Explainable AI (XAI) to szybko rozwijająca się dziedzina uczenia maszynowego, która ma na celu wyjaśnienie, jak podejmowane są decyzje przez modele AI. Oto kilka pakietów Pythona, które mogą pomóc w wyjaśnianiu twoich modeli:

SHAP (SHapley Additive exPlanations): SHAP łączy teorię gier z lokalnymi wyjaśnieniami, łącząc kilka wcześniejszych metod [LIME, DeepLIFT, itp.] i reprezentując jedyną możliwą spójną i lokalnie dokładną metodę atrybucji cech opartą na oczekiwaniach.

LIME (Local Interpretable Model-Agnostic Explanations): LIME to nowa technika wyjaśniania, która wyjaśnia prognozy dowolnego klasyfikatora w zrozumiały i wierny sposób, ucząc się modelu interpretowalnego lokalnie wokół prognozy.

eli5: eli5 to pakiet Pythona, który pomaga debugować klasyfikatory uczenia maszynowego i wyjaśniać ich prognozy.

PyCaret: PyCaret to otwarta biblioteka uczenia maszynowego w Pythonie, która automatyzuje przepływy pracy uczenia maszynowego.

Alibi: Alibi to otwarta biblioteka Pythona skierowana na inspekcję i interpretację modeli uczenia maszynowego.

DALEX: DALEX to zestaw narzędzi do wyjaśniania, eksploracji i debugowania modeli predykcyjnych. Główna funkcja explain() tworzy opakowanie wokół modelu predykcyjnego. Wynikowy obiekt ma zestaw ogólnych metod do eksploracji zachowania modelu.

XAI: XAI to biblioteka uczenia maszynowego, która została zaprojektowana z myślą o wyjaśnialności AI. XAI zawiera różne narzędzia, które umożliwiają analizę i ocenę danych i modeli.

Sztuczna inteligencja (AI) jest obecnie wykorzystywana w szerokim zakresie zaawansowanych zastosowań, jednak rezultaty wielu modeli AI są trudne do zrozumienia i zaufania ze względu na ich "czarną skrzynkę". Zazwyczaj istotne jest zrozumienie przyczyn podejmowania decyzji przez model AI. Dlatego pojawiła się potrzeba metod eXplainable AI (XAI) mających na celu zwiększenie zaufania do modeli AI. XAI stała się popularnym tematem badawczym w dziedzinie AI w ostatnich latach. Istniejące prace przeglądowe poruszały koncepcje XAI, jej ogólne terminy i metody wyjaśniania ex post, ale nie było żadnych recenzji, które zajęły się metodami oceny, dostępnymi narzędziami, zbiorami danych XAI i innymi związanymi aspektami. Dlatego w tej kompleksowej pracy przedstawiamy czytelnikom przegląd obecnych badań i trendów w tej dynamicznie rozwijającej się dziedzinie na przykładzie studium przypadku. Praca rozpoczyna się od wyjaśnienia tła XAI, wspólnych definicji i podsumowania ostatnio proponowanych technik XAI dla nadzorowanego uczenia maszynowego. Przegląd dzieli techniki XAI na cztery osie za pomocą hierarchicznego systemu kategoryzacji: (i) wyjaśnialność danych, (ii) wyjaśnialność modelu, (iii) wyjaśnialność ex post, oraz (iv) ocena wyjaśnień. Wprowadzamy również dostępne metryki oceny, a także otwarte pakiety oprogramowania i zbiory danych, wraz z kierunkami przyszłych badań. Następnie zarysowano znaczenie wyjaśnialności w kontekście wymogów prawnych, punktów widzenia użytkowników i orientacji aplikacji, określonych jako zagadnienia XAI. Niniejszy artykuł popiera dostosowywanie treści wyjaśnień do konkretnych typów użytkowników. Przeprowadzono badanie technik XAI i oceny, analizując 410 istotnych artykułów, opublikowanych między styczniem 2016 a październikiem 2022 roku, w renomowanych czasopismach i korzystając z szerokiej gamy baz danych naukowych jako źródła informacji. Artykuł skierowany jest do badaczy XAI, którzy interesują się sprawianiem, by ich modele AI były bardziej godne zaufania, a także do badaczy z innych dziedzin, którzy poszukują skutecznych metod XAI do wykonywania zadań z pewnością siebie, jednocześnie komunikując znaczenie danych.

