowanych normami społeczno-etycznymi, bowiem na tym polu zdolności AI chyba najwyraźniej nie dorównują (jeszcze) zdolnościom człowieka[17]. Pierwsze

[14] H. Harreis, T. Koullias, R. Roberts, *Generative AI: Unlocking the future of fashion*, McKinsey&Company, March 8, 2023; https://www.mckinsey.com/industries/retail/our-insights/generative-ai-unlocking-the-future-of-fashion [dostęp: 28.11.2023]; T.T. Eapen, D.J. Finkenstadt, J. Folk, L. Venkataswamy, *How Generative AI Can Augment Human Creativity*, „Harvard Business Review", June 16, 2023; https://hbr.org/2023/07/how-generative-ai-can-augment-human-creativity [dostęp: 28.11.2023].

[15] Generatywne systemy AI można trenować na sekwencjach aminokwasów lub dowolnych innych molekułach. Systemy wykorzystywane w biologii syntetycznej, takie jak AlphaFold, służą m.in. do modelowania i symulowania struktury białek, przewidywania ekspresji nowych sekwencji DNA oraz projektowania nowej generacji spersonalizowanych leków. Zob. W.D. Heaven, *AI is dreaming up drugs that no one has ever seen. Now we've got to see if they work*, „MIT Technology Review", February 15, 2023; https://www.technologyreview.com/2023/02/15/1067904/ai-automation-drug-development/ [dostęp: 28.11.2023].

[16] K. Michalski, *Autonomizacja techniki i niepożądane skutki eliminowania człowieka jako źródła błędów*, „Filo-Sofija. Z Problemów Współczesnej Filozofii" 2017, 39/I (2017/4/I), tom 6: Człowiek i maszyna, s. 79-95.

[17] Pomimo postępu w dziedzinie AI istnieją fundamentalne ograniczenia jej możliwości, takie jak brak emocjonalności i empatii, wyobraźni, nielogicznego/nieschematycznego zachowania, kreatywności czy intuicji – a przede wszystkim brak specyficznego dla istot ludzkich poziomu samoświadomości i zdolności do samorefleksji. Te unikalne cechy ludzkiego umysłu stanowią kategoryczną różnicę między inteligencją maszynową a inteligencją ludzką. Konwencjonalne systemy oparte na AI mają ograniczone możliwości adaptacyjne i ograniczoną elastyczność, co jest szczególnie widoczne w dynamicznie zmieniających się środowiskach. Ograniczenia inteligencji obliczeniowej uświadomił ludzkości paradoks Moraveca, uważany za największe odkrycie w dziedzinie AI. Hans Moravec zauważył, że

próby wprowadzania AI do tych zastosowań ograniczone były przez ograniczenia samej AI, a szczególnie jej małej elastyczności przy szerokim wachlarzu możliwych typów danych wejściowych. W skrócie: AI nie była w stanie odpowiednio reagować na złożone problemy świata realnego.

Jednakże z początkiem XXI wieku zaczęto wykorzystywać głębokie sieci neuronowe zdolne do uczenia się modeli generatywnych w środowisku złożonych danych, takich jak obrazy, m.in. dzięki wykorzystaniu takich rozwiązań, jak autokoder wariacyjny i generatywna sieć kontradyktoryjna. Wcześniej sieci neuronowe były zwykle szkolone z użyciem modeli dyskryminacyjnych. W 2017 r. sieć transformatorów umożliwiła udoskonalenie modeli generatywnych w porównaniu ze starszymi modelami pamięci długo-krótkoterminowej[18], co doprowadziło do powstania w 2018 r. pierwszego wstępnie wytrenowanego transformatora generatywnego (GPT), znanego jako GPT-1. W 2023 roku META wprowadziła na rynek model sztucznej inteligencji ImageBind, który łączy dane z tekstu, obrazów, wideo, danych termowizyjnych, danych 3D, dźwięku i ruchu, co ma pozwolić na bardziej wciągające, generatywne treści AI. Funkcje generatywnej sztucznej inteligencji zostały zintegrowane z wieloma istniejącymi wcześniej, komercyjnie dostępnymi produktami, takimi jak Microsoft Office, Adobe Photoshop czy oprogramowania CAD. Wiele generatywnych modeli sztucznej inteligencji jest również dostępnych w postaci oprogramowania typu *open source* (np. Stable Diffusion czy model językowy LLaMA). Wiodącą platformą sharingową udostępniającą oprogramowania *open source* tego typu jest Hugging Face.

Mniejsze generatywne modele sztucznej inteligencji zawierające do kilku miliardów parametrów można uruchamiać na smartfonach, urządzeniach wbudowanych i komputerach osobistych, natomiast większe modele z dziesiątkami miliardów parametrów mogą wymagać akceleratorów, takich jak chipy GPU, na przykład jak te produkowane przez firmy NVIDIA i AMD lub silnik Neural Engine stosowany w produktach Apple. Modele językowe z setkami miliardów

maszyny i algorytmy oparte na AI mogą z łatwością wykonywać zadania, które ludzie uważają za intelektualnie złożone (np. rozwiązywanie równań matematycznych lub granie w gry strategiczne), ale mają ogromne trudności z wykonywaniem zadań percepcyjnych i psychomotorycznych, które ludzie uważają za instynktowne i prostsze, takich jak rozpoznawanie obiektów czy poruszanie się w zróżnicowanym, wielodomenowym środowisku – zadań, z którymi świetnie radzi sobie roczne dziecko. Zob. H. Moravec, *Mind Children*, Boston 1988. W związku z tym wśród kognitywistów coraz mniej zwolenników znajduje teza, że myślenie można sprowadzić do przetwarzania symbolicznego. Zasadnicza różnica pomiędzy inteligencją maszynową a intelektem ludzkim oraz ograniczenia tej pierwszej sugerują, że choć jej wpływ na życie indywidualne i zbiorowe będzie szybko wzrastał, to jednak prawdopodobnie nie zdetronizuje ona w najbliższym czasie ludzkiego umysłu.

[18] Por. Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P.S. Yu, L. Sun, *A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT*, arXiv:2303.04226 [cs.AI], March 7, 2023,

parametrów, takie jak GPT-4 lub PaLM, zwykle działają na komputerach w centrach danych wyposażonych w układy procesorów graficznych (takich jak H100 firmy NVIDIA) lub chipy akceleratora AI (takie jak TPU firmy Google). Te bardzo duże modele są zazwyczaj dostępne jako usługi w chmurze.

Zagrożenia i obawy związane z ekspansją GenAI

Żywiołowa ekspansja GenAI na coraz więcej sfer życia budzi wiele obaw i zastrzeżeń, bowiem w wyścigu na innowacje firmy wprowadzają nowe produkty na rynek, zanim nauka zdąży w porę rozpoznać potencjały oddziaływań tych produktów na jednostkę i społeczeństwo[19]. Część z obaw odnosi się do problemów wspólnych technologicznym innowacjom każdego typu, do których odnosi się paradoks Collingridge`a[20] i które mają swoich zyskujących i tracących. Ponieważ

[19] Zaostrzająca się globalna konkurencja czyni z innowacyjności wysoko ceniony czynnik przewagi w biznesie, zmuszając przedsiębiorstwa do pochopnego wdrażania innowacji, zanim nauka w pełni rozpozna ich konsekwencje. Ponieważ wiele niepożądanych skutków ubocznych innowacyjnych produktów ujawnia się dopiero w późniejszych fazach ich cyklu życia po tym, gdy przedsiębiorstwo poniosło już określone koszty inwestycyjne, które chciałoby możliwie szybko zamortyzować, zrozumiała jest niechęć firm do rezygnacji z zysków czerpanych ze sprzedaży takich produktów wynikająca z obawy przed stratami finansowymi i utratą udziałów w rynku. Pokusa nieodpowiedzialności jest tym większa, im większe są wątpliwości co do istnienia niepodważalnych naukowych dowodów potwierdzających szkodliwość produktów. Producenci wiedzą, jakie są rzeczywiste możliwości poznawcze nauki, więc zamiast angażować pomysły, siły i środki w udoskonalanie produktów i poprawę ich bezpieczeństwa, dokładają wszelkich starań, aby ewentualne społecznie niepożądane skutki ich działalności pozostawały niewykrywalne dla nauki albo ich znajomość pozostała nieznana dla pozostałej części społeczeństwa. Zamiast wykosztowywać się na zapewnianie zgodności produktów z obowiązującymi wymaganiami i społecznymi oczekiwaniami, przedsiębiorstwa wolą przeznaczać stosunkowo niewielkie sumy na wynagrodzenia dla lobbystów, którzy wywierając wpływ na procesy polityczne (np. stanowienie prawa) o wiele skuteczniej zadbają o to, aby właściwe produkty trafiły do sprzedaży na mocy specjalnie w tym celu wprowadzonych regulacji. Inną opcją jest przeznaczanie pewnej kwoty na odszkodowania, co jest mniej kosztowne niż usuwanie usterek czy wycofywanie produktu. Ponieważ operatorzy niebezpiecznych instalacji przemysłowych i dostawcy produktów mogą w wielu krajach świata liczyć na przychylność władz, które czerpią profity z opodatkowania takiej działalności, a na zagrożenia z przedwczoraj rozpoznane wczoraj zareagują jutro, natomiast ich ewentualne interwencje spowodują skutki dopiero pojutrze, analiza istniejących uwarunkowań skłania raczej do pesymistycznego wniosku o istnieniu zaskakującej konstelacji interesów w utrzymaniu bezpieczeństwa systemów technicznych na najniższym możliwym poziomie. Zob. K. Michalski, M. Jurgilewicz, *Konflikty technologiczne. Nowa architektura zagrożeń w epoce wielkich wyzwań*, Warszawa 2021, s. 98-102.

