P H I L O S O P H I C A L D I S C O U R S E S
Prace Naukowe Uniwersytetu Jana Długosza w Częstochowie
2025, t. VII

http://dx.doi.org/10.16926/pd.2025.07.09

Marzena FORNAL
https://orcid.org/0000-0002-0263-9081
Uniwersytet Jana Długosza w Częstochowie

Godna zaufania sztuczna inteligencja w medycynie

Streszczenie

Artykuł podejmuje analizę kluczowych wyzwań etycznych wynikających z implementacji systemów sztucznej inteligencji w medycynie, ze szczególnym uwzględnieniem koncepcji „godnej zaufania AI" (*trustworthy* AI). Autor omawia ramy regulacyjne, takie jak AI Act, rekomendacje WHO i OECD, wskazując ich ograniczenia w kontekście praktyki klinicznej. Artykuł prezentuje także wyzwania prawne dotyczące odpowiedzialności za decyzje AI w medycynie, w tym problem „black-box liability" i luki w istniejących regulacjach. Całość uzupełnia propozycja kierunków rozwoju etycznych ram projektowania AI, które integrują wymogi prawne z wartościami moralnymi, aby wspierać zarówno bezpieczeństwo, jak i podmiotowość pacjenta.

Słowa kluczowe: sztuczna inteligencja w medycynie, godna zaufania sztuczna inteligencja, etyka troski, pryncypializm.

Współczesna medycyna stosuje szeroki wachlarz narzędzi opartych na sztucznej inteligencji w celu wsparcia procesu diagnostycznego i terapeutycznego. Systemy uczące się potrafią analizować ogromne zbiory danych medycznych, rozpoznawać subtelne wzorce w obrazach diagnostycznych, przewidywać przebieg chorób oraz optymalizować dobór metod leczenia. Rozwój tych technologii otwiera nowe możliwości w zakresie poprawy skuteczności terapii, skrócenia czasu diagnozy oraz personalizacji opieki nad pacjentem. Jednocześnie jednak pojawia się szereg pytań o aspekty etyczne związane z ich wdrażaniem – od kwestii odpowiedzialności za decyzje kliniczne podejmowane przez algorytmy, poprzez ochronę prywatności danych medycznych, po zagadnienia równego dostępu do nowoczesnych rozwiązań. W artykule podjęta zostanie analiza kluczowych wyzwań etycznych wynikających z implementacji sztucznej inteligencji w praktyce medycznej, ze szczególnym uwzględnieniem wartości, które powinny przyświecać projektowaniu i stosowaniu tych systemów.

Definicje sztucznej inteligencji

John McCarthy – uznany za twórcę pojęcia „sztuczna inteligencja" – w klasycznej definicji z 1955 roku, zawartej w propozycji projektu badawczego *Dartmouth Summer Research Project on Artificial Intelligence*, określał AI jako „naukę i inżynierię tworzenia inteligentnych maszyn, zwłaszcza programów komputerowych" (McCarthy i in. 1955). McCarthy postrzegał AI jako dziedzinę nauki, której celem jest konstruowanie maszyn zdolnych do wykonywania działań, które – gdyby były realizowane przez człowieka – uznalibyśmy za inteligentne. Jego ujęcie miało zatem charakter porównawczy, odwołując się do ludzkich zdolności poznawczych i zachowań, które można odwzorować w systemach maszynowych.

Współcześnie możemy dostrzec zmianę perspektywy badawczej, w ramach której akcent zostaje przesunięty na funkcjonalność systemów sztucznej inteligencji, rezygnując z odniesień do ludzkiego myślenia. Przykładem ilustrującym tę zmianę jest definicja sztucznej inteligencji zaproponowana przez Kaplana i Haenleina: „Sztuczna inteligencja to zdolność systemu do prawidłowej interpretacji danych zewnętrznych, uczenia się na ich podstawie oraz wykorzystywania tej wiedzy do osiągania konkretnych celów i zadań poprzez elastyczne dostosowywanie się" (2019, s. 17). Definicja ta jest często przytaczana w literaturze, ponieważ kładzie nacisk na proces uczenia się i adaptacji, podkreśla zdolność interpretacji danych oraz działania w oparciu o zdobytą wiedzę.

Problematyka definicji sztucznej inteligencji nadal stanowi wyzwanie nie tylko dla naukowców ale również dla prawodawców. Próby podania legalnej definicji sztucznej inteligencji pojawiają się obecnie na poziomie UE oraz w wielu opracowaniach organizacji międzynarodowych, m.in. OECD i UNESCO .

Przyjęty w maju 2019 r. przez Radę OECD (Organisation for Economic Co-operation and Development) dokument *Recommendation of the Council on Artificial Intelligence* stanowi pierwszy międzyrządowy standard w dziedzinie AI, zatwierdzony przez 42 państwa, w tym wszystkie kraje UE oraz m.in. USA, Kanadę, Japonię i Australię. Ma on charakter miękkiego prawa (*soft law*) – zawiera zbiór zasad i zaleceń, które państwa sygnatariusze zobowiązały się wdrażać w krajowych strategiach dotyczących sztucznej inteligencji. W dokumencie tym AI zdefiniowano następująco: „System sztucznej inteligencji to system oparty na maszynie, który dla wyraźnie określonego zestawu celów otrzymuje dane wejściowe, przetwarza je za pomocą algorytmów i modeli, generując wyniki – takie jak przewidywania, rekomendacje lub decyzje – które mogą wpływać na środowisko fizyczne lub wirtualne" (OECD 2019, s. 7).

Innym istotnym dokumentem jest *Recommendation on the Ethics of Artificial Intelligence* – pierwszy globalny standard etyczny dotyczący AI, przyjęty jednogłośnie w listopadzie 2021 r. przez 193 państwa członkowskie UNESCO. Również ma on status miękkiego prawa. W rekomendacji tej AI została zdefiniowana

w następujący sposób: „Systemy sztucznej inteligencji to systemy technologiczne zdolne do przetwarzania danych w sposób autonomiczny lub półautonomiczny, w celu generowania wyników, takich jak przewidywania, zalecenia lub decyzje, które mogą wpływać na środowisko fizyczne lub wirtualne" (UNESCO 2021). Definicja przyjęta przez UNESCO jest podobna strukturalnie do tej z OECD, jednak dodaje element autonomii lub półautonomii oraz osadza pojęcie AI w kontekście wartości etycznych i praw człowieka.

AI Act – Rozporządzenie (*UE*) *2024/1689 w sprawie sztucznej inteligencji* przyjęty 13 czerwca 2024 r. i opublikowany w Dzienniku Urzędowym UE 12 lipca 2024 r., jest pierwszym na świecie kompleksowym aktem prawnym regulującym sztuczną inteligencję w oparciu o analizę ryzyka. Ma charakter rozporządzenia, co oznacza, że obowiązuje bezpośrednio we wszystkich państwach członkowskich UE, bez potrzeby implementacji do prawa krajowego. W akcie tym przyjęto następującą definicję AI: „System sztucznej inteligencji oznacza system oparty na maszynie, zaprojektowany do działania z różnym stopniem autonomii, który może, dla wyraźnie określonych celów, generować wyniki, takie jak przewidywania, zalecenia lub decyzje, wpływające na środowisko, z którym wchodzi w interakcję" (Parlament Europejski i Rada Unii Europejskiej 2024). Definicja ta rozwija ujęcie zaproponowane wcześniej przez OECD, wprowadzając element różnego stopnia autonomii oraz dodając pojęcie „środowiska interakcji", co pozwala precyzyjniej określić zakres regulacji.

Analiza przedstawionych definicji sztucznej inteligencji wskazuje na ewolucję rozumienia tego pojęcia od jego wczesnych, naukowo-teoretycznych ujęć, po współczesne definicje osadzone w kontekście prawnym i etycznym. Mimo różnic szczegółowych, współczesne akty prawne i rekomendacje organizacji międzynarodowych coraz bardziej zbliżają się do spójnego, funkcjonalnego rozumienia sztucznej inteligencji, co ma ułatwić opracowanie jednolitych zasad jej oceny i nadzoru.