1.2. Cel XAI

Głównym celem XAI jest uzyskanie modeli zrozumiałych dla ludzi, zwłaszcza w aplikacjach z sektorów wrażliwych, takich jak wojsko, bankowość i aplikacje medyczne, ponieważ specjaliści branżowi potrzebują pomocy w skutecznym rozwiązywaniu problemów, ale chcą także otrzymać sensowne wyniki, aby zrozumieć i zaufać tym rozwiązaniom. Przeglądanie odpowiednich wyników nie jest korzystne tylko dla specjalistów branżowych, ale także dla deweloperów, gdy wyniki okazują się nieprawidłowe, co skłania ich do zbadania systemu. Metody AI umożliwiają: (i) ocenę obecnej wiedzy, (ii) rozwijanie wiedzy oraz (iii) rozwijanie nowych założeń/teorii [22].

Ponadto cele metod XAI, które badacze chcieliby osiągnąć poprzez wyjaśnialność, to usprawiedliwienie, kontrola, poprawa i odkrycie [21]. Poniższa lista podsumowuje korzyści wynikające z otwarcia okna na te systemy "czarnej skrzynki" [18]:

• Wzmacnianie osób w walce z negatywnymi skutkami zautomatyzowanego podejmowania decyzji.

• Pomoc osobom w podejmowaniu bardziej świadomych wyborów.

• Ujawnianie i ochrona podatności bezpieczeństwa.

• Integracja algorytmów z ludzkimi wartościami jest ważnym celem.

• Poprawa standardów branżowych dotyczących rozwoju produktów zasilanych przez AI, co z kolei zwiększa zaufanie konsumentów i firm.

• Wprowadzenie polityki prawa do wyjaśnień.

Aby model został zaakceptowany przez użytkowników końcowych i branże, musi być godny zaufania [23]. Budowanie wiarygodnego modelu jest jednak trudne. Kilka czynników, które przyczyniają się do wiarygodności modelu, to sprawiedliwość [24], odporność [25], interpretowalność [26] oraz wyjaśnialność/interpretowalność [27]. Wyjaśnialność jest jednym z najważniejszych aspektów. Istniejące badania skupiały się wyłącznie na dostarczaniu lepszych wyjaśnień i wglądu w celu przyszłych badań. Badacze zaproponowali różne strategie wyjaśniania modeli AI, używając zrozumiałego tekstu [28], matematyki [29] lub wizualizacji [15]. W następnym podrozdziale omówimy nasze motywacje do przeprowadzenia tego badania.

2 . Studia tła

Zainteresowanie badaniami w dziedzinie XAI odżyło. W 2019 roku Mueller i inni [32] opublikowali systematyczną analizę metod XAI i systemów wyjaśniania, które zostały sklasyfikowane na trzy generacje: (i) Systemy pierwszej generacji próbowały opisać w sposób explicite wewnętrzny proces działania systemu poprzez integrowanie wiedzy ekspertów w reguły poprzez przekształcanie tych reguł w wyrażenia języka naturalnego, takie jak te stosowane w systemach ekspertowych z początku lat 70., (ii) Systemy drugiej generacji to systemy człowiek-komputer, które zapewniają pomoc poznawczą, skupiając się na ludzkiej wiedzy i zdolnościach rozumowania z początku lat 2000, i (iii) Systemy trzeciej generacji starają się wyjaśnić wewnętrzne działanie systemów tak, jak pierwsza generacja. Jednak systemy trzeciej generacji stały się głównie systemami czarnej skrzynki od około 2012 roku. Dzięki ulepszonej technologii komputerowej, kilka nowych koncepcji wyjaśniania wyborów stało się bardziej wykonalnych. Te pomysły pojawiły się w odpowiedzi na potrzebę przede wszystkim odpowiedzialnych, sprawiedliwych i godnych zaufania procesów i decyzji. Trzy generacje systemów inteligentnych zostaną szczegółowo omówione w następujących paragrafach.

Systemy Pierwszej Generacji. Badacze interesowali się zrozumieniem działania systemów AI od wczesnych stadiów rozwoju systemów AI. Chandrasekaran i inni [33], oraz Swartout i inni [34,35] byli wśród pierwszych, którzy opisali proces podejmowania decyzji w opartych na wiedzy i systemach ekspertowych. Systemy ekspertowe, ale także Zbiory Rozmyte i System (FSS) [36–38] w latach 70., Sztuczne Sieci Neuronowe (ANN) [39] i Sieci Bayesowskie (BN) [40] w latach 80., a także systemy rekomendacji [41,42] w latach 2000 wzbudziły zainteresowanie wyjaśnieniami inteligentnych systemów. Pomimo ich matematycznej poprawności, wcześniejsze prace są niedokładne. Jednak zainspirowały one kolejne badania nad zrozumieniem systemów inteligentnych.