[20] Paradoks Collingridge`a zwany także dylematem kontroli polega na tym, że we wczesnej fazie rozwoju innowacyjnej technologii, gdy możliwe jest sterujące oddziaływanie na nią, brakuje potrzebnej do tego wiedzy o konsekwencjach jej zastosowań czerpanej z doświadczenia, a gdy taka wiedza wreszcie się pojawi, jest zwykle za późno na interwencje sterujące, bowiem technologia żyje już własnym życiem i dokonuje ekspansji zgodnie z własną logiką. Zob. D. Collingridge, *The Social Control of Technology*, London 1980.

te uniwersalne zagadnienia mają obszerne i kompleksowe opracowania[21], nie warto poświęcać im tutaj specjalnej uwagi z powodu ograniczeń objętościowych artykułu. Warto natomiast skupić się na zagrożeniach nowych, specyficznych dla GenAI, wynikających z jej dotychczasowych lub przewidywalnych zastosowań. Chodzi więc o kompleks globalnych problemów, które są nowe i inne niż zagrożenia stwarzane dotąd przez technologie jądrowe, chemiczne, bio- lub nanotechnologie i wobec których w związku z tym tradycyjne sposoby zapewniania bezpieczeństwa (regulacje prawne, rozwiązania organizacyjno-instytucjonalne, procedury i praktyki etc.) wydają się być w wielu aspektach niewydolne.

Żywiołowy rozwój GenAI wzbudza obawy zarówno rządów, przedsiębiorstw, jak i osób prywatnych, ponieważ na świecie rośnie świadomość przełomowości tej technologii oraz jej potencjału transformacyjnego. Obawy te celnie wyartykułował Sekretarz Generalny ONZ António Guterres w trakcie posiedzenia Rady Bezpieczeństwa Organizacji Narodów Zjednoczonych w lipcu 2023 r., stwierdzając, że „GenAI ma ogromny potencjał wyrządzania dobra i zła na dużą skalę" i że AI może „turbodoładować" globalny wzrost gospodarczy i już do roku 2030 wygenerować od 10 do 15 bilionów dolarów wartości dodanej w globalnej gospodarce, ale jej złośliwe użycie „może spowodować przerażające zniszczenia, powszechną traumę i głębokie szkody psychiczne na niewyobrażalną skalę"[22]. Poczucie zagrożenia w obliczu rosnącej potęgi AI oraz w obliczu umacniającej się pozycji globalnych gigantów technologicznych dostarczających produkty z tego asortymentu mobilizuje coraz więcej środowisk do protestów oraz działań prawnych. Coraz głośniej słychać też wezwania do nałożenia globalnych restrykcyjnych ograniczeń na deweloperów GenAI albo nawet całkowitego wstrzymania eksperymentów nad AI. Tonacja tych obaw, sprzeciwów i wynikających z nich postulatów nie jest jednak jednorodna, co wynika z przeciwstawnych interesów oraz odmiennych politycznych wizji i kultur bezpieczeństwa, przez pryzmat których rozpatruje się konsekwencje upowszechniania GenAI.

Zagrożenia dla cyberbezpieczeństwa

Jednym z obszarów wpływu GenAI na bezpieczeństwo jest cyberbezpieczeństwo. Skojarzenie GenAI z cyberbezpieczeństwem wydaje się oczywiste, jako że zarówno ich geneza, jak i obszar działania, to przestrzeń cyfrowa. Nauki o cyberbezpieczeństwie są dyscypliną rozległą, angażującą wiedzę zarówno sprzętową

[21] Zob. np. K. Michalski, *Technology Assessment. Ocena technologii – nowe wyzwania dla filozofii nauki i ogólnej metodologii nauk*, Rzeszów 2019.

[22] Zob. A. Guterres, *UNO-Secretary-General's remarks to the Security Council on Artificial Intelligence*, https://www.un.org/sg/en/content/sg/speeches/2023-07-18/secretary-generals-remarks-the-security-council-artificial-intelligence, [dostęp: 28.11.2023].

i programową, jaki i nauki o człowieku jako twórcy, użytkowniku i protektorze świata cyfrowego.

Przedmiotem nauk o cyberbezpieczeństwie jest bezpieczeństwo informacji. Jego główne priorytety to tzw. CIA (ang. Confidentiality, Integrity, Availability) czyli „zachowanie poufności, integralności i dostępności informacji". Wywiązanie się z tych zadań ważne jest zarówno dla poszczególnych ludzi, ale również dla przedsiębiorstw, szczególnie tych, które zarządzają danymi swoich klientów/użytkowników. W przeciwnym wypadku grozi im nie tylko utrata danych, ale poniesienie konsekwencji prawnych, materialnych i wizerunkowych za niewłaściwą ochronę tych danych. Zalecenia i wytyczne odnoszące się do bezpieczeństwa informacji ujednolicono w standardach ISO 27000, które są na bieżąco aktualizowane i uzupełniane w celu pomocy organizacjom dowolnej wielkości i branży w systematycznej i rozsądnej kosztowo ochronie informacji poprzez przyjęcie zintegrowanego Systemu Zarządzania Bezpieczeństwem Informacji (zgodnie z ISO/IEC 27001). Organizacje stosujące te standardy wzmacniają swoją zdolność do ochrony przed cyberatakami i zapobiegają niepożądanemu dostępowi do wrażliwych i poufnych informacji oraz ich utracie.

Wraz z pojawieniem się GenAI część celów polityk cyberbezpieczeństwa stała się trudniejsza do osiągnięcia. Poufność informacji jest zagrożona przez cyberataki, które z jednej strony mogą mieć charakter techniczny, gdzie GenAI może generować na podstawie *scrapingu* z mediów społecznościowych zestawów prawdopodobnych haseł do ataków słownikowych w celu uzyskania nieuprawnionej autoryzacji, czy też wykorzystania różnego rodzaju *exploitów* powstałych przy tworzeniu narzędzi do przechowywania, przetwarzania i przesyłania danych. Niedawne badania przeprowadzone w 2023 r. wykazały, że GenAI ma słabe punkty, którymi przestępcy mogą manipulować w celu wydobycia szkodliwych informacji z pominięciem zabezpieczeń. W badaniu przedstawiono przykładowe ataki przeprowadzane na ChatGPT, w tym Jailbreaks i psychologię odwrotną. Ponadto złośliwe osoby mogą wykorzystywać ChatGPT do ataków socjotechnicznych i ataków phishingowych, ujawniając szkodliwą stronę tych technologii. GenAI rozszerza możliwości *spoofingu* czy *phishingu*. Spoofing rozumiany był do tej pory jako podszywanie się pod czyjś email, numer telefonu podczas przekazywania wiadomości tekstowych czy pod numer telefonu instytucji, takiej jak np. bank, gdzie odbiorca informacji nie znał głosu dzwoniącej osoby. Obecnie GenAI może tworzyć nie tylko treści, ale imitować głos a nawet czyjś wygląd na podstawie próbek głosu czy zdjęć. W czasach cyfrowego ekshibicjonizmu te próbki są powszechnie dostępne. Zadzwonienie więc do kogoś z numeru osoby jej znanej i posłużenie się głosem lub wyglądem takiej osoby w celu wyłudzenia informacji pozwalającej na nieuprawnioną autoryzację jest możliwe i finansowo osiągalne. Phishing został zaś wzmocniony o masowe tworzenie spersonalizowa-

nych informacji dobranych do odbiorców na podstawie informacji dostępnych na przykład w mediach społecznościowych. Problem dotyczy zarówno osób prywatnych, gdzie wyłudzone dane służą do przejmowania kont, ale też instytucji, gdzie pracują ludzie, którzy są zazwyczaj najsłabszym ogniwem sieci bezpieczeństwa cybernetycznego.

Integralność informacji ma zabezpieczać ich prawdziwość, czyli stan relacji między tym, co prezentują a tym, co reprezentują. Naruszenie tej relacji, choć ważne, nie jest jednak aż tak istotne, jak jego konsekwencje w postaci wpływu na działanie osób posługujących się nimi. Ponieważ ludzie działają zgodnie z tym, co wiedzą, zmiana tej wiedzy na jakiś temat skutkuje zmianami działań lub postaw. Przykładowo manipulacje za pomocą Deepfake[23] wytworzonych przez GenAI mają na celu zmianę zachowań tych, którzy są ich odbiorcami. Wojna w Ukrainie pokazała nieznane dotąd możliwości GenAI wykorzystywane do modelowania przekonań i postaw społecznych.

GenAI może zagrozić również dostępności danych i to w dwojaki sposób. Po pierwsze, umożliwiając nieautoryzowany dostęp, grozi ich usunięciem lub zmodyfikowaniem. Po drugie, może nastąpić taka produkcja danych fałszywych, że prawdziwe staną się niewidoczne dla wyszukującego oraz niemożliwe do zweryfikowania. Z tym zjawiskiem ściśle współdziała *opinio communis*, że prawda leży gdzieś pośrodku, co skutkuje uznaniem za prawdziwe poglądów odpowiednio często propagowanych w sieci. GenAI jest na masową skalę wykorzystywana jako narzędzie takiego ilościowego uwierzytelniania.

Zagrożenia dla opisywanych filarów bezpieczeństwa są wielowymiarowe. Na przykład nadzór nad użytkownikiem zagraża poufności danych go dotyczących. Dzieje się to ze względu na możliwość włamań do stref zastrzeżonych, przechowujących dane tej osoby. Innym przykładem jest ciągła inwigilacja możliwa dzięki narzędziom informatycznym do śledzenia i zbierania informacji o każdym użytkowniku. Przechwycone dane można dalej wykorzystywać do dalszych działań z użyciem GenAI, na przykład spoofingu. Zasypując Internet bezwartościowymi lub zmanipulowanymi treściami można ponadto ograniczyć realny dostęp do prawdziwych lub wartościowych informacji. Cenzura następuje tu przez zasyp

[23] Deepfake – będące połączeniem słów „głębokie uczenie się" i „fałszywość" – to media generowane przez sztuczną inteligencję, które przedstawiają osobę na istniejącym obrazie lub filmie i zastępują ją podobizną innej osoby za pomocą generatywnej AI. Deepfakes wzbudziły szerokie zainteresowanie i obawy związane z ich wykorzystaniem do tworzenia filmów pornograficznych ze sfingowanym udziałem celebrytów czy też tzw. pornografii zemsty, a także do różnego rodzaju oszustw oraz szerzenia dezinformacji dla zasiania nieufności, wywołania społecznej dezorientacji albo zainicjowania społecznych konfliktów. Sztucznie generowane fałszywe obrazy z pola walki (np. dokumentujące rzekome zbrodnie) propagowane w mediach społecznościowych są powszechnie wykorzystywane przez strony konfliktu w ramach wojny hybrydowej do osłabiania albo wzmacniania morale, wywoływania wybuchów paniki lub masowych dezercji.