Zastosowania AI w medycynie

W załączniku I do wspomnianego wyżej AI Act znajduje się lista technik i podejść zaliczanych do gałęzi sztucznej inteligencji. Dokument ten wskazuje, że systemy AI mogą wykorzystywać między innymi: 1) metody oparte na logice i wiedzy (systemy ekspertowe, programowanie oparte na regułach), 2) metody statystyczne, optymalizacyjne i wyszukiwania oraz 3) uczenie maszynowe (w tym *deep learning*) (Parlament Europejski i Rada Unii Europejskiej 2024).

a) Metody symboliczne – oparte na logice i wiedzy

Historycznie najstarszą gałęzią sztucznej inteligencji stanowią metody oparte na logice i wiedzy, których rozwój w latach 50. XX wieku przebiegał pod

silnym wpływem paradygmatu symbolicznego. Podejście to opierało się na założeniu, że inteligencję można modelować poprzez operowanie na symbolach reprezentujących elementy świata i relacje między nimi, zgodnie z regułami logiki formalnej. W konsekwencji powstawały systemy przetwarzające jawnie zakodowaną wiedzę, których działanie można było w całości wyjaśnić poprzez analizę ich struktur logicznych i algorytmów. Przykładami tego typu AI są: a) techniki przetwarzania języka naturalnego (NLP) opartego na regułach, b) systemy oparte na wiedzy, c) programy i komputery szachowe (m.in. słynny *DeepBlue*) (Flasiński 2024, s. 328).

Jednym z wczesnych obszarów badań nad SI była komputerowa analiza języka naturalnego. W latach 50. i 60. podejmowano pierwsze próby automatycznego tłumaczenia (m.in. eksperyment Georgetown-IBM z 1954 r., w którym komputer przetłumaczył 60 zdań z rosyjskiego na angielski). Choć wczesne systemy NLP wykorzystywały głównie proste metody przetwarzania leksykalnego i składniowego, stanowiły fundament dla dalszych prac nad reprezentacją semantyki i pragmatyki języka. Rozwój systemów symbolicznych w latach 60. i 70. XX wieku obejmował także próby stworzenia interaktywnych programów konwersacyjnych (chatbotów), które umożliwiały komunikację w języku naturalnym. Ich działanie opierało się na prostych regułach przetwarzania tekstu – dopasowywaniu wzorców i podstawianiu przygotowanych wcześniej odpowiedzi (Baek i in. 2025). Najbardziej znanym przykładem była ELIZA (Weizenbaum 1966), która symulowała rozmowę z psychoterapeutą w stylu terapii skoncentrowanej na kliencie Carla Rogersa. Choć jej mechanizm działania był wyłącznie syntaktyczny – program nie rozumiał sensu wypowiedzi – ELIZA stała się inspiracją dla dalszych badań nad konwersacyjnymi interfejsami w medycynie.

Pierwsze medyczne adaptacje chatbotów regułowych pojawiły się w latach 70. i 80., głównie jako systemy wywiadu medycznego. Mogły one zadawać pacjentowi pytania o objawy w ustalonej kolejności, a następnie – na podstawie z góry określonych reguł – kierować użytkownika do odpowiednich zaleceń lub sugerować lekarzowi możliwe rozpoznania. Systemy tego typu były wykorzystywane m.in. do wstępnej oceny ryzyka chorób zakaźnych, zbierania danych w badaniach epidemiologicznych oraz w prostych programach edukacyjnych dla pacjentów. W porównaniu do współczesnych modeli generatywnych, chatboty regułowe były „sztywne" – każdy możliwy scenariusz dialogu musiał być zaprogramowany ręcznie. Ich przewagą była natomiast przewidywalność i pełna kontrola odpowiedzi, co w kontekście medycznym zmniejszało ryzyko wygenerowania niebezpiecznych lub błędnych informacji (Kaul i in. 2020).

Koncepcja systemów opartych na wiedzy (*knowledge-based systems*) wywodzi się z przekonania, że skuteczność SI zależy od jakości i kompletności bazy wiedzy dziedzinowej oraz mechanizmu wnioskowania. Pierwsze tego typu systemy, rozwijane pod koniec lat 50., wykorzystywały reguły produkcji typu „je-

żeli–to" (*if–then*), umożliwiające generowanie wniosków na podstawie przesłanek. Choć ich możliwości były ograniczone, stanowiły one prototypy późniejszych systemów eksperckich, które w kolejnych dekadach znalazły zastosowanie m.in. w medycynie.

Już w latach 60. i 70. XX wieku rozpoczęto próby adaptacji metod przetwarzania symbolicznego do potrzeb klinicznych. Jednym z pierwszych i najbardziej znanych projektów był system MYCIN (Uniwersytet Stanforda, początek lat 70.), który analizował dane pacjenta (objawy, wyniki badań, masa ciała) i na tej podstawie generował listę prawdopodobnych patogenów bakteryjnych oraz sugerował schemat antybiotykoterapii. System ten działał na bazie około 600 reguł eksperckich, stosując tzw. *backward chaining* (wnioskowanie wsteczne), i choć nigdy nie został wdrożony w praktyce klinicznej ze względu na ograniczenia prawne, jego dokładność w doborze antybiotyków dorównywała specjalistom chorób zakaźnych. Innym pionierskim przykładem był CASNET (*Causal-Associational Network*) opracowany na Rutgers University w latach 70., wykorzystywany w diagnostyce i leczeniu jaskry. CASNET opierał się na modelu sieci, w której reprezentowano relacje między objawami, zmianami patologicznymi i postępem choroby, umożliwiając lekarzowi zarówno ocenę stanu pacjenta, jak i planowanie terapii. Równolegle rozwijano systemy takie jak INTERNIST-1 (Uniwersytet w Pittsburghu), które posiadały znacznie szerszą bazę wiedzy, obejmującą kilkaset chorób wewnętrznych, i służyły wspomaganiu lekarzy podstawowej opieki zdrowotnej w procesie diagnostycznym (Kaul i in. 2020).

Te wczesne systemy symboliczne w medycynie miały jednak istotne ograniczenia: wymagały żmudnego tworzenia i aktualizowania baz wiedzy przez ekspertów, były mało elastyczne w przypadku nowych chorób lub nietypowych objawów, a ich skuteczność zależała od kompletności wprowadzonych reguł. Mimo to odegrały kluczową rolę w ukształtowaniu podstaw sztucznej inteligencji w medycynie oraz zainspirowały rozwój późniejszych systemów eksperckich i metod opartych na uczeniu maszynowym.

b) Metody heurystyczne – optymalizacyjne i wyszukiwania

W sztucznej inteligencji metody heurystyczne to podejścia, które wykorzystują uproszczone reguły, strategie lub funkcje oceny w celu szybszego znalezienia dobrego rozwiązania problemu, bez gwarancji jego globalnej optymalności. Są one szczególnie przydatne w zadaniach, gdzie przestrzeń możliwych rozwiązań jest bardzo duża, a metody dokładne byłyby zbyt czasochłonne lub obliczeniowo kosztowne. Heurystyki przyjmują różne formy, w zależności od rodzaju zadania. W algorytmach wyszukiwania mogą stanowić funkcję szacującą „odległość" od celu, jak w klasycznym algorytmie A*, co pozwala ukierunkować proces poszukiwania najbardziej optymalnych rozwiązań (Hart i in. 1968). W syste-

mach opartych na rozpoznawaniu podobieństwa, takich jak *case-based reasoning* (CBR) czy *content-based image retrieval* (CBIR), heurystyka występuje w postaci miar podobieństwa, które określają, jak bardzo dany przypadek lub obraz przypomina przypadki już znane (Müller i in. 2004). W obszarze optymalizacji szczególną rolę odgrywają tzw. metaheurystyki — ogólne strategie przeszukiwania przestrzeni rozwiązań inspirowane procesami naturalnymi lub społecznymi, takimi jak algorytmy genetyczne, symulowane wyżarzanie, optymalizacja rojowa czy algorytmy mrówkowe (Eiben, Smith, 2015). Istnieją również heurystyki domenowe, opracowane na podstawie wiedzy ekspertów z danej dziedziny, które pozwalają zawęzić obszar poszukiwań do rozwiązań najbardziej prawdopodobnych (Shortliffe, Cimino 2013).