Systemy Drugiej Generacji. Ta generacja przyniosła budowę bardziej potężnego systemu AI. Jednak zbudowane modele stały się złożone. Proces podejmowania decyzji w systemach AI był taki, że ludzie, w tym eksperci dziedzinowi, nie do końca go rozumieli, gdy chodziło o potężne modele klasyfikacyjne ML szkolone na dużych zbiorach danych z infrastrukturą o wysokiej wydajności [43]. Innym ryzykiem związanym z tymi technikami jest to, że mogą one przypadkowo podejmować błędne decyzje w wyniku uprzednich artefaktów lub fałszywych korelacji w danych. Jest to szczególnie istotna kwestia, jeśli chodzi o wykorzystanie tych systemów w aplikacjach o wysokim ryzyku, takich jak pojazdy autonomiczne i medycyna, gdzie pojedyncza błędna decyzja może skutkować śmiercią człowieka [44].

Systemy Trzeciej Generacji. Obecne postępy w dziedzinie AI, jej zastosowanie w różnorodnych dziedzinach, obawy dotyczące niestosownego wykorzystania [45], braku przejrzystości [46] i niezamierzonych uprzedzeń [47] w modelach są głównymi powodami zwiększonego zainteresowania badaniami XAI. Ma to wpływ nie tylko na ilość informacji, która może kierować etycznymi decyzjami, ale także na odpowiedzialność, bezpieczeństwo i odpowiedzialność przemysłową tych systemów XAI. Ponadto nowe przepisy uchwalane przez różne kraje oznaczają jeszcze większą potrzebę badań XAI w celu budowania zaufania do modeli AI [48]. Modele AI wykorzystywane w wrażliwych obszarach badań naukowych, takich jak zdrowie, biologia i nauki społeczno-ekonomiczne, muszą być w stanie wyjaśnić swoje wyniki, aby umożliwić odkrywanie naukowe i postępy w badaniach.

Istniejące prace XAI można kategoryzować w różny sposób, w tym zastosowania XAI, wielodyscyplinarny łączenia metod, a także wyjaśnialność poprzez modyfikację wewnętrznej funkcjonalności, między innymi. Następny podrozdział przejdzie przez przeanalizowaną literaturę.

2.2. Powiązane badania

Mimo szybkiego wzrostu liczby badań nad XAI (patrz tabela 1), nadal brakuje wszechstronnych przeglądów i systematycznej klasyfikacji tych badań, z wyjątkiem [82]. Istnieje wiele artykułów przeglądowych na temat XAI, ale większość z tych przeglądów koncentruje się na ogólnych technikach XAI, ich znaczeniu i podejściach do oceny. Na przykład Doshi i inni [15] przedstawiają drogę do zdefiniowania i rygorystycznej oceny interpretowalności. Ich głównym wkładem jest taksonomia służąca do oceny interpretowalności. Autorzy skupili się tylko na jednym elemencie wyjaśnialności, tj. interpretowalności i jej związanych z nią technikach oceny. Abdul i inni [79] stworzyli sieć cytowań na podstawie obszernego korpusu badań nad wyjaśnialnością opartych na 289 głównych artykułach i 12 412 cytowanych publikacjach. Jednak ich przegląd dotyczył głównie badań z zakresu interfejsu człowiek-komputer (HCI), które skupiają się na wyjaśnialności. Adadi i inni [21] próbowali dostarczyć informacji na temat idei, motywów i konsekwencji leżących u podstaw badań XAI, aby zrozumieć ważne tematy w dziedzinie XAI.