informacji lub przez wychwytywanie i usuwanie niechcianych treści. To też są domeny idealnie dostosowane do możliwości GenAI. Warto wspomnieć również o cyberpsychologii, która pozwala przy użyciu narzędzi AI określać profile psychologiczne i dobierać narzędzia kierowania osobami o znanych profilach psychologicznych. Z pomocą GenAI można również na wielką skalę tworzyć fikcyjne osoby (fake persona), będące agregatem cech osób obserwowanych. Zdolność GenAI do tworzenia realistycznych fałszywych treści jest wykorzystywana w wielu rodzajach cyberprzestępczości, m.in. w phishingu. Do dezinformacji i oszustw wykorzystuje się na masową skalę spreparowane pliki wideo i audio typu Deepfake. Ponadto modele wielkojęzyczne i inne formy sztucznej inteligencji generującej tekst stosuje się na szeroką skalę w celu tworzenia fałszywych recenzji w witrynach handlu elektronicznego w celu zwiększenia zaufania do produktów lub sprzedawców. Takie fikcyjne tożsamości mogą z powodzeniem posłużyć do działań propagandowych i kształtowania opinii (trolling), jeśli ich liczebność zapewni przewagę określonym poglądom albo uda się takie osoby wyposażyć w społeczny autorytet. Fikcyjne osoby mogą również być – w dużym stopniu bezkarnie – wykorzystywane do całkowicie realnych złośliwych działań (przestępczych, terrorystycznych, szpiegowskich, antypaństwowych itp.) – na przykład wyłudzania danych lub pieniędzy, absorbowania organów państwa fikcyjnymi skargami lub wnioskami w celu ich przeciążenia, a nawet doprowadzenia do krytycznej niewydolności albo też wyłudzania określonych uprawnień lub dostępu do cennych zasobów lub systemów o dużym znaczeniu dla funkcjonowania państwa lub społeczeństwa poprzez zatrudnianie takiej fikcyjnej osoby na stanowiskach pracy zdalnej.

Utożsamianie bezpieczeństwa – nawet jeśli tylko w sferze techniki - z cyberbezpieczeństwem nie wydaje się być jednak uzasadnione. Choć zarówno geneza GenAI jak i obszar jej działania to przestrzeń wirtualna, łańcuchy szkód, jakie powoduje, są jak najbardziej realne, a ich konsekwencje ponoszone są przez społeczeństwa zarówno w sposób kolektywny, jak i dystrybutywny (jednostki).

Pozacybernetyczne zagrożenia związane z ekspansją GenAI

Jedną z najczęściej zgłaszanych obaw w kontekście ekspansji GenAI jest likwidacja miejsc pracy w zawodach kreatywnych. Pod wpływem obserwacji dotychczasowych globalnych zmian na rynkach pracy[24] wzrasta świadomość, że

[24] Przykładowo szacuje się, że do kwietnia 2023 sztuczna inteligencja generująca obrazy spowodowała w Chinach likwidację 70% stanowisk pracy dla ilustratorów gier wideo, por. V. Zhou, *AI is already taking video game illustrators' jobs in China*, „Rest of World", April 11, 2023, https://restofworld.org/2023/ai-china-video-game-layoffs-illustrators/ [dostęp: 28.11.2023]; J. Carter, *China's game art industry reportedly decimated by growing AI use*, „Game Developer", April 11, 2023; https://www.gamedeveloper.com/art/china-s-game-art-industry-re-

dalsze udoskonalanie GenAI oznacza w praktyce egzystencjalne zagrożenie dla osób zajmujących się zawodowo pracą twórczą. W większości takich branż może wkrótce dokonać się to, co stało się z miejscami pracy w przemyśle najpierw pod wpływem automatyzacji produkcji, a następnie 4 Rewolucji Przemysłowej, ze stanowiskami pracy biurowej z chwilą upowszechnienia komputerów osobistych i pakietu Office czy ze stanowiskami obsługi klienta w handlu i bankowości wraz z upowszechnieniem usług e-commerce i e-banking. Problemy te nie są więc całkiem nowe i można je sprowadzić do głównego pytania, wokół którego ogniskowała się debata towarzysząca rozwojowi sztucznej inteligencji od chwili jej powstania: czy wszystkie zadania, które mogą wykonywać komputery, powinny być – z korzyścią dla człowieka i wobec wielowymiarowych różnic między komputerami a ludźmi, przede wszystkim wobec różnicy między „bezdusznie nieomylnymi" operacjami obliczeniowymi inteligentnych maszyn a ludzką fantazją, afektywnością i jakościowymi ocenami opartymi na wartościach – faktycznie przez nie wykonywane. Coraz bardziej realna staje się w tej sytuacji dystopia, której ponurą wizję nakreślił niegdyś Jeremy Rifkin[25].

Obawy przed nasileniem się uprzedzeń i dyskryminacji (rasowej, etnicznej, płciowej, światopoglądowej czy też związanej ze stanem zdrowia itp.)

Modele GenAI mają skłonność do powielania, utrwalania i wzmacniania stereotypów i uprzedzeń kulturowych presuponowanych w danych źródłowych. Np. przeszkolenie modeli generatywnych na zbiorze danych obciążonych treściami o charakterze patologicznym, rasistowskim albo seksistowskim może skutkować nieproporcjonalnym odzwierciedlaniem fałszywych albo społecznie szkodliwych wzorców lub normatywów (np. przekonań, że zawody lepiej wynagradzane i kojarzone z większym prestiżem są wyłączną domeną białych mężczyzn[26]). Coraz częściej dyskutowanym problemem w kontekście wpływu GenAI na rynki pracy jest przyszła sytuacja grup mających obecnie niedostateczny udział w globalnym rynku pracy oraz niedostatecznie reprezentowanych zwłaszcza w cenionych zawodach. Istnieje realne niebezpieczeństwo, że wykorzystanie GenAI do wspomagania procesów doboru i rekrutacji pracowników wzmocni istniejące uprze-

portedly-decimated-ai-art-use [dostęp: 22.11.2023]. Analogicznie sztuczna inteligencja generująca tekst, muzykę czy głos jest postrzegana jako poważne zagrożenie dla takich grup twórców, jak pisarze i dziennikarze, kompozytorzy, instrumentaliści, aktorzy czy lektorzy. Zagrożone są również stanowiska pracy projektantów, a nawet programistów.

25 J. Rifkin, *Koniec pracy. Schyłek siły roboczej na świecie i początek ery postrynkowej*, Wrocław 2001.

26 OpenAI, *Reducing bias and improving safety in DALL·E 2*, July 18, 2022, https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2 [dostęp: 22.11.2023].

dzenia do pewnych obszarów geograficznych i grup społecznych (np. kobiet, niepełnosprawnych, przedstawicieli określonych ras, grup etnicznych lub wyznań, osób z tzw. tłem imigracyjnym czy też osób o zbyt niskim IQ albo o określonych niepożądanych cechach wyglądu). W efekcie spodziewane są jeszcze większa koncentracja podaży zatrudnienia i dalsze przemieszczenia stanowisk pracy na globalnym rynku. Aby zapobiec wpływom uprzedzeń i zjawiskom dyskryminacyjnym, potrzeba regulacji wymuszających stosowanie różnych metod korygowania błędu systematycznego, takich jak np. zmiana wejściowego wsadu danych szkoleniowych[27] czy odpowiednie ich wagowanie[28], a także regulacji gwarantujących przejrzystość procesów rekrutacji i selekcji kandydatów, planowanie włączające i wzmacnianie spersonalizowanej edukacji w celu kompensacji dotychczasowych nierówności w dostępie do rynków pracy, zawodów i ofert[29].

Zagrożenia dla dziennikarstwa i misji środków przekazu

Nadużywanie GenAI do niskobudżetowej produkcji przekazu (kompilowanie oraz parafrazowanie newsów i innych treści) stało się od kilku lat prawdziwą plagą w dziennikarstwie, zwłaszcza internetowym. Odkąd pojawienie się tzw. nowych mediów opartych na bezpłatnym dostępie do treści zmieniło sposób finansowania jej produkcji, której koszty – zamiast z dochodów pochodzących ze sprzedaży gazet lub abonamentów – pokrywa się obecnie w główne mierze z dochodów z reklamy, nastąpiło zauważalne obniżenie lotów w profesjonalnym dziennikarstwie. Ponieważ wyceny usług reklamowych w internetowych serwisach informacyjnych zależą od popularności serwisu mierzonej liczbą odwiedzających w czasie rzeczywistym, wydawcy konkurujący między sobą o uwagę użytkowników wypełniają swoje serwisy podobnymi – optymalnie stargetowanymi – treściami i stosują podobne strategie przyciągania oparte na aktualności i kompletności treści, intrygujących lub sensacyjnie brzmiących tytułach oraz korzystnym pozycjonowaniu stron. GenAI wyświadcza w kompletowaniu treści nieocenione usługi i jest na świecie równie powszechnie, co nietransparentnie używana do produkcji dziennikarskiego chłamu – niskiej jakości artykułów pełnych błędów i niesprawdzonych, a często również fałszywych, choć realistycznie, „łudząco prawdziwie" brzmiących informacji[30]. Coraz więcej wydawców w ramach polityki jawności

[27] Por. J. Traylor, *No quick fix: How OpenAI's DALL·E 2 illustrated the challenges of bias in AI*, „NBC News" July 27, 2022.

[28] OpenAI, *DALL·E 2 pre-training mitigations*, June 28, 2022; https://openai.com/research/dall-e-2-pre-training-mitigations [dostęp: 22.11.2023].

[29] S. Gupta, *Underrepresented groups in countries around the world are worried about AI being a threat to jobs*, Fast Company, October 31, 2023; https://www.fastcompany.com/90975180/ai-threat-jobs-fears-survey-seven-countries-job-seekers-hr [dostęp: 28.11.2023].

[30] Wobec znudzenia użytkowników wszechobecnością tych samych treści i pogoni za sensacją wydawcy serwisów dziennikarskich coraz częściej posuwają się do fingowania faktów. Na

nie tylko przyznaje się do wykorzystywania GenAI do produkcji treści, ale także ujawnia standardy, jakie przy tym stosuje[31]. Nadużywanie GenAI do produkcji przekazu nie tylko skutkuje zauważalnym pogorszeniem się jakości dziennikarstwa oraz ograniczeniem pluralizmu i różnorodności mediów – pluralizmu będącego koniecznym warunkiem możliwości zdrowych procesów opiniotwórczych w społeczeństwie, stanowiących przeciwwagę dla współczesnych działań propagandowych, dezinformacyjnych i manipulacyjnych, ale także degradacją profesji pełniącej tradycyjnie ważne misje społeczne o dużym znaczeniu dla prawidłowego funkcjonowania demokracji (dzięki swojej działalności opiniotwórczej media są wszak uznawane za czwartą władzę!). Poważny problem w tym kontekście – powszechne zjawisko pasożytnictwa deweloperów GenAI i wydawców wykorzystujących algorytmy do produkcji treści, wiążące się z naruszaniem praw autorskich twórców treści wykorzystywanych do trenowania takich algorytmów – z racji swojej kompleksowości wymaga osobnego omówienia.

