Journal of KONBiN 2024
Volume 54, Issue 2
DOI 10.5604/01.3001.0054.6490

Janusz SZPYTKO ORCID 0000-0001-7064-0183, szpytko@agh.edu.pl – corresponding author
AGH University of Krakow (Akademia Górniczo-Hutnicza), Poland

BEZPIECZEŃSTWO TECHNIKI Z MASZYNOWĄ INTELIGENCJĄ

Safety of technology with machine intelligence

Streszczenie: *Skutkiem cyfryzacji przemysłu i środowisk aktywności człowieka, a ponadto rozwoju maszynowej inteligencji, zasadne jest podjęcie debaty w zakresie bezpieczeństwa techniki z maszynową inteligencją. Maszynowa inteligencja jest swojego rodzaju przedłużeniem i wzmocnieniem człowieka w zakresie budowy określonych zasobów danych i wiedzy z ukierunkowaniem na ich celowe i bezpieczne wykorzystanie. Przedmiotem wypowiedzi jest analiza wybranego piśmiennictwa w zakresie bezpieczeństwa techniki z maszynową inteligencją.*
Słowa kluczowe: *bezpieczeństwo techniki, maszynowa inteligencja, krytyczne zasoby*

Abstract*: As a result of the digitalization of industry and human activity environments, and also the development of machine intelligence, it is reasonable to start a debate on the safety of technology with machine intelligence. Machine intelligence is a kind of extension and reinforcement of humans in the area of building specific data and knowledge resources with a focus on their purposeful and safe use for specific needs. The subject of the statement is the analysis of selected literature on the safety of technology with machine intelligence.*
Keywords: safety of technology, machine intelligence, critical resources

Received: June 8, 2024/ Revised: June 24, 2024/ Accepted: June 25, 2024/ Published: June 28, 2024

1. Wstęp

Bezpieczeństwo techniki jest właściwością obiektu (lub systemu) charakteryzującą się jego odpornością na powstawanie sytuacji niebezpiecznych [14]. Rozróżnia się stany obiektu, w których występują:

1. zawodność bezpieczeństwa (prawdopodobieństwo warunkowego wystąpienia uszkodzenia obiektu lub błędu jego działania skutkującego zagrożeniem jego bezpieczeństwa oraz obiektów z nim współpracujących, środowiska i życia ludzkiego),
2. zawodność funkcjonowania (prawdopodobieństwo warunkowego wystąpienia uszkodzenia obiektu lub błędu jego działania skutkującego jedynie przerwą w jego funkcjonowaniu lub niepełne funkcjonowanie),
3. niezawodność bezpieczeństwa (prawdopodobieństwo warunkowego niewystąpienia uszkodzenia obiektu lub błędu jego działania skutkującego zagrożeniem jego bezpieczeństwa oraz obiektów z nim współpracujących, środowiska i życia ludzkiego).

Bezpieczeństwo techniki z maszynową inteligencją jest połączeniem interdyscyplinarnej wiedzy i praktyki w zakresie niezawodnej eksploatacji obiektów (systemów) w sposób pierwotnie przewidziany (bez niezamierzonych zdarzeń).

Aktualne zastosowania maszynowej inteligencji w procesach eksploatacji obiektów (systemów) technicznych są ukierunkowane na monitorowanie i identyfikowanie zagrożeń bezpieczeństwa w miejscu użytkowania, a w szczególności:

1. monitorowanie zagrożeń i kontrolę obiektów,
2. identyfikację i minimalizację niekorzystnych zdarzeń w czasie rzeczywistym,
3. identyfikację i minimalizację błędów ludzkich w czasie rzeczywistym,
4. realizację ćwiczeń praktycznych z użyciem rzeczywistości wirtualnej z możliwymi scenariuszami zagrożeń bezpieczeństwa,
5. analizę i syntezę możliwych zagrożeń bezpieczeństwa,
6. monitorowanie stanu psycho-fizycznego operatorów w określonych warunkach na potrzeby minimalizacji zagrożeń bezpieczeństwa.

Przedmiotem wypowiedzi jest analiza wybranego piśmiennictwa w zakresie bezpieczeństwa techniki z maszynową inteligencją, zwłaszcza dla zasobów krytycznych.