Ponadto Guidotti i inni [18] przeanalizowali różnorodne podejścia do wyjaśniania dużych modeli czarnej skrzynki, obejmujące techniki eksploracji danych i ML. Autorzy przedstawili obszerną taksonomię metod wyjaśniania dla systemów, które borykają się z problemem czarnej skrzynki. Ich praca kompleksowo oceniła modele ML pod względem XAI, jednak skupiła się tylko na procesach interpretowalności, pomijając inne elementy wyjaśnialności, takie jak ocena. W rezultacie, mimo wszechstronnego technicznego przeglądu rozważanych podejść, trudno było od razu uzyskać ogólne zrozumienie XAI. Samek i inni [80] opisali dwie metody interpretacji wyników modelu. W ich podejściu najpierw oblicza się czułość wyniku w odniesieniu do zmian w danych wejściowych. Drugim krokiem jest rozłożenie decyzji wynikowej na zmienne wejściowe. Dosilović i inni [64] podkreślili najnowsze osiągnięcia w dziedzinie XAI, aby zapewnić uczciwe porównanie między interpretowalnością a wyjaśnialnością w nadzorowanym ML. Lipton [28] zdefiniował właściwości i techniki modelu oraz pojęcie interpretowalności dla nadzorowanego uczenia maszynowego w kontekście przejrzystości identyfikacyjnej dla ludzi oraz wyjaśnień późnych.

W ostatnich latach Carvalho i inni [77] badali interpretowalność ML, skupiając się na uznanych technikach i metrykach. Vilone i inni [75] podzielili popularne metody XAI na cztery kategorie: artykuły przeglądowe, teorie i koncepcje, metodyki i oceny. Arrieta i inni [63] omówili osiągnięcia XAI pod względem wysiłku i wkładów. W ich przeglądzie omówiono dwa podejścia taksonomiczne do wyjaśnialności: (i) przejrzystość modelu ML, i (ii) wyjaśnialność późna. Linardatos i inni [73] przeprowadzili badania skupiające się na technikach interpretowalności ML, w szczególności analizie literatury i taksonomii metod interpretowalności, a także analizowali związki z implementacjami programistycznymi. Ponadto Li i inni [26] opisali i zdefiniowali dwa kluczowe pojęcia XAI: interpretacje i interpretowalność. Autorzy użyli nowej taksonomii do opisu architektury kilku algorytmów interpretacyjnych, a także podkreślili pewne inicjatywy badawcze dotyczące interpretacji. Poza próbą zrozumienia wyników interpretacji, przeszli dalej, aby zbadać określone miary wydajności do oceny algorytmów interpretacji. Langer i inni [74] przyjrzeli się głównym grupom interesariuszy, które poszukują wyjaśnialności AI, oraz ich potrzebom. Confalonieri i inni [83] przedstawili perspektywę historyczną XAI, gdzie przeanalizowali ewolucję pojęcia wyjaśnialności od systemów ekspertowych do uczenia maszynowego i systemów rekomendacyjnych, aż po AI neuro-symboliczną.

Podsumowując, tabela 1 zawiera podsumowanie istniejących artykułów przeglądowych, z których możemy wyciągnąć dwie główne wnioski. Po pierwsze, większość przeglądów dotyczyła trendów badawczych, podstawowych koncepcji XAI (i pojęć związanych z nimi), ich obaw oraz wyjaśnialności późnej. Pomimo faktu, że wielu badaczy skupiło się na kwestiach XAI i terminologii z nią związanej, wciąż pojawiają się regulacje rządowe dotyczące narzucania wyjaśnialności, a także nieakceptowalne i niekonsekwentne definicje ze strony społeczności XAI. Po drugie, liczni badacze zidentyfikowali trzy istotne i trwające wyzwania związane z XAI: (i) brak narzędzi XAI, (ii) różne osie lub wymiary wyjaśnialności, oraz (iii) konieczność poważnego potraktowania oceny XAI (zarówno automatycznych metryk, jak i oceny ludzkiej) w przyszłych kierunkach.

3. Pojęcia wyjaśnialności i ważna nomenklatura

Sztuczna inteligencja (AI) to potężna technologia o szerokim zakresie zastosowań. AI osiągnęła dużą dokładność nie tylko dzięki poprawie wydajności sprzętu, ale także dzięki wykorzystaniu bardziej zaawansowanych algorytmów, takich jak te stosowane w nowoczesnych metodach Deep Learning (DL). Ze względu na złożoną naturę używanych algorytmów, te nowoczesne systemy AI nie są w stanie wyjaśnić swoich decyzji w prosty sposób, co ogranicza ich praktyczne zastosowanie [84]. W rezultacie AI musi zmierzyć się z tym problemem czarnej skrzynki, nawet jeśli deweloperzy muszą zrezygnować z wydajności. Konieczność wyjaśnienia AI i zachęcenia do jego przyjęcia przez wielu interesariuszy zainspirowała powstanie XAI jako nowej dziedziny badań. Ta sekcja jest zorganizowana w następujący sposób: (i) zdefiniowano koncepcję XAI poprzez badania wstępne, (ii) wyjaśniono związane z XAI terminy, oraz (iii) zbadano kompromis między dokładnością a interpretowalnością.

