[Artykuł recenzowany]

MICHAŁ ARASZKIEWICZ

Sztuczna Inteligencja i prawo do wyjaśnienia[1]

AUTOR
jest prawnikiem, pracownikiem Katedry Teorii Prawa Wydziału Prawa i Administracji Uniwersytetu Jagiellońskiego.

SŁOWA KLUCZOWE
Sztuczna Inteligencja i prawo, wyjaśnialna Sztuczna Inteligencja, maszynowe uczenie się, prawo do wyjaśnienia, wpływ algorytmów na społeczeństwo

DOI:
10.26368/17332265-044-4-2018-3

ABSTRAKT
Celem artykułu jest przedstawienie podstawowych problemów związanych z koncepcją wyjaśnialności (interpretowalności, przejrzystości) algorytmów Sztucznej Inteligencji (SI), ze szczególnym uwzględnieniem kwestii stosowania systemów uczących się do przetwarzania informacji prawnej. Funkcjonowanie programów komputerowych uczących się jest obecnie doniosłym problemem społecznym, ponieważ na podstawie uzyskiwanych przez nie rezultatów są podejmowane decyzje w różnych sferach życia społecznego (działalność gospodarcza, stosunki pracy, więziennictwo), a sposób ich działania jest trudny do wytłumaczenia nawet dla specjalistów i w zasadzie niemożliwy do zrozumienia dla osób, których dotykają skutki tych decyzji. W związku z tym postuluje się skonstruowanie podmiotowego prawa do wyjaśnienia działania algorytmu. Uwarunkowania przetwarzania informacji prawnej stawiają wyższe wymagania – nie tylko wyjaśnienia, ale także racjonalnego uzasadnienia decyzji. Aktualne jest pytanie o to, czy algorytmy SI mogą być pomocne w poszukiwaniu takich uzasadnień.

W historii badań nad Sztuczną Inteligencją naprzemiennie występują okresy wzmożonego zainteresowania społeczeństwa tą problematyką oraz okresy tak zwanych zim, kiedy zazwyczaj – w związku z niespełnionymi oczekiwaniami – dziedzina ta staje się przede wszystkim obiektem zainteresowania ekspertów (Nilsson 2010). Obecnie mamy do czynienia z kolejnym okresem pierwszego typu, ale z uwagi na dynamikę rozwoju technologii informatycznych oraz ich rolę we współczesnym społeczeństwie dyskusja wokół SI ma szerszy zasięg niż w przeszłości. Pojęcie SI weszło do dyskursu potocznego. Zostały stworzone systemy informatyczne, które naturalnie są określane zarówno przez laików, jak i przez specjalistów jako wyposażone w elementy inteligencji. Chodzi tu przede wszystkim o programy komputerowe wykorzystujące różne mechanizmy maszynowego uczenia się (*machine learning* – przystępne wprowadzenie do problematyki w: Alpaydin 2016) i charakteryzujące się pewną autonomią, to znaczy zdolnością działania w środowisku bez stałego nadzoru człowieka (por. Russell, Norvig 2016, s. 38–39). Podkreślenia wymaga, że współcześnie motywacją do dyskusji na temat SI są nie tylko spektakularne osiągnięcia technologii, jak program IBM *Watson*, który uzyskał przewagę nad mistrzami teleturnieju *Jeopardy!* (https://www.ibm.com) czy reprezentujący nieosiągalny dla człowieka

poziom w starożytną grę GO program *AlphaGo Zero* (https://deepmind.com). Sednem obecnego sporu dotyczącego roli SI w społeczeństwie jest wszechobecność programów komputerowych wyposażonych w mechanizmy uczenia się, stosowanych między innymi w reklamie i marketingu, diagnostyce medycznej, przy dokonywaniu oceny pracowników i potencjalnych pracowników, w związku z rozpoznawaniem wniosków kredytowych czy w różnego rodzaju programach przetwarzających język naturalny, na przykład chatbotach. Szczególne kontrowersje są związane z programami komputerowymi stosowanymi w sytuacjach, w których występuje znaczne ryzyko naruszenia przez maszyny takich wartości jak zdrowie i życie człowieka – mamy tu na myśli algorytmy stosowane w autonomicznych pojazdach (Prakken 2017) i robotach wojskowych (Bode, Huelss 2018). Debata dotycząca tych zagadnień ma szeroki zasięg i uczestniczą w niej przedstawiciele różnych grup społecznych. Bardziej ekspercki charakter ma dyskusja dotycząca roli SI w wymiarze dostępu do informacji prawnej i wsparcia wykonywania różnych zadań w związku ze świadczeniem pomocy prawnej i stosowaniem prawa. Niemniej jednak kilka istotnych rezultatów, dotyczących szczególnie oceny jakości kontraktów czy przewidywania pewnych decyzji prawnych, stało się już przedmiotem zainteresowania szerokiej publiczności i podlega komercjalizacji, jak program *Ross*, wyszukujący informacje dotyczące zagadnień prawnych w zbiorach orzeczeń (https://rossintelligence.com), czy program *Case Crunch*, nakierowany na przewidywanie rozstrzygnięć prawnych (https://www.case-crunch.com). Zasięg stosowania technologii SI i jej dynamiczny rozwój stanowią asumpt do dokonywania starannej analizy ryzyka związanego z tym zjawiskiem, jak również do rozważania jego społecznych, moralnych i prawnych implikacji.