W praktyce klinicznej metody heurystyczne znajdują szerokie zastosowanie. W planowaniu radioterapii mogą posłużyć do automatycznego wyznaczania rozkładu dawki, zapewniającego maksymalną skuteczność leczenia przy ochronie tkanek zdrowych (Mohan i in. 2000). W kardiologii mogą wspierać dobór parametrów programowania urządzeń implantowalnych, takich jak stymulatory serca, w celu poprawy bezpieczeństwa i efektywności terapii (Wilkoff i in. 2015), a w chirurgii robotycznej planowanie trajektorii narzędzi, skracając czas operacji i zmniejszając ryzyko powikłań (Taylor, Menciassi 2016; Misra, Reed, Okamura 2010). W diagnostyce obrazowej heurystyki mogą kierować wyszukiwaniem podobnych przypadków w dużych bazach danych obrazów medycznych, co wspomaga proces diagnostyczny, zwłaszcza w rzadkich lub nietypowych przypadkach (Begum i in. 2011). Tego typu rozwiązania stosowane są również w zarządzaniu ochroną zdrowia — od harmonogramowania bloków operacyjnych (Cardoen, Demeulemeester, Beliën 2010; Ernst i in. 2004), przez planowanie tras karetek (Brotcorne, Laporte, Semet 2003; Aringhieri i in. 2017), po zarządzanie zapasami leków (De Vries, Huijsma 2011; Kelle, Woosley, Schneider 2012). Dzięki swojej elastyczności i zdolności do działania w ograniczonym czasie metody heurystyczne stanowią istotny komponent systemów AI w medycynie, często współpracując z metodami uczenia maszynowego w hybrydowych rozwiązaniach wspomagających decyzje kliniczne (Deo 2015).

c) Płytkie uczenie maszynowe – metody statystyczne i estymacje baysowskie

Pojęcie uczenia maszynowego (*machine learning*, ML) odnosi się do klasy metod obliczeniowych, w których systemy nabywają zdolność wykonywania zadań poprzez identyfikowanie wzorców w danych empirycznych, a nie poprzez odgórne programowanie reguł (Mitchell 1997; Bishop 2006). W ujęciu funkcjonalnym oznacza to, że model „uczy się" odwzorowania pomiędzy przestrzenią wejść a przestrzenią wyjść na podstawie przykładów uczących, a następnie wy-

korzystuje to odwzorowanie do przewidywania, klasyfikacji bądź rekomendowania w nowych, wcześniej nieobserwowanych przypadkach. Kluczową cechą ML jest zatem zdolność do generalizacji — przenoszenia wiedzy nabytej na zbiorze treningowym na populację przypadków spoza tego zbioru (Goodfellow, Bengio, Courville 2016).

W obrębie ML wyróżnia się m.in. nurt określany jako „uczenie płytkie" (*shallow* ML). Obejmuje ono klasyczne algorytmy statystyczno–obliczeniowe, takie jak regresja logistyczna, maszyny wektorów nośnych (SVM), drzewa decyzyjne i ich kompozycje (np. las losowy), metody oparte na najbliższym sąsiedztwie (k-NN) czy klasyfikatory bayesowskie (Hastie, Tibshirani, Friedman 2009). Modele te działają zwykle w oparciu o niewielką liczbę warstw przetwarzania i wymagają uprzedniego zaprojektowania reprezentacji danych przez człowieka (tzw. inżynierię cech) (Domingos 2012). Dopiero na tak przygotowanej reprezentacji algorytm uczy się rozróżniać klasy lub przewidywać ryzyko. Do zalet tego podejścia należy umiarkowane zapotrzebowanie na dane i moc obliczeniową oraz stosunkowo wysoka przejrzystość — łatwiej jest prześledzić, które cechy wpłynęły na decyzję modelu (Murphy 2012). Ograniczeniem jest natomiast trudność w uchwyceniu wszystkich istotnych wzorców w przypadku bardzo złożonych, „surowych" danych, takich jak obrazy medyczne wysokiej rozdzielczości czy długie nagrania sygnałów.

Metody statystyczne płytkiego ML od lat stanowią fundament analizy danych klinicznych. Regresja logistyczna leży u podstaw kalkulatorów ryzyka, takich jak model Framingham, szacujący 10-letnie ryzyko choroby wieńcowej na podstawie wieku, płci, ciśnienia i poziomu cholesterolu (D'Agostino i in. 2008). W chorobach przewlekłych regresja liniowa pozwala modelować tempo zmian parametrów, np. spadku eGFR w przewlekłej chorobie nerek, co ułatwia planowanie dializoterapii (Grams i in. 2022). Podejścia bayesowskie, jak system DXplain, generują listę możliwych rozpoznań z przypisanymi prawdopodobieństwami, wspierając diagnostykę chorób rzadkich (Barnett i in. 1987). W onkologii bayesowska analiza przeżycia umożliwia wcześniejsze i bardziej ostrożne wnioskowanie o skuteczności terapii, uwzględniając cenzorowanie danych (Biard i in. 2021). Integrację danych z różnych źródeł, np. MRI i biomarkerów, wspiera analiza dyskryminacyjna (LDA), klasyfikująca pacjentów i identyfikująca cechy o największym znaczeniu różnicującym (López i in. 2011).

Wspólną cechą metod wymienionych w punktach a), b) i c) jest praca na cechach przygotowanych przez człowieka (np. parametrach EKG, prostych miarach morfologii zmiany), co sprzyja przejrzystości: łatwo ustalić, które informacje zadecydowały o wyniku i jak duża jest niepewność prognozy. Dzięki temu klinicysta nie obcuje z „czarną skrzynką", lecz z narzędziem, którego rozumowanie można audytować i komunikować pacjentowi. Granice między tymi rodzinami metod bywają jednak płynne. Przykładowo, gdy parametry optymalizacyjne są dosto-

sowywane automatycznie na podstawie oznaczonych przykładów (jak w k-NN, SVM czy LDA operujących na przygotowanych zestawach cech), wówczas metoda heurystyczna staje się jednocześnie formą płytkiego uczenia maszynowego. Odwrotnie — modele uczące się można włączać w procesy optymalizacyjne, aby generowały rozwiązania spełniające określone wymagania kliniczne, na przykład zapewniały najlepszy kompromis między skutecznością a bezpieczeństwem terapii.

Z perspektywy regulacyjnej AI Act (Rozporządzenie UE 2024/1689) przyjmuje definicję funkcjonalną systemu sztucznej inteligencji: jest to system maszynowy, który — dla jawnych lub niejawnych celów — dokonuje wnioskowania na podstawie danych wejściowych. Tak sformułowana definicja obejmuje szerokie spektrum technologii, ponieważ klasyfikacja danego rozwiązania jako AI nie zależy od rodzaju zastosowanego algorytmu, lecz od pełnionej funkcji — a więc od tego, czy generuje on decyzje, rekomendacje lub treści intelektualne wpływające na środowisko fizyczne lub wirtualne. W praktyce oznacza to, że pod regulacje AI Act mogą podlegać zarówno systemy oparte na uczeniu maszynowym (ML), jak i podejścia czysto statystyczne, estymacje bayesowskie czy metody wyszukiwania i optymalizacji (wskazane w załącznikach do wniosku z 2021 r.), o ile realizują funkcję inferencyjną.