3.1. Koncepcje XAI

Van Lent i inni [85] stworzyli koncepcję XAI w 2004 roku, aby scharakteryzować zdolność swojego systemu do wyjaśniania działań kontrolowanych przez AI jednostek w aplikacjach symulacyjnych gier. Środowisko naukowe i praktycy niedawno ożywili zainteresowanie tematem XAI [86,87]. Kilka grup badawczych badało pojęcie wyjaśnialności w podejmowaniu decyzji przez AI. Każda społeczność badawcza podchodzi jednak do tego zagadnienia z różnych perspektyw i udziela wyjaśnień o różnych znaczeniach. Słowo wyjaśnialność, w kontekście pojęć AI, oznacza funkcjonalną wiedzę na temat modelu, którego celem jest próba opisania zachowania czarnej skrzynki modelu [28]. Często używa się go zamiennie z terminem interpretowalność w literaturze. Wyjaśnialność wyraża to, co dzieje się w modelu, zapewniając czytelne dla człowieka wyjaśnienie decyzji modelu. Jednak trudno jest opracować precyzyjny opis tego, co kwalifikuje się jako wyjaśnienie. Oto kilka najbardziej uznanych definicji wyjaśnienia [53]:

• Wyjaśnienie to przypisanie przyczynowości [88].

• Wyjaśnienie to opisanie czegoś i udzielenie odpowiedzi na pytanie, dlaczego to opisanie czegoś jest poprawne [89].

• Wyjaśnienie to proces znajdowania lub tworzenia wspólnego znaczenia [90].

Nowsza, szeroko akceptowana definicja wyjaśnialnej sztucznej inteligencji została podana w [63], gdzie nacisk kładzie się na odbiorcę wyjaśnienia: wyjaśnialna sztuczna inteligencja to taka, która dostarcza szczegółów lub powodów, aby jej funkcjonowanie było jasne lub łatwe do zrozumienia. W zależności od konkretnego odbiorcy, Miller [52] uważa również, że obecnie dostępne wyjaśnienia są zbyt statyczne. Idealne wyjaśnienie to takie, w którym osoba wyjaśniająca i odbiorca wyjaśnień wzajemnie oddziałują ze sobą. Autor zasugerował, że wyjaśnienia są społeczne i powinny być komunikowane w interaktywny sposób z użytkownikami. W tym samym duchu Grice stworzył zasady współpracy [91] oraz cztery maksymy, których należy przestrzegać podczas wyjaśnień:

1 Jakość: Upewnij się, że wyjaśnienie ma dobrą jakość z następującymi właściwościami:

• Nie dostarczaj przypadkowego wyjaśnienia, które może nie być prawdziwe, oraz

• Nie dostarczaj wyjaśnienia, które nie ma wystarczających dowodów.

2 Ilość: Dostarcz odpowiednią ilość informacji w wyjaśnieniu z następującymi właściwościami:

• Wyjaśnienie musi być informacyjne, to znaczy, dostarczaj tyle informacji, ile jest potrzebne, oraz

• Jednocześnie, nie dostarczaj więcej informacji, niż jest to konieczne.

3 Relacja: Wyjaśnienie musi zawierać tylko informacje istotne dla dyskusji. Ta maksyma może być wykorzystana do poprawy ilości wyjaśnienia.

4 Sposób: Zamiast tego, co jest dane, sposób odnosi się do tego, jak informacje są dostarczane. Grice [91] podzielił to na kilka maksym:

• Unikaj niejednoznacznego języka w wyjaśnieniu.

• Unikaj niejasności w wyjaśnieniu.

• Unikaj długotrwałości za pomocą zwięzłego wyjaśnienia.

• Unikaj informacji, które nie są uporządkowane.

5 Zorientowany na kontekst: Wyjaśnienia dla programistów różnią się od tych dla regulatorów, którzy różnią się od tych dla użytkowników końcowych [92].

Ponadto, XAI nie jest jednolitą jednostką; obejmuje kilka powiązanych ze sobą zasad. Zgodnie z przeanalizowanymi badaniami, istnieją różne koncepcje przyczyniające się do wyjaśniania systemów AI. Chociaż mogą wystąpić pewne nakładanie się między tymi koncepcjami, uważamy, że odzwierciedlają one wiele motywów wyjaśnialności. W następnym akapicie omówimy kilka koncepcji, które wyjaśniają standardową definicję.