2. Stan wiedzy i techniki

Skutkiem cyfryzacji przemysłu i środowisk aktywności człowieka, a ponadto rozwoju maszynowej inteligencji, zasadne jest podjęcie debaty w zakresie bezpieczeństwa techniki z maszynową inteligencją. Maszynowa inteligencja wykorzystuje specjalistyczne algorytmy umożliwiające realizację złożonych operacji matematycznych (klasyfikacji, regresji) na dużych zbiorach danych i wiedzy umożliwiających budowanie modeli predykcyjnych na potrzeby decyzyjne.

Rozwiązywanie problemów związanych z budową bezpiecznej i niezawodnej inteligencji maszynowej jest przedmiotem pracy [15]. W pracy [3] autor stwierdza, że maszynowa inteligencja ma potencjał minimalizacji ryzyka w przypadku wystąpienia krytycznych dla bezpieczeństwa zdarzeń w rezultacie wbudowanych mechanizmów uczenia maszynowego (Machine Learning, ML). Niedogodnością uczenia maszynowego jest brak mechanizmu rozumowania kontekstowego w sytuacjach z dużym zmiennym w czasie poziomem niepewności i ograniczonego dostępu do informacji oraz wymaganiem adaptacji do nieznanego otoczenia. Konsekwencją mogą być zdarzenia krytyczne dla bezpieczeństwa, gdyż obiekty techniczne z włączoną maszynową inteligencją mogą działać w trybie wcześniej zdeterminowanym przez projektanta. Dlatego też niezbędne jest, aby warunki eksploatacji obiektów technicznych z maszynową inteligencją były każdorazowo szczegółowo opisane przez producenta. Większość modeli sztucznej inteligencji, w szczególności uczenia maszynowego, to modele statystyczne [1]. W praktyce mamy do czynienia ze zdarzeniami dynamicznymi, które wymagają wypracowania innych modeli.

Autorzy pracy [11] stwierdzają, że sztuczna inteligencja (AI) umożliwi rozwój autonomicznych systemów bezpieczeństwa, w których algorytmy uczenia maszynowego (ML) uczą się zoptymalizowanych i bezpiecznych rozwiązań. Niezbędne jest jednakże poszukiwanie rozwiązań dla systemów bezpieczeństwa krytycznego, które powinny obejmować: multidyscyplinarne podejście, pełny cykl życia produktu z uwzględnieniem zmian w jego fazach życia rejestrowanych (dynamiczne zbieranie danych eksploatacyjnych w celu aktualizacji modelu uczenia maszynowego) w długim okresie eksploatacji.

Stosowane modele maszynowej inteligencji jako wsparcie procesów decyzyjnych w systemach krytycznych wykorzystują różne modele. W pracy [6] dokonano analizy skuteczności stosowania modelu rozmyto-eksperckiego (Fuzzy Adaptive Networks, FAN), sieci neuronowych (Artificial Neural Network, ANN) i adaptacyjnego systemu wnioskowania neuro-rozmytego (Adaptive Neuro-Fuzzy Inference Systems, ANFIS). Autorzy stwierdzili, że technika hybrydowa (synergistyczna kombinacja technik maszynowej inteligencji) wykazuje mniejsze błędy przy mniejszej liczbie iteracji i może być też stosowana do szacowania ryzyk dla krytycznej infrastruktury.

Obiekty (i systemy) techniczne są coraz bardziej złożone, dlatego też ich analizowanie z uwagi na możliwe zagrożenia bezpieczeństwa funkcjonalnego jest zagadnieniem trudnym i błędne decyzje są możliwe. Autorzy w pracy [5] zaproponowali połączenie metody analizy drzewa błędów (Fault Tree Analysis, FTA) z uczeniem maszynowym (Machine Learning, ML) w celu identyfikacji nieprawidłowości w procesie eksploatacji obiektu technicznego w sytuacjach zagrożenia bezpieczeństwa. W przypadku zdarzeń nienormalnych w odniesieniu do modelu normalnego zachowania obiektu, decyzję podejmuje operator z wykorzystaniem własnych doświadczeń i podejmuje próbę korekty modelu drzewa błędów procesu eksploatacji obiektu. Skuteczność proponowanego podejścia pokazano na hipotetycznym przykładzie systemu dystrybucji paliwa lotniczego (Aircraft Fuel Distribution System, AFDS).