W obszarze ochrony zdrowia ostateczna klasyfikacja ryzyka systemu AI jest często powiązana z przepisami dotyczącymi wyrobów medycznych (MDR/IVDR). Zgodnie z AI Act, jeżeli system jest wyrobem medycznym lub wspiera podejmowanie decyzji klinicznych, automatycznie zostaje zakwalifikowany jako „high-risk" (art. 6), co wiąże się z koniecznością spełnienia szeregu wymogów, m.in. prowadzenia szczegółowej dokumentacji, oceny ryzyka oraz zapewnienia udziału człowieka w procesie decyzyjnym (*human-in-the-loop*). Takie podejście znacząco poszerza zakres regulacji, obejmując nimi nie tylko zaawansowane modele ML, lecz także relatywnie proste narzędzia oparte na optymalizacji dawki czy wyszukiwaniu podobnych przypadków.

d) Głębokie uczenie maszynowe

Uczenie głębokie (*deep learning*, DL) to poddziedzina uczenia maszynowego, która wykorzystuje wielowarstwowe modele obliczeniowe do automatycznego wydobywania i przekształcania cech z danych wejściowych w celu realizacji zadań takich jak klasyfikacja, regresja, detekcja obiektów czy generowanie treści. Głębokie modele charakteryzują się hierarchiczną strukturą przetwarzania, w której kolejne warstwy uczą się coraz bardziej złożonych i abstrakcyjnych reprezentacji danych. Kluczową cechą tego podejścia jest zdolność modelu do samodzielnej ekstrakcji cech (feature extraction) bez konieczności ręcznego pro-

jektowania ich przez człowieka — w przeciwieństwie do wielu klasycznych metod uczenia maszynowego (LeCun, Bengio, Hinton 2015).

Podstawowym typem modeli stosowanych w uczeniu głębokim są sztuczne sieci neuronowe (*artificial neural networks*, ANN), inspirowane strukturą i zasadami działania ludzkiego mózgu. Ich historia sięga lat 40. XX wieku, kiedy McCulloch i Pitts (1943) zaproponowali pierwszy formalny model neuronu. Na przestrzeni dekad rozwój ANN doprowadził do powstania współczesnych, wysoce złożonych sieci głębokich, stanowiących fundament dla wyspecjalizowanych architektur, takich jak konwolucyjne sieci neuronowe (*convolutional neural networks*, CNN), rekurencyjne sieci neuronowe (*recurrent neural networks*, RNN) czy sieci transformatorowe (*transformers*). Jednak DL nie musi być ograniczone do architektur neuronowych (Schmidhuber 2015).

Konwolucyjne sieci neuronowe stanowią wyspecjalizowany typ głębokich sieci neuronowych, który szybko zyskał status standardu w analizie obrazów. Ich architektura projektowana jest tak, by automatycznie wydobywać cechy z obrazów, zachowując jednocześnie strukturę przestrzenną danych (LeCun, Bengio, Hinton 2015). W medycynie CNN odgrywają kluczową rolę w diagnostyce obrazowej — od MRI, CT, RTG, USG po PET. Dzięki wyżej omawianym właściwościom sieci te osiągają często dokładność porównywalną z ludzkimi specjalistami lub ją przewyższają. Na przykład model opracowany przez McKinney i in. (2020) do analizy mammografii przewyższył radiologów w wykrywaniu raka piersi, redukując fałszywe alarmy o około 5,7%. Podobnie, system Google Health do rozpoznawania retinopatii cukrzycowej osiągnął wartość AUC powyżej 0,99, co świadczy o niezwykle wysokiej precyzji diagnostycznej (Gulshan i in. 2016). Z kolei w diagnostyce raka płuc system DeepLung osiągnął wydajność diagnostyczną porównywalną z doświadczonymi radiologami (Zhu i in. 2018).W radioterapii CNN są wykorzystywane do automatycznej segmentacji narządów krytycznych, co znacząco usprawnia planowanie leczenia (Isensee i in. 2021). W neurologii CNN wspierają analizę obrazów mózgu, m.in. w chorobie Alzheimera, a także, gdzie dokonują segmentacji guzów mózgu (Korolev i in. 2017).

Rekurencyjne sieci neuronowe (RNN) to szczególny rodzaj sztucznych sieci neuronowych zaprojektowany z myślą o przetwarzaniu danych sekwencyjnych, w których kolejność elementów ma istotne znaczenie. W przeciwieństwie do klasycznych sieci neuronowych, które traktują każde wejście niezależnie, RNN posiadają mechanizm pętli zwrotnych umożliwiający przekazywanie stanu ukrytego z poprzednich kroków czasowych do bieżącego. Dzięki temu mogą modelować zależności czasowe oraz kontekstowe, co jest szczególnie istotne w analizie sygnałów fizjologicznych, mowy czy dokumentacji medycznej (Lipton, Berkowitz, Elkan 2016; Sherstinsky 2020). RNN znalazły szerokie zastosowanie w analizie sygnałów EKG, gdzie umożliwiają zarówno wykrywanie arytmii, jak i przewidywanie incydentów sercowych. Przykładem jest praca Hannuna i in. (2019),

w której model LSTM osiągnął dokładność w klasyfikacji 14 typów arytmii przewyższającą kardiologów. Równie często RRN stosowane jest w analizie EEG, m.in. do detekcji napadów padaczkowych oraz monitorowania głębokości znieczulenia w czasie rzeczywistym (Mirowski i in. 2009). W obszarze przetwarzania języka naturalnego (NLP) w medycynie, RNN umożliwiają automatyczną ekstrakcję informacji z elektronicznej dokumentacji medycznej, identyfikację diagnoz czy śledzenie przebiegu leczenia pacjenta (Rajkomar i in. 2018).

Architektura Transformer to przełomowy typ sieci neuronowych zaprezentowany przez Vaswaniego i współautorów w pracy *Attention Is All You Need* (2017), który zrewolucjonizował przetwarzanie sekwencji danych. W odróżnieniu od rekurencyjnych sieci neuronowych (RNN), które przetwarzają dane krok po kroku, transformatory wykorzystują mechanizm uwagi, umożliwiający równoczesne uwzględnianie relacji między wszystkimi elementami sekwencji (Vaswani i in. 2017; Khan i in. 2022). W medycynie transformatory odgrywają coraz większą rolę, zwłaszcza w analizie tekstów klinicznych i integracji danych z różnych źródeł. Modele takie jak BioBERT (Lee i in. 2020) i ClinicalBERT (Alsentzer i in. 2019) zostały dostosowane do przetwarzania języka biomedycznego, umożliwiając automatyczną ekstrakcję informacji z elektronicznej dokumentacji medycznej (EHR), prognozowanie ryzyka powikłań, identyfikację diagnoz czy ocenę stanu pacjenta na podstawie opisów klinicznych. W jednostkach intensywnej terapii modele te wspierają systemy wspomagania decyzji klinicznych, przewidując m.in. ryzyko sepsy czy śmiertelności pacjentów (Huang i in. 2022). Znaczącym obszarem rozwoju jest generowanie i analiza raportów medycznych. Transformatory potrafią automatycznie tworzyć opisy badań obrazowych, np. radiogramów klatki piersiowej (Boecking i in. 2022), a także streszczać historię choroby w formie użytecznej dla konsylium lekarskiego. W zastosowaniach multimodalnych łączą dane tekstowe z obrazowymi i sygnałowymi – przykładem są modele BioViL i MedViT, które integrują opis radiologiczny z obrazem MRI lub RTG w celu lepszej detekcji nieprawidłowości. Ponadto transformatory stosowane są w analizie danych genomowych i proteomicznych, gdzie mechanizm uwagi pozwala na wykrywanie istotnych wzorców w sekwencjach biologicznych, a także w modelach przewidujących odpowiedź pacjenta na leczenie (Ji i in. 2021).

Dzięki zdolności do pracy z dużymi, złożonymi i różnorodnymi zbiorami danych, transformatory – w formie dużych modeli multimodalnych– stanowią obecnie jeden z najważniejszych kierunków rozwoju sztucznej inteligencji w medycynie. Z tego też powodu, w styczniu 2024 roku Światowa Organizacja Zdrowia (World Health Organization, WHO) opublikowała kompleksowe wytyczne dotyczące etyki i zarządzania LMM w sektorze zdrowia. Dokument ten zawiera ponad czterdzieści szczegółowych rekomendacji skierowanych do rządów, twórców technologii oraz dostawców usług medycznych, których celem jest zapewnienie odpowiedzialnego, bezpiecznego i etycznego wdrażania LMM w praktyce kli-

nicznej i badaniach (WHO 2024). WHO podkreśla, że LMM mogą znaleźć zastosowanie w wielu krytycznych obszarach ochrony zdrowia, takich jak diagnostyka i opieka kliniczna, aplikacje wspierające pacjentów, zadania administracyjne, edukacja medyczna, prowadzenie badań naukowych czy rozwój leków. Jednocześnie wskazuje na poważne ryzyka związane z ich stosowaniem: możliwość generowania błędnych lub stronniczych treści, zagrożenia dla prywatności i bezpieczeństwa danych, a także potencjalny wpływ na jakość relacji lekarz–pacjent oraz na umiejętności kliniczne personelu (Gardhouse 2024).