Naruszenia praw autorskich

Generacyjne systemy AI są szkolone na dużych, powszechnie dostępnych zbiorach danych, z których część stanowią dzieła chronione prawem autorskim. Duże firmy technologiczne budują swoje modele biznesowe i uzyskują korzyści finansowe korzystając z treści, których powstanie wymaga zaangażowania twórców, wydawców czy dziennikarzy. Właściciele praw autorskich do dzieł wykorzystywanych przez Big Tech do trenowania generatywnych algorytmów uznają to za naruszenie ich praw i domagają się pilnego wprowadzenia regulacji zapewniających ochronę ich interesów[32]. Skoro narzędzia AI przynoszące gigan-

przykład w kwietniu 2023 r. niemiecki tabloid „Die Aktuelle" opublikował sfingowany, wygenerowany przez GenAI wywiad z byłym kierowcą wyścigowym Michaelem Schumacherem, który od 2013 r. nie występował publicznie po urazie mózgu w wypadku na nartach. Por. R. Traisman, *A magazine touted Michael Schumacher's first interview in years. It was actually AI*, https://www.npr.org/2023/04/28/1172473999/michael-schumacher-ai-interview-german-magazine [dostęp: 28.11.2023].

31 Zob. np. K. Viner, A. Bateson, *The Guardian's approach to generative AI*, „The Guardian" June 16, 2023, https://web.archive.org/web/20240103230448/https:/www.theguardian.com/help/insideguardian/2023/jun/16/the-guardians-approach-to-generative-ai [dostęp: 28.11.2023].

32 Ch.T. Zirpoli, *Generative Artificial Intelligence and Copyright Law*, Congressional Research Service. LSB10922, September 29, 2023, https://crsreports.congress.gov/product/pdf/LSB/LSB10922 [dostęp: 07.12.2023]. W Polsce organizacje twórców, dziennikarzy i wydawców wystąpiły w marcu 2024 r. z apelem do premiera ws. uczciwych warunków korzystania z wytwarzanych przez nich treści w związku z projektem ustawy implementującej dyrektywy UE o prawie autorskim i prawach pokrewnych. Zaproponowane przepisy wzmacniają pozycję największych międzynarodowych firm technologicznych i nie chronią interesów twórców przed bezumownym korzystaniem z ich wytworów. W przygotowanym przez polski rząd projekcie noweli ustawy o prawie autorskim i prawach pokrewnych próbuje się przemycić

tom technologicznym spore zyski funkcjonują dzięki wytworom czyjejś pracy, oczekiwanie wypłaty godziwego wynagrodzenia za tą pracę wydaje się być uzasadnione. Biorąc dodatkowo pod uwagę fakt, że treści (np. teksty, obrazy, muzyka itp.) wygenerowane przez AI wykazują daleko idące podobieństwa do objętych prawami autorskimi treści użytych do szkolenia, z którymi te pierwsze zwykle konkurują, żądania klarownych uregulowań ograniczających bezumowne korzystanie przez deweloperów i użytkowników GenAI z owoców cudzej twórczości są przekonujące. Zwolennicy zachowania obecnej swobody korzystania z ogólnodostępnych treści przekonują jednak, że – po pierwsze – taki sposób wykorzystania treści wiąże się z ich przekształcaniem (w różnym zakresie) i nie polega na zabronionym prawnie publicznym udostępnianiu oryginalnych dzieł oraz że – po drugie – wytwory GenAI w świetle przepisów obowiązujących w większości krajów same również nie posiadają zdolności do ochrony prawnej[33] i mogą być

niekorzystne dla twórców zapisy dotyczące wynagradzania twórców za ich treści wykorzystywane do szkolenia algorytmów. Wynika z nich, że duże firmy technologiczne w dalszym ciągu będą mogły za darmo trenować swoje algorytmy na treściach twórców, czerpiąc z tego zyski, bowiem rząd usunął zapisy zawarte w pierwszej wersji dokumentu nakładające na Big Tech obowiązek płacenia tantiem za treści wykorzystywane przez producentów AI. Problem dotyczy nie tylko uchwalenia odpowiednich przepisów prawa, ale również ich skutecznego egzekwowania. Ponieważ rząd nie zapewnia w obecnym projekcie arbitrażu pod nadzorem organu państwowego, który pozwalałby twórcom na realne dochodzenie praw do wynagrodzenia w przypadku nieosiągnięcia porozumień z dostawcami aplikacji, giganty technologiczne nie mają interesu w tym, żeby porozumieć się z twórcami. Dlatego też istnienie wiarygodnego organu umocowanego do rozstrzygania ewentualnych sporów wydaje się konieczne dla zapewnienia warunków do negocjowania sprawiedliwego wynagrodzenia. Zamiast tego projekt dopuszcza praktycznie niczym nie ograniczone korzystanie z treści dostępnych w Internecie do trenowania algorytmów AI. Opóźniona o niemal 3 lata implementacja – zamiast wzmacniać prawną pozycję twórców treści – faworyzuje interesy globalnych gigantów technologicznych. W ostatniej wersji projektu ustawy projektodawca wycofał się też z propozycji wyłączenia możliwości trenowania algorytmów GenAI przez podmioty komercyjne. Powinnością prawodawcy jest dbałość o równowagę między postępem technologicznym a ochroną twórców. Aby zapewnić twórcom, dziennikarzom i wydawcom możliwość realnego dochodzenia praw do wynagrodzenia za treści wykorzystywane do szkolenia algorytmów, państwa powinny zobligować dostawców generatywnych aplikacji do uzyskiwania zgód twórców na takie wykorzystanie ich twórczości oraz określić ustawowo terminy na negocjacje warunków finansowych, a w razie nieosiągnięcia porozumienia również terminy interwencji i zakres uprawnień organu arbitrażowego właściwego do rozstrzygania tego typu sporów. Dotychczasowe doświadczenia wielu państw uświadamiają, jak brak takich zapisów nieuchronnie prowadzi w praktyce do sporów sądowych, których wysokie koszty i przewlekłość skutecznie zniechęcą większość poszkodowanych do dochodzenia swoich roszczeń, zwłaszcza w przypadku niewielkiej wartości przedmiotu sporu.

[33] Amerykański Urząd ds. Praw Autorskich orzekł niedawno, że dzieła stworzone przez sztuczną inteligencję bez udziału człowieka nie mogą być objęte prawami autorskimi, ponieważ nie spełniają wymagań określonych w obowiązujących definicjach autorstwa. Zob. B. Brittain, *AI-generated art cannot receive copyrights, US court says*, "Reuters", August 21, 2023; https://www.reuters.com/legal/ai-generated-art-cannot-receive-copyrights-us-court-says-2023-08-21/ [dostęp: 30.12.2023]. Urząd zbiera jednak opinie i rozpoczął konsultacje,

w związku z tym wykorzystywane bezumownie w zasadzie bez żadnych ograniczeń, co w pewnym sensie przywraca balans w działaniach wykorzystujących GenAI. Brak jednoznacznych, wewnętrznie spójnych regulacji w większości krajów skutkuje szybko rosnącą liczbą procesów sądowych związanych z wykorzystaniem materiałów chronionych prawem autorskim w szkoleniach algorytmów[34]. Problem kojarzy się ze znanymi i szeroko dyskutowanymi w nauce i na forach publicznych kontrowersjami wokół granic plagiatu – nieuprawnionych zapożyczeń, w przypadku których mimo istnienia w większości krajów zagęszczonych regulacji wyznaczających granice dopuszczalności podobieństw oraz zaawansowanych narzędzi do wykrywania nadużyć udaje się ograniczyć skalę nadużyć w tekstach akademickich oraz twórczości artystycznej tylko w pewnej mierze, a uznanie kradzieży własności intelektualnej za naruszenie prawa prywatnego i brak organów administracyjnych uprawnionych do rozpatrywania odnośnych skarg z urzędu albo do ścigania sprawców z oskarżenia publicznego sprawia, że poszkodowani muszą dochodzić swoich roszczeń na własną rękę w kosztownych i przewlekłych konfrontacjach sądowych, których rezultat zwykle jest niepewny. Tak nieskuteczne sposoby egzekwowania praw autorskich zniechęcają część poszkodowanych i powodują, że w cywilizowanym świecie większość drobnych deliktów tego typu uchodzi sprawcom płazem.

Wymagania bezpieczeństwa i wyzwania związane z GenAI. Dotychczasowe działania ochronne (regulacje prawne i delegowanie odpowiedzialności) i ich niewystarczalność

Większość analityków bezpieczeństwa specjalizujących się w szeroko rozumianych technologiach cyfrowych uznaje dotychczasowe regulacje odnoszące się

aby stwierdzić, czy odnośne przepisy wymagają rewizji i aktualizacji wobec wymagań i wyzwań związanych z ekspansją GenAI. Zob. E. David, *US Copyright Office wants to hear what people think about AI and copyright*, „The Verge", August 29, 2023: https://www.theverge.com/2023/8/29/23851126/us-copyright-office-ai-public-comments [dostęp: 30.12.2023]. Problem zdolności wytworów GenAI do prawnej ochrony (np. patentowej albo na mocy prawa autorskiego) pod względem strukturalnym sprowadza się do kontrowersji znanych z debat politycznych wokół zdolności patentowej GMO.