W pracy [12] autor zaproponował koncepcję (model) sztucznej inteligencji zorientowanej na człowieka HCAI (Human-Centered Artificial Intelligence), który z założenia powinien wspomagać proces projektowania charakteryzujący się niezawodnością, bezpieczeństwem i zaufaniem w zakresie eksploatacji (Reliable, Safe and Trustworthy, RST). W pracy [8] autorzy proponują wprowadzenie do praktyki eksploatacji obiektu technicznego limitowanego czasu życia określonej wersji maszynowej inteligencji i następnie zastąpienie jej nową generacją ze skuteczniejszym rozwiązaniem w zakresie zapewnienia niezawodności bezpieczeństwa w procesie użytkowania.

W pracy [13] autorzy dokonali przeglądu literatury w zakresie zastosowania sztucznej inteligencji w systemach o znaczeniu krytycznym dla bezpieczeństwa. W pracy [10] zidentyfikowano szkody wyrządzane przez systemy sztucznej inteligencji i przedstawiono wizję zorientowanego na człowieka i wrażliwego na kontekst rozwiązania. Autorzy pracy [16] dokonali analizy zarejestrowanych wybranych awarii systemów z zastosowaniem AI. W konkluzji stwierdzili, że niezbędne jest wypracowanie mechanizmów bezpieczeństwa i szukanie narzędzi zapewnienia bezpieczeństwa w obiektach samodoskonalących się.

W pracy [7] autor podjął debatę w zakresie bezpieczeństwa w przyszłej wersji sztucznej inteligencji i przeprowadził dyskusję na temat etyki maszyn i bezpieczeństwa maszynowej inteligencji i możliwych zagrożeń bezpieczeństwa.

Nowe technologie mają bezpośrednie przełożenie na ewolucję otoczenia i miejsca pracy operatorów obiektów technicznych i uczestników celowych procesów technologicznych i wymagają bezpiecznego współdziałania z ludźmi w środowisku pracy, przykładowo z robotami współpracującymi typu cobot (collaborative robot). Ewolucji ulegają struktura siły roboczej, rynek pracy, formy zatrudnienia i organizacji pracy, czego skutkiem są nowe kategorie zagrożeń i wyzwań w zakresie bezpieczeństwa, w tym zdrowia pracowników (operatorów).

Przykładowo maszynowa inteligencja umożliwia skuteczniejsze kompleksowe projektowane opieki okołooperacyjnej, interwencji zdrowotnych lub zabiegu chirurgicznego dla pacjentów korzystających ze świadczeń zdrowotnych [9]. Sztuczna inteligencja umożliwia identyfikację zagrożeń dla zdrowia pacjentów i w tym zakresie poprawia profil bezpieczeństwa w odniesieniu do opieki zdrowotnej [2].

Maszynowa inteligencja umożliwia zrozumienie nowych zagrożeń w miejscu pracy i w rezultacie może wspomagać i regulować (sterować) procesy zapewnienia bezpieczeństwa i higieny w środowiskach aktywności ludzi, minimalizować możliwe zagrożenia bezpieczeństwa oraz zapewniać dostęp do pracy w nowych środowiskach osobom ze specyficznymi potrzebami, jak również umożliwiać dostęp do nauki osobom ze szczególnymi potrzebami edukacyjnymi, co może przyczynić się do wyrównywania szans w życiu społecznym i zawodowym.

Projektowanie bezpieczeństwa z użyciem maszynowej inteligencji i nowych technik komunikacji zorientowanych na człowieka w miejscu pracy ma istotne znaczenie dla zapewnienia dobrostanu pracowników przy jednoczesnym maksymalnym wykorzystaniu transformacji cyfrowej i budowy obszaru dla innowacyjności [4].

3. Uwagi końcowe

Zainteresowanie biznesu narzędziami maszynowej inteligencji, w szczególności w technice, istotnie przyspiesza. Użycie zaawansowanych algorytmów umożliwia uzyskanie określonej przewagi w prowadzonym biznesie, ale równocześnie stanowi czynnik przyspieszający rozwój istniejących procesów z ukierunkowaniem na innowacyjność. Maszynowa inteligencja jest równocześnie źródłem obaw użytkowania w zakresie zapewnienia określonego poziomu bezpieczeństwa z uwagi na wykorzystanie danych (w tym wrażliwych) i odpowiedzialność za generowane treści i procesy decyzyjne (w tym sterowania).

Zapewnienie bezpieczeństwa eksploatacyjnego systemów technicznych używających maszynowej inteligencji jest zagadnieniem niezmiernie trudnym z uwagi na złożoność i interdyscyplinarność oraz dynamikę i trudne do przewidzenia kierunki rozwoju.