W AI Act (Regulacja UE 2024/1689) nie wspomina się wprost o LLM, ale ich status regulacyjny wynika z ogólnej definicji systemu AI oraz reżimu opartego na ocenie ryzyka. AI Act wyróżnia osobną kategorię dla „*general-purpose* AI" (ogólnego zastosowania), dla której obowiązują specjalne wymogi dotyczące transparentności i ewaluacji. LMM, ze względu na wszechstronność, najprawdopodobniej wpisują się w tę kategorię (Gstrein 2024). Jeśli LMM są wykorzystywane do zadań wpływających na zdrowie i bezpieczeństwo pacjentów (np. wspomaganie diagnostyki klinicznej), klasyfikowane są dodatkowo jako *high-risk* AI. Oznacza to, że podlegają surowym obowiązkom dotyczącym bezpieczeństwa, przejrzystości, dokumentacji, nadzoru człowieka i oceny ryzyka. Warto jednak podkreślić, że otwarty charakter definicji oraz brak doprecyzowania statusu LMM rodzi wyzwania praktyczne. Przykładowo, w 2024 roku Meta zrezygnowała z publicznego udostępnienia swojej multimodalnej wersji modelu LLaMA w Unii Europejskiej, wskazując na niepewność regulacyjną w kontekście AI Act oraz wymogów RODO (Weatherbed 2024).

Godna zaufania sztuczna inteligencja – czyli jaka?

Pojęcie „godnej zaufania sztucznej inteligencji" (*trustworth*y AI) zostało wypracowano na gruncie dokumentów i inicjatyw międzynarodowych – zarówno na poziomie Unii Europejskiej, jak i polityk krajowych, oraz standardów branżowych. Choć szczegółowe definicje różnią się w zależności od źródła, wspólny jest ich normatywno-proceduralny charakter – chodzi o zestaw wymogów prawnych i technicznych, których spełnienie ma zagwarantować, że system AI będzie działał w sposób bezpieczny, zgodny z prawem i wartościami społecznymi.

W wytycznych *Ethics Guidelines for Trustworthy AI* (HLEG 2019), opracowanych dla Komisji Europejskiej, wskazano trzy filary godnej zaufania AI:

— Zgodność z prawem – przestrzeganie obowiązujących regulacji.
— Etyczność – działanie w zgodzie z zasadami etycznymi i wartościami społecznymi.
— Solidność techniczna – bezpieczeństwo, niezawodność, odporność na błędy i ataki.

Te filary rozwinięto w siedmiu wymiarach oceny, obejmujących m.in. nadzór człowieka (*human-in-the-loop*), ochronę prywatności i danych, przejrzystość i wyjaśnialność algorytmów, brak dyskryminacji, pozytywny wpływ społeczny i środowiskowy oraz jasne przypisanie odpowiedzialności. Narzędzie ALTAI (*Assessment List for Trustworthy AI*) przekłada te zasady na pytania kontrolne, pozwalając ocenić w praktyce, czy system spełnia wymogi nadzoru, odporności, przejrzystości, rozliczalności i poszanowania praw człowieka. W *Ramowej Konwencji Rady Europy o AI, prawach człowieka, demokracji i rządach prawa* (2024) akcentuje się konieczność zapewnienia zgodności z prawami człowieka, zasady proporcjonalności, przejrzystości w sektorach wrażliwych, oceny wpływu przed wdrożeniem oraz możliwości kwestionowania decyzji AI.

W polskich dokumentach strategicznych – *Polityce dla rozwoju AI* oraz projekcie *Polityki AI 2025–2030* – zaadaptowano unijne założenia, podkreślając m.in., że odpowiedzialność za działanie systemów AI zawsze ponosi człowiek lub instytucja, a nie sam „system", oraz że kluczowym elementem budowania zaufania jest przeprowadzanie certyfikacji.

Synteza tych źródeł pozwala wyróżnić osiem kluczowych cech „godnej zaufania" AI:

1. Zgodność z prawem i prawami człowieka.
2. Etyczność i brak dyskryminacji.
3. Przejrzystość i wyjaśnialność algorytmów i decyzji.
4. Solidność techniczna i bezpieczeństwo, w tym cyberbezpieczeństwo.
5. Ochrona prywatności i właściwe zarządzanie danymi.
6. Nadzór człowieka i możliwość ingerencji w decyzje AI.
7. Odpowiedzialność i rozliczalność twórców, dostawców i użytkowników.
8. Stałe monitorowanie skuteczności i ryzyka po wdrożeniu.

W kontekście medycznym pojęcie godnej zaufania sztucznej inteligencji w tych dokumentach ma dodatkowe, bardziej rygorystyczne znaczenie, bo dotyczy systemów wysokiego ryzyka wpływających na zdrowie i życie ludzi. Zgodnie z wytycznymi *Ethics Guidelines for Trustworthy AI* (HELG 2019), kluczowym wymogiem jest zapewnienie bezpieczeństwa pacjenta poprzez minimalizację ryzyka błędnej diagnozy lub terapii. Konieczne jest prowadzenie walidacji klinicznej algorytmów na reprezentatywnych zbiorach danych, uwzględniających zróżnicowanie populacji pacjentów pod względem wieku, płci oraz chorób współistniejących. Szczególny nacisk kładzie się na wyjaśnialność (*explainability*) systemów, czyli dostarczanie diagnoz lub rekomendacji w formie zrozumiałej zarówno dla lekarza, jak i pacjenta. Istotną zasadą jest także *human-in-the-loop*, czyli obowiązek nadzoru i weryfikacji decyzji AI przez personel medyczny, oraz zapewnienie ochrony danych zdrowotnych zgodnie z wymogami RODO.

Narzędzie samooceny ALTAI przekłada te zasady na pytania kontrolne, obejmujące m.in. kwestię nadzoru klinicznego, walidacji międzyinstytucjonalnej, do-

kumentowania procesu uczenia algorytmu i jego modyfikacji, a także wdrażania systemów zgłaszania incydentów związanych z użyciem AI. Podobny kierunek wyznacza *Ramowa Konwencja Rady Europy o sztucznej inteligencji, prawach człowieka, demokracji i rządach prawa* (2024), która nakłada obowiązek oceny wpływu na prawa człowieka i zdrowie przed wdrożeniem systemu, wymaga proporcjonalności zastosowania AI względem celu klinicznego, gwarantuje pacjentowi prawo do informacji o wykorzystaniu AI w procesie diagnostycznym oraz możliwość zakwestionowania jej decyzji i uzyskania drugiej opinii.

W polskich dokumentach strategicznych (*Polityka dla rozwoju AI, projekt Polityki AI 2025–2030*) podkreśla się, że sztuczna inteligencja w medycynie powinna wspierać pracę lekarza, a nie go zastępować, oraz że niezbędnym elementem budowania zaufania jest certyfikacja kliniczna zgodna z MDR/IVDR. Ważnym priorytetem jest eliminacja dyskryminacji w algorytmach oraz zapewnienie interoperacyjności z krajowymi systemami elektronicznej dokumentacji medycznej. Z kolei *Biała Księga AI w praktyce klinicznej* (2022) akcentuje konieczność oceny klinicznej wiarygodności systemów w badaniach prospektywnych, zapewnienia odtwarzalności wyników (*traceability*), unikania opóźnień w podejmowaniu decyzji ratujących życie oraz szkolenia użytkowników w zakresie działania i ograniczeń AI.