Rezultaty generowane przez programy uczące się coraz częściej spełniają oczekiwania użytkowników, ale problematyczne okazuje się wyjaśnienie w konkretnej sytuacji, dlaczego uzyskiwane rezultaty są właśnie takie, a nie inne. Co więcej, niektóre spośród rezultatów oceniane są jako niepokojące, utrwalają bowiem pewne błędy zawarte w zbiorach danych, w tym również różnego rodzaju stereotypy, albo prowadzą do oczywiście niesprawiedliwych, stronniczych (*biased*) rozstrzygnięć (por. ogólne wprowadzenie na stronie internetowej IBM: https://www.research.ibm.com). Istotne w tym zakresie staje się zagadnienie wyjaśnialności albo – inaczej – interpretowalności lub przejrzystości (wyrazy te stanowią odpowiedniki angielskich terminów *explainability*, *interpretability* oraz *transparency*) operacji obliczeniowych wykonywanych przez Sztuczną Inteligencję. Celem niniejszego artykułu jest omówienie zagadnienia wyjaśnialności rezultatów uzyskiwanych przez programy wyposażone w elementy SI w wymiarze stosowania tej technologii do przetwarzania informacji prawnej i dostępu do informacji prawnej.

Należy postulować, aby zagadnienia związane z wpływem algorytmów SI na decyzje dotyczące ważnych dla ludzi dóbr oraz praw jednostek stały się przedmiotem szerokiej debaty społecznej. W tej dyskusji szczególną rolę do odegrania mają organizacje pozarządowe skoncentrowane na problematyce transparencji decyzji podejmowanych zarówno przez władze publiczne, jak i przez przedsiębiorców

w ich relacjach z konsumentami. Programy komputerowe wyposażone w elementy SI są bowiem tworzone przede wszystkim w celu maksymalizacji efektywności procesów. Przedsiębiorstwa informatyczne są naturalnie zainteresowane tworzeniem produktów, które pozwolą ich klientom, zarówno z sektora publicznego, jak i z sektora prywatnego, ograniczyć koszty funkcjonowania oraz dokonać optymalizacji tych parametrów, które są istotne z punktu widzenia osiągania ich celów. Z kolei podmioty posługujące się tym oprogramowaniem czują się zwolnione z obowiązku wyjaśniania takiego czy innego sposobu działania lub podjęcia takiej lub innej decyzji, ponieważ twierdzą, że została ona zasugerowana przez „obiektywny" algorytm. W konsekwencji ostateczny odbiorca danego działania lub decyzji ma znikomy wpływ na przebieg danego procesu albo jego rezultat. Dlatego niezbędne jest, aby organizacje pozarządowe zainteresowane ochroną praw obywateli (szczególnie: konsumentów) podjęły działania mające na celu monitorowanie działania algorytmów i dążenie do implementacji polityk, które zapewnią lepszą wyjaśnialność działania algorytmów. Przedmiotowa debata toczy się już od dłuższego czasu w państwach zachodnich (a także azjatyckich), ze względu zaś na wzrastającą doniosłość społeczną problemów związanych z SI przybiera na intensywności. Jako przykład warto wymienić ogłoszenie przez koalicję The Public Voice, z datą 23 października 2018 roku, *Uniwersalnych wytycznych dla Sztucznej Inteligencji* (https://thepublicvoice.org). Celem tego dokumentu jest zwrócenie uwagi na wzrastającą rolę programów komputerowych wyposażonych w elementy SI we współczesnych społeczeństwach oraz sformułowanie zasad, które powinny być przestrzegane w związku z ich funkcjonowaniem. Wśród dwunastu zaproponowanych zasad na pierwszym miejscu zostało wymienione prawo do przejrzystości (*transparency*), opisane następująco: „Każdy ma prawo poznać podstawę decyzji Sztucznej Inteligencji, która go dotyczy. To prawo obejmuje dostęp do czynników, logiki oraz technik, które wytworzyły dany rezultat" [tłum. własne – M.A.]. Postulowane prawo do przejrzystości (interpretowalności, wyjaśnialności) dotyczy szczególnie tych decyzji podejmowanych przez SI, które mają znaczenie w odniesieniu do praw lub obowiązków osób, na przykład w związku ze stosowaniem algorytmów SI w celu wsparcia procesu sądowego, udzielenia pomocy prawnej lub podjęcia decyzji co do statusu danej osoby, choćby więźnia (por. program Public Safety Assessment stosowany w amerykańskim systemie penitencjarnym http://www.arnoldfoundation.org).