3.2. Nomenklatura XAI

Problem czarnej skrzynki w AI odnosi się do trudności systemu w udzieleniu sensownego wyjaśnienia, jak system doszedł do decyzji. Słowa czarna skrzynka, szara skrzynka i biała skrzynka są używane w informatyce i inżynierii do określania różnych poziomów zamknięcia wewnętrznych elementów systemu.

Zasada wyjaśnialności jest ściśle związana z zasadą interpretowalności. Metody interpretowalne są wyjaśnialne, jeśli ludzie mogą zrozumieć ich działanie. Pomimo że wyjaśnialny jest słowem kluczowym w nomenklaturze XAI, termin interpretowalny jest bardziej powszechnie stosowany w społeczności ML niż wyjaśnialny. Powiązane terminy są zdefiniowane następująco:

Definicja 3.2.1 (Wyjaśnialność). Proces wyjaśniania lub ujawniania mechanizmów podejmowania decyzji przez modele. Użytkownik może zobaczyć, jak dane wejściowe i wyjściowe są matematycznie powiązane. Odnosi się to do zdolności do zrozumienia, dlaczego modele AI podejmują swoje decyzje. Zdolność do dokonywania automatycznych interpretacji i opisywania wewnętrznych mechanizmów systemu AI w ludzkiej formie nazywana jest wyjaśnialnością. Wyjaśnialna technika podsumowuje powody decyzji modelu AI. Ponadto „wyjaśnialność pośrednia modelu” odnosi się do metod/algorytmów używanych do wyjaśnienia decyzji modelu AI [21,26,27].

Definicja 3.2.2 dotyczy koncepcji wyjaśnialności w sztucznej inteligencji oraz związanej z nimi terminologii. Podkreśla kluczowe definicje i elementy, które przyczyniają się do zrozumienia pojęcia wyjaśnialnej sztucznej inteligencji (XAI). Oto podsumowanie głównych punktów:

Interpretowalność odnosi się do zrozumienia sposobu podejmowania decyzji przez modele AI, zwłaszcza w przypadku skomplikowanych modeli „czarnej skrzynki”. Techniki wewnętrzne modelu mają na celu ujawnienie wewnętrznych właściwości modeli głębokiego uczenia się.

Przejrzystość jest istotna w dostarczaniu dla ludzi czytelnych wyjaśnień decyzji modelu. Pomaga w ocenie jakości decyzji modelu i w obronie przed atakami adwersaryjnymi.

Sprawiedliwość dotyczy zdolności modeli AI do podejmowania bezstronnych decyzji, nie faworyzując żadnych konkretnych grup reprezentowanych w dystrybucji danych wejściowych. Ważne jest rozwiązywanie uprzedzeń w zbiorach danych i algorytmach, aby osiągnąć sprawiedliwość.

Odporność mierzy wrażliwość systemu na zmiany wejściowe i jego zdolność do poprawnego funkcjonowania w sytuacjach niepewności. Utrzymywanie stabilności i radzenie sobie z atakami adwersaryjnymi to kluczowe czynniki zapewnienia odporności.

Różne inne koncepcje, takie jak kompletność, pouczającość, usprawiedliwienie, monotoniczność, odwracalność, prostota, niezawodność, przenaszalność, satysfakcja, stabilność i odpowiedzialność, przyczyniają się do ogólnego zrozumienia XAI.

Dyskusja podkreśla wzajemne powiązania tych koncepcji i jak razem przyczyniają się do budowania zaufania i przejrzystości w systemach AI, promując odpowiedzialne praktyki AI.

Co do kompromisu między dokładnością a interpretowalnością, zaznaczono, że skomplikowane modele zazwyczaj są mniej interpretowalne, dlatego wyjaśnialność staje się istotna dla zrozumienia tak skomplikowanych systemów. Osiągnięcie równowagi między dokładnością a interpretowalnością jest istotnym aspektem skutecznej analizy, zwłaszcza że organizacje coraz bardziej polegają na modelach AI do różnych zastosowań.

Ponadto, omówiono znaczenie wyjaśnialności w kontekście przepisów rządowych, perspektyw użytkowników i praktycznych zastosowań, podkreślając jej znaczenie w zapewnianiu zaufania godnym systemom AI.