Normy ISO i CEN, w tym ISO/IEC 42001 oraz ISO 14971, precyzują wymagania operacyjne dotyczące analizy ryzyka klinicznego, wdrażania procedur ciągłego monitorowania skuteczności i bezpieczeństwa systemów po ich implementacji, prowadzenia dokumentacji umożliwiającej audyt oraz stosowania mechanizmów awaryjnych, takich jak manualne przejęcie kontroli nad decyzją AI.

W syntetycznym ujęciu godna zaufania sztuczna inteligencja w medycynie powinna charakteryzować się następującymi cechami:

1. Bezpieczeństwo pacjenta i minimalizacja ryzyka klinicznego.
2. Walidacja kliniczna na reprezentatywnych i wysokiej jakości danych.
3. Transparentność dla lekarza i pacjenta (wyjaśnialność, oznaczenie użycia AI).
4. Ochrona prywatności zgodnie z RODO (szczególna kategoria danych).
5. Nadzór lekarza i prawo pacjenta do drugiej opinii.
6. Brak dyskryminacji i równość dostępu.
7. Śledzenie i audytowalność procesów decyzyjnych AI.
8. Stałe monitorowanie skuteczności po wdrożeniu w praktyce klinicznej.

Obszary problemowe

Bazując na analizie obowiązujących dokumentów regulacyjnych i strategicznych dotyczących sztucznej inteligencji w ochronie zdrowia, można wyróżnić kilka kluczowych obszarów problemowych, które wymagają dalszego doprecyzowania oraz zmian legislacyjnych:

1. Luka w operacjonalizacji kluczowych pojęć

Chociaż takie dokumenty jak *Ethics Guidelines for Trustworthy AI* (HLEG 2019), ALTAI czy AI Act odwołują się do wartości takich jak „przejrzystość", „wyjaśnialność" czy „odpowiedzialność", brakuje im precyzyjnych, mierzalnych definicji tych terminów. Przejrzystość nie została jednoznacznie określona — pozostaje niejasne, czy odnosi się jedynie do dostępu do dokumentacji technicznej, czy też zawiera aspekt zrozumiałości działania systemu przez lekarza lub pacjenta. W kontekście wyjaśnialności (*explainability*), brak rozróżnienia w odniesieniu do konkretnego zastosowania (np. różnice między systemami decyzyjnymi dla lekarzy a systemami administracyjnymi) utrudnia ustalenie odpowiedniego poziomu wyjaśnienia. Odpowiedzialność prawna pozostaje pojęciem ogólnym — bez wskazania konkretnych mechanizmów prawnych egzekwowania w razie błędu AI. W rezultacie brak precyzyjnych wskaźników i metod oceny stanowi poważną barierę dla organów regulacyjnych i audytorów w egzekwowaniu wymogów (Markus, Kors, Rijnbeek, 2020; Cheong 2024).

2. Luka w walidacji klinicznej systemów AI

Chociaż walidacja kliniczna jest często wskazywana jako warunek dopuszczenia systemu do użycia, brak jednoznacznych, obowiązkowych protokołów badań stanowi istotne wyzwanie. Brakuje standaryzacji metodyki testów klinicznych — szczegółowych wymogów co do liczebności próby pacjentów, zróżnicowania demograficznego czy zakresu ośrodków. Wiele systemów jest dopuszczanych wyłącznie na podstawie badań retrospektywnych lub danych z pojedynczych klinik, co zwiększa ryzyko *overfittingu* i ograniczonej generalizacji wyników. Ponadto, brak wymogu badań typu *real-world evidence* uniemożliwia pełną ocenę działania AI w rzeczywistych warunkach klinicznych. Brakuje także kryteriów wymagających międzynarodowej porównywalności wyników walidacji — co ogranicza transgraniczną uznawalność certyfikacji (Morley i in. 2021; Pennestrì 2025).

3. Luka w monitorowaniu po wdrożeniu (post-market surveillance)

Dokumenty UE, takie jak AI Act oraz MDR/IVDR, a także wytyczne WHO, wskazują na potrzebę monitorowania działania AI po wdrożeniu, jednak formułują to w sposób ogólny (European Medical Device Coordination Group 2025; WHO 2024). Brakuje ustalonych minimalnych częstotliwości audytów oraz precyzyjnych metod monitorowania skuteczności i bezpieczeństwa. Producenci zachowują znaczną swobodę w doborze procedur nadzoru (MDCG, 2025). Ponadto, nie ma wymogu raportowania zdarzeń *near-misses*, czyli sytuacji potencjalnie prowadzących do błędu, które zostały wychwycone i skorygowane — choć stanowiłyby one wartościowy materiał do doskonalenia systemu (Babic et al. 2025). Równie płytko rozwinięty pozostaje mechanizm publicznego informo-

wania o problemach AI po wdrożeniu — dostęp pacjentów i środowiska medycznego do raportów o incydentach jest ograniczony (Muralidharan, 2024). Brakuje także centralnego rejestru incydentów związanych z AI w medycynie — analogicznego do systemów funkcjonujących w lotnictwie czy farmakologii (Babic i in. 2025; Dolin i in. 2025).

4. Luka w odpowiedzialności prawnej

W obecnym stanie prawnym brakuje jednoznacznych zasad rozdzielania odpowiedzialności pomiędzy producenta, dostawcę, wdrażającego i użytkownika AI w przypadku szkody pacjenta. Szczególnie problematyczna jest kwestia tzw. *black-box liability*, czyli sytuacji, gdy algorytm podejmuje decyzję w sposób nietranspatentny, a szkoda wynika z błędu systemu. Tradycyjne ramy odpowiedzialności, oparte na koncepcjach wady produktu czy zaniedbania, są niewystarczające w obliczu złożonych i autonomicznych mechanizmów decyzyjnych AI (Duffourc 2023; Ebers 2022). Problemy z tym związane będą się pogłębiać wraz z wzrastającą autonomią systemów typu LMM. EU AI Act i projekt *Dyrektywy o odpowiedzialności za sztuczną inteligencję* (AILD/PLD) wskazują kierunek zmian, m.in. przez uwzględnienie oprogramowania jako produktu i łagodzenie reguł dowodowych, lecz w sektorze zdrowia wciąż brakuje konkretnych, praktycznych wytycznych (Rada Unii Europejskiej 2022a, 2022b).

5. Luka w interoperacyjności i standaryzacji danych

Integracja AI z elektroniczną dokumentacją medyczną w UE stoi przed wyzwaniem braku pełnej harmonizacji norm interoperacyjności, takich jak HL7/FHIR czy DICOM (Bender, Sartipi 2013; Mandel i in. 2016). Dzisiejszy system ochrony zdrowia opiera się na sieci archaicznych systemów i różnorodnych standardach, co utrudnia efektywną wymianę danych między AI a systemami klinicznymi. Dodatkowo brakuje wytycznych dotyczących jakości danych treningowych, zwłaszcza w kontekście brakujących danych (*missing data*) i ich wpływu na wyniki AI (Rajkomar i in. 2019).

6. Luka w ochronie prywatności w praktyce klinicznej

Regulacje skupiają się głównie na zgodności z RODO, ale brakuje szczegółowych wytycznych odnośnie sytuacji, gdy AI wymaga dostępu do dużych, zdecentralizowanych zbiorów danych (np. w modelach *federated learning*). Choć FL pomaga chronić prywatność, nadal wymaga wsparcia zaawansowanymi technikami, takimi jak anonimizacja, szyfrowanie czy audyt (Rieke i in. 2020). Ponadto, mechanizmy uzyskiwania świadomej zgody pacjenta na ponowne użycie jego danych w szkoleniu AI są niewystarczająco określone – co rodzi wątpliwości etyczne i prawne co do autonomii pacjenta (McKeown i in. 2021).