[34] Przykładem takich sporów są niedawne pozwy złożone wobec Microsoft i OpenAI przez dziennik „The New York Times" w związku z wykorzystaniem treści będących własnością skarżących do trenowania chatbotów GPT. Zob. G. Barber, *The Generative AI Copyright Fight Is Just Getting Started*, „Wired", December 9, 2023: https://www.wired.com/story/livewired-generative-ai-copyright/ [dostęp: 30.12.2023]; A. Bruell, *New York Times Sues Microsoft and OpenAI, Alleging Copyright Infringement*, „Wall Street Journal", December 27, 2023: https://www.wsj.com/tech/ai/new-york-times-sues-microsoft-and-openai-alleging-copyright-infringement-fd85e1c4 [dostęp: 30.12.2023]; H. Hadero, D. Bauder, *The New York Times sues OpenAI and Microsoft for using its stories to train chatbots*, „Associated Press News", December 27, 2023: https://apnews.com/article/nyt-new-york-times-openai-microsoft-6ea53a8ad3efa06ee4643b697df0ba57 [dostęp: 30.12.2023].

do GenAI za dalece niewystarczające i mało skuteczne w ochronie dotychczasowych zdobyczy cywilizacyjnych ludzkości, takich jak indywidualna podmiotowość, godność i autonomia jednostki, prawa i swobody, własność, prywatność, równość, praworządność i demokracja itp. przed zagrożeniami generowanymi przez GenAI. Można przy tym wyróżnić trzy odmienne koncepcje przeciwdziałania zagrożeniom powodowanym przez GenAI oraz transformację cyfrową w ogólności. Koncepcje te bazują na trzech odmiennych kulturach bezpieczeństwa technologicznego: kulturze tradycyjnej opartej na paradygmacie podwójnego zastosowania (I), kulturze opartej na paradygmacie cyberbezpieczeństwa (II) oraz nowej kulturze propagowanej przez międzynarodowy ruch humanizmu cyfrowego (III)[35].

Kultura bezpieczeństwa oparta na paradygmacie podwójnego zastosowania (I) to najstarsza kultura bezpieczeństwa technologicznego powstała w epoce zimnej wojny. Początkowo skupiała się na międzynarodowym zarządzaniu energią jądrową, bronią chemiczną i biologiczną[36], obejmującym międzynarodowe regulacje (konwencje) oraz specjalne instytucje powołane do nadzorowania ich przestrzegania. Bezpieczeństwo w tym paradygmacie to wyłączna domena instytucji pierwszego sektora. Określenie podwójne zastosowania (*dual use*) był pierwotnie używany w odniesieniu do technologii, które można zastosować zarówno do celów wojskowych, jak i cywilnych, ale z czasem to pierwotne pojęcie przeszło ewolucję i odnosi się dzisiaj bardziej do dobroczynnego i złoczynnego wykorzystania technologii[37]. Przykładami instytucjonalizacji kultury bezpieczeństwa technologicznego opartej na paradygmacie podwójnego zastosowania są *Układ o nierozprzestrzenianiu broni jądrowej* przedłożony do ratyfikacji w 1968 r. oraz *Konwencja o zakazie broni chemicznej* (CWC), przyjęta w 1993 r. i obowiązująca od 1997 r. W przeciwieństwie do *Układu o nierozprzestrzenianiu broni jądrowej*, który nie odnosił się do cywilnych zastosowań technologii jądrowych i w którego realizację zaangażowane były wyłącznie podmioty pierwszego sektora, w realizację CWC włączony został sektor prywatny (przemysł chemiczny). Ponieważ technologie cyfrowe (IT) – podobnie jak chemikalia – są przede wszystkim domeną cywilną, w której operuje prywatny biznes, celem kultury bezpieczeństwa technologicznego opartego na paradygmacie podwójnego zastosowania jest powstrzymanie złoczynnego (złośliwego) wykorzystania zdobyczy techniki przy

[35] Zob. M. Kaldor, *Global Security Cultures*, Cambridge 2018.

[36] J. Molas-Gallart, J. P. Robinson, *Assessment of dual-use technologies in the context of European security and defence*, Report for the Scientific and Technological Options Assessment (STOA), European Parliament, Brussel 1997.

[37] J. Rath, M. Ischi, D. Perkins, *Evolution of Different Dual-use Concepts in International and National Law and Its Implications on Research Ethics and Governance*, „Science and Engineering Ethics" 2014, 20, s. 769-790.

jednoczesnym zachowaniu ich pełnej użyteczności do celów dobroczynnych[38]. Cel ten próbuje się osiągać poprzez współpracę międzynarodową państw i globalnych gigantów IT.

Kultura bezpieczeństwa wywodząca się z tradycji cyberbezpieczeństwa (II) różni się od paradygmatu podwójnego zastosowania przerzuceniem odpowiedzialności za bezpieczeństwo na podmioty drugiego sektora – prywatne firmy technologiczne. Podejście wypracowane w USA w toku współpracy między rządem a firmami z Doliny Krzemowej[39] opiera się na technicznych udoskonaleniach i (w większości kosmetycznych) poprawkach produktów IT wzmacniających ochronę i odporność tych produktów na różnego typu „cyberzagrożenia" rozpoznawane na bieżąco, najczęściej po szkodzie. W ramach tego minimalistycznego, „leseferystycznego" paradygmatu, przerzucającego z instytucji państwa na sektor komercyjny koszty bezpieczeństwa razem z dużym zakresem samowoli i władzy, założono, że wielkie firmy IT skutecznie zapobiegną zagrożeniom dla bezpieczeństwa narodowego w wystarczającym stopniu, jeśli prywatny przemysł specjalizujący się w produktach zapewniających cyberbezpieczeństwo stanie się wysoce rentowny[40]. Elementem tej doktryny jest osobliwa narracja o cyberzagrożeniach, kształtowana przez publiczne informowanie o zagrożeniach dla bezpieczeństwa narodowego ze strony produktów i systemów opartych na technologii cyfrowej. Mimo że nigdzie dotąd zagrożenia cybernetyczne nie okazały się rzeczywistymi zagrożeniami dla bezpieczeństwa narodowego, narracja ta utrzymuje w społeczeństwie odpowiednio wysoki poziomu lęku, który skłania interesariuszy do akceptacji wysokich wydatków na ochronę przed cyberzagrożeniami, a także akceptacji działań inwigilacyjnych, które głęboko ingerują w prywatność obywateli. Głównym wyzwaniem dla ochrony praw jednostki w ramach kultury bezpieczeństwa opartej na paradygmacie cyberbezpieczeństwa jest to, że zarówno infrastruktura, jak i wszystkie rozwiązania zapewniające ochronę przed zagrożeniami, znajdują się w rękach sektora prywatnego[41]. To dominujące obecnie globalne podejście do regulacji innowacji cyfrowych, faworyzujące interesy korporacji BigTech kosztem marginalizacji interesów społeczeństwa obywatelskiego, jest obecnie coraz głośniej krytykowane jako niewystarczające z punktu widzenia ochrony podmiotowości, autonomii i praw jednostki przed zagrożeniami ze strony AI oraz technologicznych oligopoli. Dlatego postuluje się pozytywną zmianę orien-

[38] C. McLeish, P. Nightingale, *Biosecurity, bioterrorism and the governance of science: The increasing convergence of science and security policy*, „Research Policy" 2007, 36, s. 1635-1654.

[39] M. Dunn Cavelty, *Cyber-Security and US Threat Politics: US Efforts to Secure the Information Age*, London 2008.

[40] J. Penney, S. McKune, L. Gill, R. J. Deibert, *Advancing human-rights-by-design in the dual-use technology industry*, „Journal of International Affairs" 2018, 71, 2, s. 103-110.

[41] M. Bay, *What is cybersecurity? In search of an encompassing definition for the post-Snowden era*, „French Journal For Media Research" 2016, 6, s. 1-28.

tacji w kulturze bezpieczeństwa, która wzmocni bezpieczeństwo zwykłych ludzi. Zmiana ta ma się dokonać pod hasłami takimi jak społeczna ocena technologii (TA), etyka sztucznej inteligencji, humanizm cyfrowy czy odpowiedzialność w badaniach i innowacjach (RRI).

Kultura bezpieczeństwa technologicznego oparta na paradygmacie społecznej oceny technologii (TA)[42] bazuje na zaangażowaniu aktywów społecznych – interesariuszy społeczeństwa obywatelskiego (podmioty trzeciego sektora) wspomaganych przez środowiska akademickie. Koncentrując się na celach społecznych, środowiskowych i etycznych, oferuje drogę do podejmowania wyzwań stawianych przez sztuczną inteligencję bez narażania na szwank najważniejszych dotychczasowych zdobyczy cywilizacyjnych – podstawowych dóbr i praw jednostki. Kultura bezpieczeństwa technologicznego trzeciej generacji bazuje na inkluzyjnej, krytycznej analizie zagrożeń dla podstawowych dóbr i uprawnień ludzkiej jednostki, takich jak podmiotowość, autonomia, prywatność, równość, zdrowie czy dobrostan – zagrożeń „wbudowanych" w konkretne produkty, usługi i systemy. Podejście to bazuje na zaangażowaniu sektora akademickiego w kształtowanie społecznie uzgodnionych kierunków, celów i warunków transformacji cyfrowej w ramach czegoś w rodzaju technologicznego oświecenia celem stworzenia przeciwwagi dla potęgi technologicznych oligopoli. Ważnym krokiem w utrwalaniu tej kultury w procesach politycznych winno być włączanie naukowców specjalizujących się w etyce sztucznej inteligencji i ocenie technologii do gremiów doradczych i decyzyjnych. Zaczynem tej nowej kultury bezpieczeństwa technologicznego jest międzynarodowy ruch społeczny organizujący się pod hasłem „humanizmu cyfrowego". Jego najważniejsze założenia zapisano w *Manifeście wiedeńskim w sprawie humanizmu cyfrowego*[43] (2019). Celem ruchu jest prewencyjna ochrona przestrzeni humanitarnej w świecie cyfrowym poprzez jej aktywne, społecznie sprawiedliwe kształtowanie w oparciu o ogólnospołeczną refleksję nad zagrożeniami dla dotychczasowych zdobyczy cywilizacyjnych wynikającymi z wdrażania i stosowania produktów obsługiwanych przez AI.

Próbą zintegrowania elementów tych trzech odmiennych paradygmatów cyberbezpieczeństwa w trosce o społecznie uczciwe i zrównoważone korzystanie ze zdobyczy transformacji cyfrowej jest ustawa o sztucznej inteligencji[44], nad którą od kilku miesięcy pracuje UE. W grudniu 2023 r. Parlament Europejski i Rada UE osiągnęły porozumienie polityczne w sprawie ustawy, której potrzeba była podnoszona, odkąd pojawiły się pierwsze zastosowania GenAI. Projekt ustawy,

[42] Więcej na temat oceny technologii zob. K. Michalski, *Technology Assessment*.