Maszynowa inteligencja jest swoistego rodzaju przedłużeniem i wzmocnieniem człowieka z jego umysłem (wyrażanym systemem cyberfizycznym) w zakresie budowy określonych zasobów danych i wiedzy z ukierunkowaniem na ich celowe i skuteczne wykorzystanie (eksplorację) dla określonych potrzeb z wykorzystaniem symbolicznej formy zapisu i wyrażania. Budowa i stosowanie w praktyce mechanizmów zapewniających bezpieczeństwo budowy i eksploracji określonych zasobów danych i wiedzy w systemach cyberfizycznych jest oczekiwane. Powinny one uwzględniać

prywatność i minimalizację zagrożeń dla użytkowników i środowiska oraz precyzyjne regulacje prawne.

Każdą użyteczną technikę, która jest przedmiotem ewolucji, można rozpatrywać w sensie pozytywnym i negatywnym. Istnieje potrzeba wbudowywania w nowe techniki na poziomie projektowania gwarancji bezpieczeństwa funkcjonalnego, w szczególności w zakresie odporności na ataki zewnętrzne, nadużycia, skłonności do wypadków i działań o charakterze destrukcyjnym. Rozwiązania w zakresie bezpieczeństwa techniki z maszynową inteligencją również są przedmiotem ewolucji i wymagają określonych działań wyprzedzających powiązanych z aktualną wiedzą i praktyką.

4. References

1. J. Braband, H. Schäbe, “On safety assessment of artificial intelligence”, arXiv preprint arXiv:2003.00260, 2020.
2. A. Choudhury, O. Asan, “Role of artificial intelligence in patient safety outcomes: systematic literature review”, JMIR medical informatics, 8(7), e18599, 2020.
3. M. Cummings, “Rethinking the maturity of artificial intelligence in safety-critical settings”, AI Magazine, 42(1), 6-15, 2021.
4. EU-OSHA, Europejska Agencja Bezpieczeństwa i zdrowia w pracy, 2024, https://osha.europa.eu/pl (visited date 24.01.2024).
5. Y. Gheraibia, S. Kabir, K. Aslansefat, I Sorokos, Y. Papadopoulos, “Safety+ AI: A novel approach to update safety models using artificial intelligence”, IEEE Access, 7, 135855-135869, 2019.
6. A. Guzman, S. Ishida, E. Choi, A. Aoyama, “Artificial intelligence improving safety and risk analysis: A comparative analysis for critical infrastructure:, In 2016 IEEE International Conference on Industrial Engineering and Engineering Management (IEEM), IEEE, 471-475, 2016.
7. U. Koese, “Are we safe enough in the future of artificial intelligence? A discussion on machine ethics and artificial intelligence safety”, BRAIN Broad Research in Artificial Intelligence and Neuroscience, 9(2), 184-197, 2018.
8. U. Koese, P. Vasant, “Fading intelligence theory: A theory on keeping artificial intelligence safety for the future”, In 2017 International Artificial Intelligence and Data Processing Symposium, IEEE, 1-5, 2017.
9. M.S. Lee, M. M.Grabowski, G. Habboub, T.E. Mroz, “The impact of artificial intelligence on quality and safety” Global Spine Journal, 10(1_suppl), 99S-103S, 2020.
10. D. Leslie, D. “Understanding artificial intelligence ethics and safety”, arXiv preprint arXiv:1906.05684, 2019.
11. J. Perez-Cerrolaza, J., et al, “Artificial intelligence for safety-critical systems in industrial and transportation domains: A survey”, ACM Computing Surveys, 56(7), 1-40, 2024.

12. B. Shneiderman, “Human-centered artificial intelligence: Reliable, safe & trustworthy”, International Journal of Human–Computer Interaction, 36(6), 495-504, 2020.
13. Y. Wang, S.H. Chung, “Artificial intelligence in safety-critical systems: a systematic review”, Industrial Management & Data Systems, 122(2), 442-470, 2022.
14. K. Wazyńska-Fiok, J. Jawiński, *Niezawodność systemów technicznych*, PWN, Warszawa, 1990.
15. R.V. Yampolskiy (Ed.), *Artificial intelligence safety and security*, CRC Press, 2018.
16. R.V. Yampolskiy, M.S. Spellchecker, “Artificial intelligence safety and cybersecurity”, A timeline of AI failures, 2016, arXiv preprint arXiv:1610.07997.