Rozważania poniższe, skierowane do czytelnika niemającego wykształcenia w zakresie nauk technicznych, są prezentowane nieformalnie, a zatem w pewnym zakresie są uproszczone. Ponieważ polska literatura prawnicza dotycząca pojęcia SI oraz jego zastosowań w prawie obejmuje pojedyncze pozycje, poświęcone zresztą zagadnieniom szczegółowym (Petzel 2017, s. 340–365; Cyrul i in. 2014, s. 149–174), a opracowania filozoficzne i informatyczne są zazwyczaj skierowane do specjalistów, w pierwszej kolejności omawiamy podstawowe pojęcia oraz rozróżnienia dotyczące istoty badań nad SI, mając nadzieję, że przyczyni się to do ograniczenia pewnych nieporozumień dotyczących rezultatów

uzyskiwanych w tym obszarze. Przede wszystkim tłumaczymy pojęcie tak zwanej wyjaśnialnej Sztucznej Inteligencji. W drugiej części artykułu rekapitulujemy zwięźle stan badań dotyczących modelowania rozumowań prawniczych. Na zakończenie wskazujemy pewne paradoksalne konsekwencje wynikające z dyskusji nad „prawem do wyjaśnienia".

Wokół pojęcia Sztucznej Inteligencji – wyjaśnialna Sztuczna Inteligencja

Spór dotyczący pojęcia Sztucznej Inteligencji trwa od momentu ukucia tego terminu przez Johna McCarthy'ego w 1956 roku, kiedy odbyła się słynna konferencja w Dartmouth College, przez wiele osób uważana za symboliczny początek badań nad tą problematyką (Kisielewicz 2014, s. 41). Oczywiście nie sposób w tym miejscu choćby wymienić wszystkich aspektów tej dyskusji. Z uwagi na cele niniejszego opracowania wystarczające będzie odróżnienie filozoficznego i praktycznego wymiaru sporu o Sztuczną Inteligencję. W pewnym uproszczeniu wymiar filozoficzny dotyczy kwestii możliwości stworzenia programu komputerowego, który byłby wyposażony w umiejętność myślenia charakterystyczną dla człowieka. Lapidarnie kwestię tę oddaje Andrzej Kisielewicz, który określa zadanie badań nad SI jako stworzenie „autonomicznego, myślącego robota" (Kisielewicz 2014, s. 34). Takie ujęcie problemu prowadzi do kluczowych pytań dotyczących rozumienia pojęcia myślenia, a także świadomości (świadomej jaźni) charakterystycznej dla ludzi. Warto przypomnieć w tym miejscu klasyczne rozróżnienie, dokonane przez Johna R. Searle'a, na silną i słabą SI (Searle 1980). Silną SI byłby program komputerowy myślący tak jak człowiek, słaba SI to taki program komputerowy, który zachowuje się jak gdyby był inteligentny (świadomy, myślący), chociaż „w rzeczywistości" taki nie jest. Zdaniem Johna R. Searle'a i wielu innych uczonych, stworzenie silnej SI jest niemożliwe z podstawowych powodów, inni stoją jednak na stanowisku, że jest to raczej kwestia dalszego postępu technologii i badań neuronaukowych. Należy stwierdzić, że ze względu na złożoność problemów związanych z teorią świadomości rachuby niektórych futurologów, takich jak Ray Kurzweil, zapowiadających powstanie silnej SI w ciągu kilku dziesięcioleci (na przykład Kurzweil 1999), należy uznać za co najmniej wątpliwe. Dlatego w ramach niniejszego opracowania koncentrujemy się na słabym ujęciu SI i na praktycznych zastosowaniach Sztucznej Inteligencji. Można wyróżnić wiele cech, których występowanie (w odpowiednio wysokim stopniu) powoduje, że jesteśmy skłonni uznać dany program komputerowy za inteligentny, na przykład:

- zdolność wyprowadzania zasadnych wniosków z podanych przesłanek,
- umiejętność przetwarzania języka naturalnego, przejawiająca się na przykład w możliwości udzielania adekwatnych odpowiedzi na pytania czy podtrzymywania konwersacji,
- zdolność zachowania się zgodnie z narzuconymi normami i właściwe reagowanie, gdy norma zostaje naruszona,

- zdolność uczenia się, to znaczy na przykład zdolność zmiany sposobu funkcjonowania w celu optymalizacji wykonywania danego zadania lub dostosowania się do zmiennych warunków środowiskowych.

Trzy pierwsze wymienione cechy mają charakter przedmiotowy, dotyczą bowiem rozwiązywania zadań określonego rodzaju, z kolei zdolność uczenia się ma charakter metacechy w tym sensie, że może prowadzić do optymalizacji poziomu wykonania zadań przedmiotowych. W historii badań nad SI ukształtowały się dwa główne podejścia związane z kształtowaniem systemów ukierunkowanych na rozwiązywanie zadań wymagających inteligencji: podejście klasyczne (symboliczne, określane także jako GOFAI – *good old-fashioned artificial intelligence*) i podejście subsymboliczne (por. Kisielewicz 2014, s. 59–61). Jakkolwiek współcześnie aktualność tego podziału jest z wielu względów kwestionowana, stanowi on przydatne narzędzie dydaktyczne, pozwalające na uporządkowanie rozważań dotyczących Sztucznej Inteligencji. Paradygmatycznym przykładem realizacji podejścia symbolicznego są systemy ekspertowe: programy komputerowe symulujące rozumowania ekspertów w określonej, wąskiej dziedzinie. Istota działania systemów ekspertowych może być przedstawiona następująco. Użytkownik programu wprowadza pewne zdania o faktach, udzielając odpowiedzi na zadawane przez system pytania. Następnie silnik inferencyjny programu, korzystając z bazy wiedzy (zwyczaj wyrażonej jako system reguł – warunkowych wyrażeń typu „jeżeli – to") oraz informacji wprowadzonych przez użytkownika, przeprowadza wnioskowanie prowadzące do określonej konkluzji. Słynnym wczesnym przykładem systemu ekspertowego jest *MYCIN* – program komputerowy, którego celem było przedstawienie diagnozy i zaproponowanie odpowiedniej terapii na podstawie wprowadzonego przez użytkownika opisu symptomów choroby (Buchanan, Shortliffe 1984). Systemy ekspertowe często wyposażano w tak zwany moduł wyjaśniający – element pokazujący użytkownikowi, w jaki sposób program wyprowadził daną konkluzję. Rozwój badań na systemami ekspertowymi ukazał wiele istotnych ograniczeń tego podejścia, w tym szczególnie niewrażliwość na specyficzne, kontekstowe cechy danej sprawy, możność przetwarzania tylko określonych typów wiedzy (najczęściej: nadającej się do wyrażenia w formie reguł) czy brak możliwości łatwej inkorporacji nowej wiedzy przez system. Główną zaletą systemów ekspertowych jest możność przypisania znaczenia wszystkim elementom, na podstawie których przeprowadzają one rozumowania (elementy te bowiem mają charakter symboliczny) i prześledzenia toku rozumowania programu. Należy dodać, że obecnie intensywnie jest rozwijany kierunek badań związany z uwzględnianiem roli kontekstu w modelach rozumowania regułowego (na przykład Bobek, Nalepa, Ślażyński 2018).

Równolegle rozwijało się inne podejście, związane z wykonywaniem przez program operacji subsymbolicznych w tak zwanych sztucznych sieciach neuronowych (SSN). Mimo że aktualny stan badań nad takimi sieciami jest bardzo zaawansowany, zasadę działania tych obiektów dogodnie jest przedstawić na przykładzie modelu stworzonego jeszcze w latach czterdziestych XX wieku