[43] *Vienna Manifesto On Digital Humanism*, https://dighum.ec.tuwien.ac.at/wp-content/uploads/2019/05/manifesto.pdf [dostęp: 19.12.2023].

[44] *Ustawa o sztucznej inteligencji*, https://digital-strategy.ec.europa.eu/pl/policies/regulatory-framework-ai [dostęp: 06.03.2024].

stanowiącej pierwsze w historii kompleksowe ramy prawne dotyczące AI, jest obecnie procedowany przez Parlament UE. Trwają czynności związane z jego tłumaczeniem na języki krajów członkowskich. Ostatnie spekulacje sugerują, że Parlament UE rozważa dodanie GenAI do kategorii systemów „wysokiego ryzyka". Tekst Rady UE precyzuje w art. 4b ust. 1, że niektóre wymagania dotyczące systemów AI wysokiego ryzyka miałyby również zastosowanie do GenAI. Celem nowych przepisów jest wspieranie rozwoju wiarygodnej AI w Europie i poza nią poprzez zapewnienie, aby systemy AI przestrzegały praw podstawowych, bezpieczeństwa i zasad etycznych, a także poprzez przeciwdziałanie zagrożeniom związanym z potężnymi i wpływowymi modelami AI. Jest to zadanie o tyle trudne, że w wielu innych niż AI obszarach nie osiągnięto dotąd globalnego konsensusu co do spójnego systemu prawnych lub etycznych regulacji, dotyczącego relacji między bezpieczeństwem a prawami jednostki takimi jak np. prawo do prywatności[45]. Proponowane przepisy mają na celu przeciwdziałanie ryzyku stwarzanemu przez aplikacje AI, zakazanie praktyk związanych z wykorzystaniem AI, które stwarzają ryzyko uznane za niedopuszczalne, określenie wykazu zastosowań AI wysokiego ryzyka, określenie jasnych wymogów dla systemów AI do zastosowań wysokiego ryzyka, określenie szczegółowych obowiązków podmiotów wdrażających i dostawców AI dla zastosowań wysokiego ryzyka, wprowadzenie wymogu oceny zgodności przed oddaniem danego systemu AI do użytku lub wprowadzeniem do obrotu, wprowadzenie egzekwowalności przepisów po wprowadzeniu danego systemu AI do obrotu oraz ustanowienie struktury zarządzania bezpieczeństwem AI na szczeblu europejskim i krajowym. Europejski ustawodawca przyjął do kategoryzacji zastosowań AI podejście oparte na analizie ryzyka, dla której określono 4 poziomy ryzyka systemów AI: ryzyko nieakceptowalne, wysokie ryzyko, ograniczone ryzyko (systemy AI ze specjalnymi obowiązkami w zakresie transparentności) oraz ryzyko minimalne.

Wszystkie systemy AI uznane za poważne zagrożenie dla bezpieczeństwa, porządku i praw człowieka i zaklasyfikowane do kategorii ryzyka nieakceptowalnego mają zostać całkowicie zakazane. Do systemów AI zaklasyfikowanych do kategorii wysokiego ryzyka zaliczono technologie wykorzystywane w infrastrukturze krytycznej (np. transport) – chodzi o technologie mogące stanowić zagrożenie dla życia i zdrowia obywateli, edukacji lub kwalifikacji zawodowej – chodzi o technologie mogące decydować o dostępie do edukacji, dostępie do zawodu i losach karier zawodowych (np. technologie wykorzystywane do egzaminowania), wykorzystywane do zapewniania bezpieczeństwa procesów i produktów (np. zastosowania AI w chirurgii wspomaganej robotem), w rekrutacji

[45] Zob. K. Michalski, *Konflikty interesów związane z bezpieczeństwem danych osobowych i ochroną prywatności w społeczeństwie cyfrowym*, „Zeszyty Naukowe Wyższej Szkoły Informatyki, Zarządzania i Administracji w Warszawie" 2017, 15, 3/40, s. 55-78.

i zatrudnianiu oraz zarządzaniu zasobami ludzkimi/kadrami (np. oprogramowanie do sortowania/ewaluacji CV i selekcji kandydatów), w usługach prywatnych i publicznych o charakterze podstawowym (np. oprogramowania wspomagające obsługę wniosków kredytowych), w egzekucji prawa – chodzi o technologie, które mogą naruszać prawa podstawowe obywateli (np. ocena wiarygodności świadków lub dowodów wspomagana AI), w kontrolach granicznych, obsłudze migrantów, rozpatrywaniu wniosków wizowych albo azylowych, a także w administrowaniu wymiarem sprawiedliwości i procesami demokratycznymi.

Systemy AI wysokiego ryzyka mają w świetle projektowanej ustawy podlegać restrykcyjnym wymaganiom i obowiązkom, zanim będą mogły zostać wprowadzone do obrotu. Do szczegółowych wymagań, które po wprowadzeniu przepisów w życie zmuszą deweloperów i dostawców AI, których zastosowania zostaną zakwalifikowane do kategorii wysokiego ryzyka, do kosztownego dostosowania procesów inżynieryjnych i procedur produktowych pod kątem zgodności, należą: wdrożenie systemu zarządzania ryzykiem (bazującego na adekwatnej ocenie ryzyka i jego ograniczaniu) (1), dokładność, solidność, wysoki poziom odporności i (cyber-)bezpieczeństwa (2), odpowiednie zarządzanie danymi (wysoka jakość zbiorów danych zasilających system, m.in. w celu zminimalizowania wpływu uprzedzeń i ryzyka bezzasadnej dyskryminacji) (3), nadzór wykonywany przez „czynnik ludzki" (4), polityka pełnej przejrzystości w relacjach z podmiotem wdrażającym lub użytkownikiem, zobowiązująca do rzetelnego informowania (5), bieżące ewidencjonowanie i rejestrowanie działalności w celu zapewnienia identyfikowalności efektów (6), prowadzenie dokumentacji technicznej (obowiązek szczegółowego dokumentowania z uwzględnieniem wszelkich informacji na temat systemu i jego celu pod kątem oceny jego zgodności)[46] (7), a także zgłaszanie zagrożeń, nadużyć i niebezpiecznych incydentów odpowiednim organom oraz współdziałanie w eliminowaniu takich zagrożeń[47]. Deweloperom i dostawcom systemów AI wysokiego ryzyka zalecono stosowanie schematu postępowania w procesie wdrażania nowych produktów, obejmującego następujące kroki: opracowany system musi przejść ocenę zgodności i spełniać wymagania (w przypadku niektórych systemów zaangażowana w ocenę jest również certyfikowana jednostka) (1), następnie system AI podlega rejestracji w bazie danych UE (2), następnie należy wydać deklarację zgodności, a system AI powinien nosić oznakowanie CE (3). Dopiero po spełnieniu tych warunków system może zostać wprowadzony na

[46] Dokumentacja musi zapewniać właściwym organom krajowym i jednostkom notyfikowanym dostęp do wszystkich informacji niezbędnych do oceny zgodności systemu AI z tymi wymogami. Dokumentację tę należy przechowywać do dyspozycji właściwych organów krajowych przez okres 10 lat od chwili wprowadzenia systemu AI ogólnego przeznaczenia na rynek na terytorium UE lub oddaniu go do użytku na terytorium UE.

[47] *The EU Artificial Intelligence Act*, https://www.mhc.ie/hubs/legislation/the-eu-artificial-intelligence-act [dostęp: 29.12.2023].

rynek. Jeżeli w cyklu życia systemu AI zajdą istotne zmiany, należy ponowić całą procedurę począwszy od kroku (1). Przykładowo wszystkie systemy zdalnej identyfikacji biometrycznej uznano za systemy wysokiego ryzyka i objęto restrykcyjnymi wymaganiami. Stosowanie zdalnej identyfikacji biometrycznej w przestrzeniach ogólnodostępnych do celów egzekwowania prawa jest co do zasady zabronione. Od reguły tej są wyjątki, które precyzyjnie uregulowano, np. poszukiwanie zaginionego dziecka, zapobieganie konkretnemu, bezpośredniemu zagrożeniu terrorystycznemu lub wykrywanie, lokalizowanie, identyfikowanie lub ściganie sprawcy poważnego przestępstwa lub podejrzanego o popełnienie takiego przestępstwa. Takie zastosowania wymagają zgody sądu lub innego niezależnego organu oraz odpowiednich ograniczeń czasowych, zasięgu geograficznego i wyszukiwanych baz danych. Ograniczone ryzyko odnosi się natomiast do braku przejrzystości w stosowaniu AI. W związku z tym ustawa wprowadza szczególne obowiązki w zakresie przejrzystości. Wymaga ona m.in., aby użytkownicy podczas korzystania z systemów AI, takich jak np. chatboty, byli informowani, w którym momencie wchodzą w interakcję z maszyną, tak aby mogli podjąć świadomą decyzję o kontynuowaniu albo przerwaniu takiej interakcji. Dostawcy treści będą również musieli zapewnić identyfikowalność treści generowanych przez AI. Ponadto każdy tekst dziennikarski wygenerowany przez AI, opublikowany w celu informowania społeczeństwa o sprawach leżących w interesie publicznym, musi być czytelnie oznakowany jako taki. Dotyczy to również treści audio i wideo stanowiących Deepfake. Ustawa pozwala natomiast na swobodne stosowanie systemów AI o minimalnym ryzyku, do których zaliczono np. gry wideo z obsługą AI lub filtry spamu. Zdecydowana większość systemów AI stosowanych obecnie w UE należy właśnie do tej kategorii ryzyka. Niektóre systemy GenAI mają bardzo uniwersalne zastosowania i można je wykorzystywać do wielu różnych celów w każdorazowo różnych okolicznościach. Znajdują one zastosowania również jako technika bazowa (podstawowa, składowa) zintegrowana z innym systemem, który sam w sobie może stać się źródłem wysokiego ryzyka.