7. LUKA W UWZGLĘDNIANIU SPECYFIKI GENERATYWNEJ AI

W 2024 roku Światowa Organizacja Zdrowia opublikowała pierwsze wytyczne dotyczące *Large Multi-Modal Models* (LMMs) – generatywnych modeli AI zdolnych do przetwarzania różnych rodzajów danych i generowania wielorakich rezultatów (WHO 2024). Jednak w dokumentach UE i Polsce wciąż brakuje szczegółowych przepisów odnoszących się do zastosowania tego typu modeli w medycynie – na przykład w tworzeniu dokumentacji klinicznej czy wspieraniu procesu diagnostycznego. Równie istotna pozostaje luka w regulacjach dotyczących zarządzania ryzykiem występowania tzw. „halucynacji" (nieprawdziwych, generowanych przez model treści) oraz konieczności filtrowania wygenerowanych wyników pod kątem trafności i bezpieczeństwa (Nori i in. 2023).

8. LUKA W EDUKACJI I KOMPETENCJACH UŻYTKOWNIKÓW

Dokumenty akcentują potrzebę nadzoru człowieka, lecz nie precyzują minimalnych standardów szkoleniowych dla lekarzy oraz personelu medycznego w zakresie stosowania AI. Brakuje jednolitych ram edukacyjnych definiujących, jakie kompetencje powinni posiadać użytkownicy AI, by korzystać z niej odpowiedzialnie i efektywnie (Li i in. 2023; Thomas i in. 2021). Badania wyraźnie wskazują potrzebę wprowadzenia AI do nauczania klinicznego — zarówno poprzez tworzenie ram kompetencyjnych, jak i włączanie programów szkoleniowych do programów kształcenia. W dodatku aktualnie brak regulacji określających odpowiedzialność lekarzy, gdy podejmują decyzje opierając się na rekomendacjach AI, których nie są w stanie w pełni zrozumieć lub ocenić (van de Sande i in. 2021).

Dlaczego sama zgodność z regulacjami nie wystarcza?

Sama zgodność sztucznej inteligencji z regulacjami prawnymi nie stanowi wystarczającego kryterium, aby uznać system AI za technologię godną zaufania w kontekście medycyny. Ramy prawne, takie jak AI Act, raporty WHO czy rekomendacje OECD, definiują pojęcie *trustworthy* AI przede wszystkim przez pryzmat formalnych i technicznych wymogów, obejmujących m.in. wyjaśnialność, brak stronniczości, bezpieczeństwo czy zgodność z obowiązującym prawem. Są to kryteria konieczne, zapewniające minimalny poziom ochrony pacjenta, lecz w praktyce klinicznej nie obejmują one pełnego spektrum warunków, w których zaufanie może się rzeczywiście wytworzyć i utrwalić.

Zaufanie, jak podkreśla Baier (1986), ma charakter relacyjny i zawsze wiąże się z podatnością strony ufającej, która oddaje część kontroli nad swoim losem, opierając się na kompetencjach, intencjach i integralności drugiej strony. W me-

dycynie relacja ta jest z natury asymetryczna: pacjent, pozbawiony pełnej wiedzy medycznej i często działający w warunkach stresu oraz choroby, znajduje się w pozycji szczególnej zależności. Im większa jest ta asymetria, tym silniejsze – według Baier – stają się etyczne zobowiązania strony posiadającej przewagę, obejmujące troskę, dobrą wolę i działanie w najlepszym interesie drugiego człowieka.

Podobną perspektywę rozwija etyka troski (Gilligan 1982; Tronto 1993; Nedelensky 2021), dla której uznanie i ochrona podatności drugiej osoby oraz podtrzymywanie relacji umożliwiającej zachowanie godności i podmiotowości stanowią fundament odpowiedzialnego postępowania. Inspiracje te znajdują swoje źródło w myśli Emmanuela Levinasa (1969), który wskazuje, że relacja z Innym rodzi bezwarunkową odpowiedzialność – nie wynikającą z kontraktu transakcyjnego, lecz z samego faktu spotkania z drugim człowiekiem.

Uwzględnienie wymiaru relacyjnego zaufania w procesie projektowania sztucznej inteligencji w ochronie zdrowia wymaga, aby systemy te były konstruowane w sposób, który od początku integruje zasady pryncypializmu (autonomii, nieszkodzenia, dobroczynności i sprawiedliwości) oraz wartości troski i odpowiedzialności. Po pierwsze, oznacza to konieczność zapewnienia transparentności działania, obejmującej zarówno techniczną wyjaśnialność algorytmów, jak i komunikację wyników w formie zrozumiałej dla użytkownika końcowego (Komija Europejscka 2019; Floridi 2019), a także informowanie o roli jaką system AI odgrywa w procesie leczenia. Po drugie, system powinien być projektowany w duchu proaktywnej troski – tak, aby nie tylko minimalizować ryzyko błędu (zasada nieszkodzenia), lecz również wspierać dobrostan pacjenta (zasada dobroczynienia), np. przez uwzględnianie komfortu psychicznego, prywatności i możliwości uzyskania drugiej opinii (Beauchamp, Childress 2019; Vallor 2016). Po trzecie, projektowanie powinno uwzględniać zarządzanie konfliktem wartości wbudowanym w działanie systemu – poprzez implementację procedur ważenia zasad w przypadku ich kolizji, a także poprzez umożliwienie użytkownikowi (pacjentowi lub lekarzowi) podjęcia ostatecznej decyzji w sytuacjach granicznych (Childress 1997; O'Neill 2002). Po czwarte, konieczne jest przeciwdziałanie nadużyciom wynikającym z asymetrii informacyjnej pomiędzy twórcami systemu a jego użytkownikami, m.in. przez jawne informowanie o ograniczeniach danych treningowych, marginesach błędu oraz poziomie niepewności rekomendacji (Floridi 2019; WHO 2021).

Z tej perspektywy „godna zaufania AI" w medycynie to nie tylko system spełniający określone normy regulacyjne, lecz także taki, który aktywnie wzmacnia relację lekarz–pacjent, chroni godność chorego, umożliwia mu odzyskanie poczucia wpływu (np. poprzez dostęp do drugiej opinii czy możliwość zakwestionowania decyzji algorytmu) oraz przekazuje informacje w sposób zrozumiały i dostosowany do indywidualnego kontekstu. Sama zgodność z regulacjami

może zapewnić poprawność proceduralną, lecz dopiero uwzględnienie wymiaru relacyjnego pozwala mówić o technologii rzeczywiście godnej zaufania, która nie tylko optymalizuje proces diagnostyczny lub terapeutyczny, lecz także wspiera fundamentalne więzi społeczne i moralne leżące u podstaw opieki medycznej.

Bibliografia

Alsentzer E, Murphy J.R., Boag W., Weng W.H., Jindi D., Naumann T., McDermott M. (2019), *Publicly available clinical BERT embeddings*. „Proceedings of the 2nd Clinical Natural Language Processing Workshop", 72–78; http://dx.doi.org/10.18653/v1/W19-1909.

Aringhieri R., Bruni M.E., Khodaparasti S., van Essen J.T. (2017), *Emergency medical services and beyond: Addressing new challenges through a wide literature review*. „Computers & Operations Research" 78, 349–368; http://dx.doi.org/10.1016/j.cor.2016.09.016.

Babic B., Chen T., Evgeniou T., Fayard A.-L. (2025), *Managing AI risks in healthcare: From principles to practice*, „Nature Medicine" 31(1), 15–22; http://dx.doi.org/10.1038/s41591-024-03260-9.

Baek G., Cha C., Han J. (2025), *AI chatbots for psychological health for health professionals: Scoping review*, „JMIR Human Factors" 12, e67682; http://dx.doi.org/10.2196/67682.

Baier A. (1986). *Trust and antitrust*, „Ethics" 96(2), 231–260; http://dx.doi.org/10.1086/292745.

Barnett G.O., Cimino J.J., Hupp J.A., Hoffer E.P. (1987), *DXplain—An evolving diagnostic decision-support system*, „JAMA" 258(1), 67–74; http://dx.doi.org/10.1001/jama.1987.03400010073036.

Bender D., Sartipi K. (2013), *HL7 FHIR: An agile and RESTful approach to healthcare information exchange*, „Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems", 326–331; http://dx.doi.org/10.1109/CBMS.2013.6627810.

Begum S., Ahmed M.U., Funk P., Xiong N., Folke M. (2011), *Case-based reasoning systems in the health sciences: A survey of recent trends and developments*, „IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)", 41(4), 421–434; http://dx.doi.org/10.1109/TSMCC.2010.2071862.