Po wprowadzeniu na rynek systemu AI organy są odpowiedzialne za nadzór rynku, podmioty wdrażające zapewniają nadzór i monitorowanie ze strony człowieka, a dostawcy mają obowiązek monitorowania ewentualnych zagrożeń oraz nadużyć ujawnionych już po wprowadzeniu do obrotu. Zarówno dostawcy systemów AI, jak i podmioty wdrażające, mają również obowiązek zgłaszania poważnych incydentów i nieprawidłowego działania systemów odpowiednim organom. Do nadzorowania bezpieczeństwa w obszarze zastosowań AI oraz egzekwowania zapisów ustawy oraz jej wdrażania we współpracy z państwami członkowskimi powołano w lutym 2024 r. Europejski Urząd ds. AI działający w ramach Komisji. Jego celem jest stworzenie środowiska, w którym technologie AI mają zaimplementowane zasady dotyczące poszanowania ludzkiej godności,

prawa i cieszą się zaufaniem. Urząd ma wspierać również badania, innowacje oraz współpracę w dziedzinie AI między różnymi interesariuszami, a także angażować się w międzynarodowy dialog i współpracę, uznając potrzebę globalnego ujednolicenia standardów zarządzania AI. Dzięki tym wysiłkom Europejski Urząd ds. AI ma dążyć do pozycjonowania Europy jako światowego lidera w zakresie etycznego i zrównoważonego rozwoju technologii AI[48].

Unijna ustawa o AI ma zacząć obowiązywać w każdym z 27 państw członkowskich bez konieczności jej lokalnej transpozycji do prawa krajowego przez każde z nich. Wejdzie w życie 20 dni po jej opublikowaniu w Dzienniku Urzędowym i będzie w pełni stosowana 2 lata później, z pewnymi wyjątkami: zakazy wejdą w życie po sześciu miesiącach, zasady zarządzania i obowiązki dotyczące modeli sztucznej inteligencji ogólnego przeznaczenia zaczną obowiązywać po upływie 12 miesięcy, a przepisy dotyczące systemów sztucznej inteligencji – wbudowane w produkty regulowane – będą obowiązywać po 36 miesiącach. Aby ułatwić przejście na nowe ramy regulacyjne, Komisja uruchomiła pakt na rzecz sztucznej inteligencji – dobrowolną inicjatywę mającą na celu wsparcie przyszłego wdrażania – i zachęca twórców AI z Europy i spoza niej do respektowania kluczowych wymagań i obowiązków określonych w ustawie o AI jeszcze przed jej formalnym wejściem w życie.

Dotychczasowe doświadczenia z wprowadzaniem w UE w interesie społecznym regulacji odnoszących się do prężnego rynku produktów i usług IT nie napawają jednak optymizmem. Ponieważ między poszczególnymi krajami członkowskimi, w których rządzą odmienne orientacje światopoglądowe, istnieją silne rozbieżności co do wizji interwencjonizmu państwa, silniejszy wpływ na procesy legislacyjne wywiera w praktyce agresywny lobbing wielkich globalnych korporacji niż naciski elit politycznych. Pouczającym przykładem są prace nad unijnym pakietem reform sprzed 10 lat, mających zapewnić skuteczną ochronę danych i prywatności obywateli UE przed nadużyciami operatorów platform internetowych i dostawców aplikacji smartfonowych – nadużyciami związanymi z wymuszaniem zgód na bezzasadne korzystanie z danych pod groźbą odmowy dostępu do usługi, nieprzejrzyste praktyki przetwarzania takich danych, przechowywania oraz udostępniania ich osobom trzecim, w tym również transferowania na terytoria państw, w których nie obowiązują restrykcyjne przepisy dotyczące ochrony spersonalizowanych informacji (m.in. USA)[49].

[48] B. McElligott.

[49] W reakcji na skandal inwigilacyjny związany z tajnym programem PRISM, zdemaskowanym w czerwcu 2013 r. przez Edwarda Snowdena, a także wobec późniejszego orzeczenia TSUE z 6.10.2015 r. w sprawie skargi austriackiego internauty i użytkownika serwisu Facebook Maxa Schremsa (sygnatura sprawy C-362/14) – orzeczenia, w którym Trybunał stwierdził nieważność umowy Safe Harbor, zawartej w 2000 r. między Komisją Europejską a Departamentem Handlu USA, oraz zobowiązał Komisję Europejską do wprowadzenia

Podsumowanie i konkluzje

Problematyka zagrożeń wynikających z żywiołowego rozwoju i ekspansji GenAI oraz wymagań i wyzwań związanych z bezpiecznym korzystaniem ze zdobyczy tego rozwoju jest o wiele bardziej złożona, niż wskazują zagadnienia omówione w niniejszym artykule. Rzeczywiste i potencjalne zagrożenia dla bezpieczeństwa ludzi ze strony generatywnej sztucznej inteligencji obejmują również przede wszystkim ryzyka wynikające z używania ChatGPT i innych dużych modeli językowych (LLM) bez odpowiedniego zrozumienia sposobu ich działania. Ponieważ niewielu użytkowników jest świadomych rzeczywistej wartości generowanych wyników, istnieje ryzyko fałszywych oczekiwań, które mogą mieć szkodliwe konsekwencje (np. fałszywe zalecenia dla zdrowia, nieuzasadnione uprzedzenia, wyolbrzymianie albo bagatelizowanie określonych zagrożeń i podejmowanie działań nieadekwatnych do sytuacji itp.). Ochrona przed tymi i wieloma innymi zagrożeniami, związanymi czy to z naruszeniami prywatności, wizerunku i praw własności intelektualnej, czy też z dezinformacją i manipulowaniem procesami politycznymi, stanowi poważne wyzwanie dla współczesności, któremu można skutecznie stawić czoła tylko pod warunkiem zmiany dotychczasowych paradygmatów kultury bezpieczeństwa technologicznego. Dotychczasowe próby uregulowań prawnych bazujące na paradygmacie podwójnego zastosowania albo paradygmacie cyberbezpieczeństwa okazują się niewystarczające wobec żywiołowej ekspansji i rosnącej potęgi GenAI. Komisja Europejska w projekcie Ustawy o sztucznej inteligencji jako pierwsza poczyniła kroki w kierunku instytucjonalizacji elementów nowej kultury i można mieć nadzieję, że rządy i parlamenty krajów członkowskich – a za ich przykładem również decydenci w innych krajach – również pójdą w tym kierunku. Wkrótce okaże się, jaki ostateczny kurs regulacyjny obierze europejski prawodawca wobec GenAI i czy zostanie ona ryczałtem zaliczona do systemów AI „wysokiego ryzy-

przepisów skuteczniej chroniących prawa obywateli UE przed nadużyciami korporacji BigTech – Komisja skierowała do Parlamentu Europejskiego pakiet projektów dyrektyw, m.in. projekt ogólnego rozporządzenia o ochronie danych oraz tzw. nowej dyrektywy ciasteczkowej (cookies). Wiosną 2015 roku europejska platforma demaskatorska lobbyplag.eu upubliczniła przecieki z posiedzeń Komisji Parlamentu Europejskiego ds. Wolności Obywatelskich, Sprawiedliwości i Spraw Wewnętrznych (LIBE) oraz ponad 10 tys. wewnętrznych dokumentów unijnej rady ministrów, demaskując przebieg prac nad tym pakietem regulacji i aktywność czołowych amerykańskich gigantów z branży IT we wpływaniu na ostateczny kształt zapisów. Obstrukcyjne działania kilku eurodeputowanych z tej komisji, którzy poddając pod głosowanie ponad 4000 nieistotnych poprawek skutecznie opóźniali wprowadzenie w życie nowych restrykcji z korzyścią dla interesów wielkich korporacji, podają w wątpliwość szczerość politycznych deklaracji manifestujących troskę o bezpieczeństwo obywateli i poszanowanie ich praw. Zob. I. Oleksiewicz, K. Michalski, E. Sienkiewicz, *Bezpieczeństwo w społeczeństwie informacyjnym. Zagadnienia w wymiarze online i offline*, Warszawa 2017, s. 32-34.

ka". Ostatnie raporty sugerują, że unijni prawodawcy mogą domyślnie obarczyć GenAI jedynie wymaganiami przejrzystości. Oczekuje się, że w nadchodzących tygodniach Parlament UE sfinalizuje swoje stanowisko w sprawie ustawy o AI. Ten rok w każdym razie przejdzie do historii jako znaczący dla dalszego rozwoju AI i odnośnych regulacji, natomiast dopiero kolejne lata zweryfikują skuteczność wprowadzonych niebawem środków bezpieczeństwa. Całkowite ograniczenie rozwoju GenAI w interesie bezpieczeństwa jednostki spowodowałoby rezygnację z niekwestionowanych dobrodziejstw AI, straty ekonomiczne i utratę znaczenia Europy w globalnej konkurencji, a tym samym skutki odwrotne w stosunku do zamierzonych. Tak czy owak należy się jednak liczyć z niechęcią ze strony oligopoli dużych firm, choć zachowania pracowników niektórych gigantów z branży BigTech zdradzają pewne sympatie do tej nowej kultury humanizmu cyfrowego.

Regulacje prawne godzące wymagania bezpieczeństwa z wymaganiami konkurencyjności w globalnym wyścigu technologicznym mają fundamentalne znaczenie, ale same nie rozwiążą wszelkich problemów wynikających ze stosowania GenAI. To nie rządy i organizacje pierwszego sektora, ani wielkie korporacje IT, lecz społeczeństwo obywatelskie winno odgrywać kluczową rolę w kształtowaniu ram i warunków rozwoju technologii, dlatego należy wzmocnić inicjatywy wzmacniające głosy tej grupy interesariuszy (*empowering*). Należy mobilizować wszystkie dostępne aktywa społeczne do włączania się w rozwój AI, tak aby ten rozwój uwzględniał różne punkty widzenia, w szczególności perspektywy grup dotychczas marginalizowanych. Zrozumienie zagadnienia wymaga z jednej strony oddzielenia zagrożeń dla jednostek (skala mikro), społeczności/grup (skala mezo-) i całej ludzkości (skala makro), z drugiej strony zagrożeń spontanicznych, wynikających z działania systemów technicznych i zagrożeń powodowane przez czynnik ludzki (podwójne zastosowanie). Zaangażowanie społecznych interesariuszy w kształtowanie warunków dalszego rozwoju GenAI i określanie związanych z tym wymagań bezpieczeństwa wymaga świadomości społecznych korzyści związanych z zastosowaniami AI oraz społecznych kosztów i niepożądanych skutków ubocznych. Platformą takiego technologicznego oświecenia, umożliwiającą balansowanie przeciwstawnych interesów i konfrontowanie odmiennych perspektyw poznawczych, powinna być społeczna ocena technologii.

References

Bibliografia

Opracowania

Bay M., *What is cybersecurity? In search of an encompassing definition for the post-Snowden era*, „French Journal for Media Research" 2016, 6, s. 1-28.

Bergen N., Huang A., *A Brief History of Generative AI, Dichotomies, Generative AI, Navigating Towards a Better Future*, Deloitte, 2023, Issue 2.

Brynjolfsson E., Li D., Raymond L.R., *Generative AI at Work* (Working Paper 31161), Working Paper Series, National Bureau of Economic Research, Cambridge MA, April 2023.

Cao Y., Li S., Liu Y., Yan Z., Dai Y., Yu P.S., Sun L., *A Comprehensive Survey of AI-Generated Content (AIGC): A History of Generative AI from GAN to ChatGPT*, arXiv:2303.04226 [cs.AI], March 7, 2023.

Collingridge D., *The Social Control of Technology*, London 1980.

Dunn Cavelty M., *Cyber-Security and US Threat Politics: US Efforts to Secure the Information Age*, London 2008.

Kaldor M., *Global Security Cultures*, Cambridge 2018.

McLeish C., Nightingale P., *Biosecurity, bioterrorism and the governance of science: The increasing convergence of science and security policy*, „Research Policy" 2007, 36, s. 1635-1654.

Metz C., *OpenAI Plans to Up the Ante in Tech's A.I. Race*, „The New York Times" March 14, 2023.

Michalski K., *Autonomizacja techniki i niepożądane skutki eliminowania człowieka jako źródła błędów*, „Filo-Sofija. Z Problemów Współczesnej Filozofii" 2017, 39/I (2017/4/I), t.6: Człowiek i maszyna, s. 79-95.

Michalski K., *Konflikty interesów związane z bezpieczeństwem danych osobowych i ochroną prywatności w społeczeństwie cyfrowym*, „Zeszyty Naukowe Wyższej Szkoły Informatyki, Zarządzania i Administracji w Warszawie" 2017, 15, 3/40, s. 55-78

Michalski K., *Technology Assessment. Ocena technologii – nowe wyzwania dla filozofii nauki i ogólnej metodologii nauk*, Rzeszów 2019

Michalski K., Jurgilewicz M., *Konflikty technologiczne. Nowa architektura zagrożeń w epoce wielkich wyzwań*, Warszawa 2021, s. 98-102.

Molas-Gallart J., Robinson J. P., *Assessment of dual-use technologies in the context of European security and defence*, Report for the Scientific and Technological Options Assessment (STOA), European Parliament, Brussel 1997.

Moravec H., *Mind Children*, Boston 1988.

Oleksiewicz I., Michalski K., Sienkiewicz E., *Bezpieczeństwo w społeczeństwie informacyjnym. Zagadnienia w wymiarze online i offline*, Warszawa 2017, s. 32-34.

Pasick A., *Artificial Intelligence Glossary: Neural Networks and Other Terms Explained*, „The New York Times" March 27, 2023.

Penney J., McKune S., Gill L., Deibert R. J., *Advancing human-rights-by-design in the dual-use technology industry*, „Journal of International Affairs" 2018, 71, 2, s. 103-110.

Rath J., Ischi M., Perkins D., *Evolution of Different Dual-use Concepts in International and National Law and Its Implications on Research Ethics and Governance*, „Science and Engineering Ethics" 2014, 20, s. 769-790.

Roose K., *A Coming-Out Party for Generative A.I., Silicon Valley's New Craze*, „The New York Times" October 21, 2022.

Rifkin J., *Koniec pracy. Schyłek siły roboczej na świecie i początek ery postrynkowej*, Wrocław 2001

Netografia

Barber G., *The Generative AI Copyright Fight Is Just Getting Started*, „Wired" December 9, 2023, https://www.wired.com/story/livewired-generative-ai-copyright/ [dostęp: 30.12.2023].

Brittain B., *AI-generated art cannot receive copyrights, US court says*, Reuters, August 21, 2023; https://www.reuters.com/legal/ai-generated-art-cannot-receive-copyrights-us-court-says-2023-08-21/ [dostęp: 30.12.2023].

Bruell A., *New York Times Sues Microsoft and OpenAI, Alleging Copyright Infringement*, „Wall Street Journal" December 27, 2023, https://www.wsj.com/tech/ai/new-york-times-sues-microsoft-and-openai-alleging-copyright-infringement-fd85e1c4 [dostęp: 30.12.2023]

Carter J., *China's game art industry reportedly decimated by growing AI use*, „Game Developer" April 11, 2023; https://www.gamedeveloper.com/art/china-s-game-art-industry-reportedly-decimated-ai-art-use [dostęp: 22.11.2023].

Coyle J., *In Hollywood writers' battle against AI, humans win (for now)*, „Associated Press News" September 27, 2023; https://apnews.com/article/hollywood-ai-strike-wga-artificial-intelligence-39ab72582c3a15f77510c9c30a45ffc8 [dostęp: 28.11.2023].

Don't fear an AI-induced jobs apocalypse just yet, „The Economist" March 6, 2023; https://www.economist.com/business/2023/03/06/dont-fear-an-ai-induced-jobs-apocalypse-just-yet [dostęp: 22.11.2023]

Eapen T.T., Finkenstadt D.J., Folk J., Venkataswamy L., *How Generative AI Can Augment Human Creativity*, „Harvard Business Review" June 16, 2023; https://hbr.org/2023/07/how-generative-ai-can-augment-human-creativity [dostęp: 28.11.2023]

Gupta S., *Underrepresented groups in countries around the world are worried about AI being a threat to jobs*, Fast Company, October 31, 2023; https://www.fastcompany.com/90975180/ai-threat-jobs-fears-survey-seven-countries-job-seekers-hr [dostęp: 28.11.2023].

Guterres A., *UNO-Secretary-General's remarks to the Security Council on Artificial Intelligence*, https://www.un.org/sg/en/content/sg/speeches/2023-07-18/secretary-generals-remarks-the-security-council-artificial-intelligence, [dostęp: 28.11.2023].

Harreis H., Koullias T., Roberts R., *Generative AI: Unlocking the future of fashion*, McKinsey&Company, March 8, 2023 https://www.mckinsey.com/industries/retail/our-insights/generative-ai-unlocking-the-future-of-fashion [dostęp: 28.11.2023].

Heaven W.D., *AI is dreaming up drugs that no one has ever seen. Now we've got to see if they work*, „MIT Technology Review" February 15, 2023; https://www.technologyreview.com/2023/02/15/1067904/ai-automation-drug-development/ [dostęp: 28.11.2023]

Hadero H., Bauder D., The New York Times sues OpenAI and Microsoft for using its stories to train chatbots, „Associated Press News" December 27, 2023: https://apnews.com/article/nyt-new-york-times-openai-microsoft-6ea53a8ad3efa06ee4643b697df0ba57 [dostęp: 30.12.2023].

Islam A., *A History of Generative AI: From GAN to GPT-4*, „Marktechpost" March 21, 2023;

https://www.marktechpost.com/2023/03/21/a-history-of-generative-ai-from-gan-to-gpt-4/ [dostęp 18.11.2023].

Karpathy A., Abbeel P., Brockman G., Chen P., Cheung V., Duan Y., Goodfellow I., Kingma D., Ho J., Houthooft R., Salimans T., Schulman J., Sutskever I., Zaremba W., *Generative models,* "OpenAI" June 16, 2016, https://openai.com/research/generative-models [dostęp: 28.11.2023].

Lamar A., *Jay-Z's Delaware producer sparks debate over AI rights*, https://www.delawareonline.com/story/entertainment/2023/04/21/jay-zs-delaware-producer-young-guru-artificial-voice-generators-rights/70107389007/ [dostęp: 28.11.2023].

McElligott, *ChatGPT and the EU AI Act*, https://www.mhc.ie/latest/insights/chatgpt-and-the-eu-ai-act [dostęp: 29.10.2023].

OpenAI, *Reducing bias and improving safety in DALL·E 2*, July 18, 2022, https://openai.com/blog/reducing-bias-and-improving-safety-in-dall-e-2 [dostęp: 22.11.2023]

Owen S., *Synthetic Data for Better Machine Learning*, https://www.databricks.com/blog/2023/04/12/synthetic-data-better-machine-learning.html, [dostęp: 14.12.2023

Paz I., *16 najlepszych generatorów kodów AI*, https://www.morningdough.com/pl/ai-tools/best-ai--code-generators/, [dostęp: 28.11.2023].

Sharma H., *Synthetic Data Platforms: Unlocking the Power of Generative AI for Structured Data*, https://www.kdnuggets.com/2023/07/synthetic-data-platforms-unlocking-power-generative-ai-structured-data.html, [dostęp: 14.12.2023].

The EU Artificial Intelligence Act https://www.mhc.ie/hubs/legislation/the-eu-artificial-intelligence-act, [dostęp: 29.12.2023].

Traylor J., *No quick fix: How OpenAI's DALL·E 2 illustrated the challenges of bias in AI*, „NBC News" July 27, 2022, https://www.nbcnews.com/tech/tech-news/no-quick-fix-openais-dalle-2-illustrated-challenges-bias-ai-rcna39918, [dostęp: 22.11.2023].

Traisman R., *A magazine touted Michael Schumacher's first interview in years. It was actually AI*, https://www.npr.org/2023/04/28/1172473999/michael-schumacher-ai-interview-german-magazine [dostęp: 28.11.2023].

Ustawa o sztucznej inteligencji https://digital-strategy.ec.europa.eu/pl/policies/regulatory-framework-ai [dostęp:06.03.2024].

Vienna Manifesto On Digital Humanism, https://dighum.ec.tuwien.ac.at/wp-content/uploads/2019/05/manifesto.pdf [dostęp: 19.12.2023].

Viner K., Bateson A., *The Guardian's approach to generative AI*, „The Guardian" June 16, 2023, https://web.archive.org/web/20240103230448/https:/www.theguardian.com/help/insideguardian/2023/jun/16/the-guardians-approach-to-generative-ai [dostęp: 28.11.2023]

Zhou V., AI is already taking video game illustrators' jobs in China, „Rest of World" April 11, 2023; https://restofworld.org/2023/ai-china-video-game-layoffs-illustrators/ [dostęp: 28.11.2023].

Zirpoli Ch.T., *Generative Artificial Intelligence and Copyright Law*, Congressional Research Service. LSB10922, September 29, 2023: https://crsreports.congress.gov/product/pdf/LSB/LSB10922 [dostęp: 07.12.2023].