Beauchamp T.L., Childress J.F. (2019), *Principles of biomedical ethics*. 8th ed. Oxford: Oxford University Press.

Belowska-Bień K., Bień B. (2021), *Application of artificial intelligence and machine learning techniques in supporting the diagnosis and treatment of neurological diseases*, „Aktualności Neurologiczne" 21, 163–172; http://dx.doi.org/10.15557/AN.2021.0021.

Bishop C.M. (2006), *Pattern recognition and machine learning*, New York: Springer.

Biard L. i in. (2021), *Bayesian survival analysis in oncology clinical trials: Principles and practice*, „Statistical Methods in Medical Research" 30(2), 331–348; http://dx.doi.org/10.1177/1740774516673362.

Boecking B., Usuyama N., Bannur S., Castro D.C., Schwaighofer A., Hyland S.L. i in. (2022), *Making the most of text semantics to improve biomedical vision–language processing*, „Nature Machine Intelligence" 4(10), 920–931; http://dx.doi.org/10.1007/978-3-031-20059-5_1.

Brotcorne L., Laporte G., Semet F. (2003), *Ambulance location and relocation models*, „European Journal of Operational Research" 147(3), 451–463; http://dx.doi.org/10.1016/S0377-2217(02)00364-8.

Cardoen B., Demeulemeester E., Beliën J. (2010), *Operating room planning and scheduling: A literature review*, „European Journal of Operational Research" 201(3), 921–932; http://dx.doi.org/10.1016/j.ejor.2009.04.011.

Cheong B.C. (2024), *Transparency and accountability in AI systems: Safeguarding wellbeing in the age of algorithmic decision-making*, „Frontiers in Human Dynamics" 6, 1421273; http://dx.doi.org/10.3389/fhumd.2024.1421273.

Childress J.F. (1997), *Methods in bioethics: A framework for moral discourse in biomedical policy*, „Journal of Medicine and Philosophy" 22(4), 283–309; http://dx.doi.org/10.1093/jmp/22.4.283.

Deo R.C. (2015), *Machine learning in medicine*, „Circulation" 132(20), 1920–1930; http://dx.doi.org/10.1161/CIRCULATIONAHA.115.001593.

De Vries J., Huijsman R. (2011), *Supply chain management in health services: An overview*, „Supply Chain Management: An International Journal" 16(3), 159–165; http://dx.doi.org/10.1108/13598541111127146.

Domingos P. (2012), *A few useful things to know about machine learning*, „Communications of the ACM" 55(10), 78–87; http://dx.doi.org/10.1145/2347736.2347755.

Duffourc M., Gerke S. (2024), *Decoding US tort liability in healthcare's black-box AI era: Lessons from the EU*. „Stanford Technology Law Review" 27, 1–56. Dostępne na: https://ssrn.com/abstract=4569698.

Ebers M. (red.) (2022), *Algorithms and law*. Cambridge: Cambridge University Press; http://dx.doi.org/10.1017/9781108347846.

Eiben A.E., Smith J.E. (2015), *Introduction to evolutionary computing*. 2nd ed. Berlin: Springer; http://dx.doi.org/10.1007/978-3-662-44874-8.

Ernst A.T., Jiang H., Krishnamoorthy M., Sier D. (2004). *Staff scheduling and rostering: A review of applications, methods and models*. „European Journal of Operational Research" 153(1), 3–27; http://dx.doi.org/10.1016/S0377-2217(03)00095-X.

European Commission (2019), *Ethics guidelines for trustworthy AI*, High-Level Expert Group on Artificial Intelligence. Dostępne na: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

Floridi L. (2019), *The logic of information: A theory of philosophy as conceptual design*, Oxford: Oxford University Press, http://dx.doi.org/10.1093/oso/9780198833635.001.0001.

Gilligan C. (1982), *In a different voice: Psychological theory and women's development*, Cambridge, MA: Harvard University Press.

Goodfellow I., Bengio Y., Courville A. (2016), *Deep learning*, Cambridge, MA: MIT Press.

Grams M.E. i in. (2022), *Estimating time to ESRD using kidney function decline in patients with CKD*, „American Journal of Kidney Diseases" 79(4), 510–519; http://dx.doi.org/10.1053/j.ajkd.2014.07.026.

Gstrein O.J. (2024). *General-purpose AI regulation and the AI Act*, „Policy Review"; http://dx.doi.org/10.14763/2024.3.1790.

Gulshan V. i in. (2016), *Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs*, „JAMA" 316(22), 2402–2410; http://dx.doi.org/10.1001/jama.2016.17216.

Hinton G., Deng L., Yu D., Dahl G., Mohamed A., Jaitly N. i in. (2012), *Deep neural networks for acoustic modeling in speech recognition*, „IEEE Signal Processing Magazine" 29(6), 82–97; http://dx.doi.org/10.1109/MSP.2012.2205597.

Holzinger A., Biemann C., Pattichis C.S., Kell D.B. (2017), *What do we need to build explainable AI systems for the medical domain? Review of the state-of-the-art and needed next steps*. *arXiv preprint*, arXiv:1712.09923.

Jiang F. i in. (2017), *Artificial intelligence in healthcare: Past, present and future*, „Stroke and Vascular Neurology" 2(4), 230–243; http://dx.doi.org/10.1136/svn-2017-000101.

Kourou K., Exarchos T.P., Exarchos K.P., Karamouzis M.V., Fotiadis D.I. (2015), *Machine learning applications in cancer prognosis and prediction*, „Computational and Structural Biotechnology Journal" 13, 8–17; http://dx.doi.org/10.1016/j.csbj.2014.11.005.

Levinás E. (1969), *Totalité et infini*, Haga: Martinus Nijhoff.

Liu X., Rivera S.C., Moher D., Calvert M.J., Denniston A.K. (2020), *Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: The CONSORT-AI extension*. *BMJ*, 370, m3164; http://dx.doi.org/10.1136/bmj.m3164.

Nedelensky R. (2021), *Care ethics and artificial intelligence: Reframing responsibility for machine learning systems*, „AI and Ethics" 1(4), 517–528; http://dx.doi.org/10.1007/s43681-021-00043-8.

O'Neill O. (2002), *A question of trust*, Cambridge: Cambridge University Press, http://dx.doi.org/10.1017/CBO9780511613814.

Rada Unii Europejskiej (2022a), *Artificial Intelligence Act: General approach*. Dostępne na: https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf.

Rada Unii Europejskiej (2022b), *Proposal for a directive on adapting non-contractual civil liability rules to artificial intelligence* (*AI Liability Directive*). Dostępne na: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52022PC0496.

Tronto J.C. (1993), *Moral boundaries: A political argument for an ethic of care*, Nowy Jork: Routledge.

Vallor S. (2016), *Technology and the virtues: A philosophical guide to a future worth wanting*, Oxford: Oxford University Press; http://dx.doi.org/10.1093/acprof:oso/9780190498511.001.0001.

WHO (2021), *Ethics and governance of artificial intelligence for health: WHO guidance*, Genewa: World Health Organization. Dostępne na: https://www.who.int/publications/i/item/9789240029200.

WHO (2024), *Ethics and governance of artificial intelligence for health: Large multi-modal models*, Genewa: World Health Organization. Dostępne na: https://www.who.int/publications/i/item/9789240095724.

Trustworthy Artificial Intelligence in Medicine

Abstract

The article examines the issue of ethical evaluation and implementation of artificial intelligence systems in medicine, with particular emphasis on the concept of "trustworthy AI." It discusses regulatory frameworks such as the AI Act, and the recommendations of the WHO and OECD, highlighting their limitations in the context of clinical practice. The article also presents legal challenges related to liability for AI-driven decisions in medicine, including the problem of "black-box liability" and gaps in existing regulations. The discussion is complemented by a proposal for the development of ethical design frameworks for AI that integrate legal requirements with moral values, in order to support both patient safety and patient autonomy.

Keywords: Artifical inteligence in medicine, trustworthy artifical inteligence, principialism, ethics of care.