(McCulloch, Pitts 1943). Sztuczny neuron jest zdefiniowany jako obiekt mający *n* wejść z przypisanymi wagami oraz jedno wyjście. Sygnał jest podawany do wejść neuronu, następnie zaś sumowany i przekształcany przez pewną funkcję. Jeżeli wartość otrzymana w neuronie przekroczy pewną wartość progową, to neuron zostaje aktywowany i produkuje sygnał wyjściowy – w przeciwnym wypadku nie zostaje aktywowany. Sztuczne sieci neuronowe konstruuje się przez łączenie sztucznych neuronów (o rodzajach sztucznych sieci neuronowych – zob. Osowski 2013). Utworzone w ten sposób struktury znajdują zastosowanie w rozwiązywaniu zadań, których kryteria poprawności trudno precyzyjnie opisać za pomocą zdań języka, na przykład rozpoznawanie twarzy wymaga analizy położenia znacznej liczby charakterystycznych punktów. Sieci neuronowe podlegają procesowi treningu (uczenia się). Niewytrenowana sieć neuronowa popełnia zazwyczaj liczne błędy (kontynuując przykład: nie rozpoznaje twarzy danej osoby A albo mylnie rozpoznaje twarz osoby B jako twarz osoby A). W procesie treningu dokonuje się zmian wag na wejściach do neuronów, aby maksymalizować trafność udzielanych odpowiedzi. Istotne jest, że zmianom wartości liczbowych przyjmowanych przez wagi nie sposób przypisać jakiegokolwiek znaczenia – w tym sensie są to operacje o charakterze subsymbolicznym.

Maszynowe uczenie się może być realizowane z wykorzystanie różnych podejść i narzędzi (Mohri, Rostamizadeh, Talwalkar 2012), ale stosowanie sztucznych sieci neuronowych w takich systemach jest powszechne i niejednokrotnie prowadzi do efektywnych wyników. Sztuczne sieci neuronowe są określane jako uniwersalne aproksymatory funkcji wielu zmiennych (Osowski 2013, s. 15). Po poddaniu odpowiedniemu treningowi (który współcześnie często odbywa się także bez nadzoru człowieka) systemy uczące się wykorzystujące sztuczne sieci neuronowe osiągają doniosłe rezultaty, na przykład w zadaniach związanych z identyfikacją i klasyfikacją wzorów oraz obiektów czy przewidywania zdarzeń.

Rezultaty osiągane przez systemy korzystające z mechanizmów maszynowego uczenia się spowodowały ich rozpowszechnienie w różnych sferach działalności gospodarczej i życia publicznego. Szybko dostrzeżono jednak, że działają one w znacznym zakresie na zasadzie czarnej skrzynki – dostarczają wyników odpowiadających oczekiwaniom, ale podanie wyjaśnienia, dlaczego wyniki te są takie, a nie inne, prowadzi do istotnych trudności. Co więcej, w wielu wypadkach (wszędzie tam, gdzie od wyniku działania algorytmu zależą decyzje dotyczące ważnych dla człowieka wartości) społeczne oczekiwania dotyczą nie tylko matematycznych formuł, na podstawie których program wygenerował wynik, ale – docelowo – przekonującej, wyrażonej w języku naturalnym argumentacji na rzecz takiego, a nie innego rezultatu. Problem w tym, że mechanizmy *machine learning* nie są oparte na jakościowej argumentacji, lecz na ilościowej analizie danych. Wyraźna jest przy tym zależność, że osiąganie przez program wyników optymalizujących założone kryteria wymaga posłużenia się bardziej skomplikowanymi, a zatem trudniej zrozumiałymi dla człowieka rozwiązaniami obliczeniowymi.

W związku z tymi zagadnieniami zostało ukute pojęcie „wyjaśnialnej", „przejrzystej" lub „interpretowalnej" Sztucznej Inteligencji (*explainable AI*, XAI; *transparent AI*; *interpretable AI*), które obecnie jest przedmiotem ożywionej debaty (Molnar 2018; Miller 2017; Doshi-Velez, Kim 2017; Lundberg, Lee 2017).

Dyskusja ta dotyczy szczególnie rozumienia terminów „wyjaśnialność" czy „interpretowalność" w omawianym kontekście. Przedstawiciele nauk komputerowych jeszcze nie korzystają w dostatecznym stopniu z bogatego dorobku metodologii nauk w zakresie pojęcia wyjaśnienia, chociaż pojawiają się już w tym obszarze interesujące opracowania, nawiązujące do wyjaśnienia w naukach społecznych (Miller 2017). Nie jest również wyraźnie podkreślana różnica między wyjaśnieniem kauzalnym takiego, a nie innego zachowania algorytmu, i rzeczywistym lub możliwym normatywnym uzasadnieniem decyzji, która powinna być podjęta w danej sytuacji, jakiej dotyczy działanie algorytmu. Można się jednak spodziewać, że te deficyty metodologiczne zostaną zniwelowane w nieodległej przyszłości, ponieważ dyskusja nad XAI przybiera coraz bardziej interdyscyplinarny charakter.

Należy zatem wyróżnić kilka płaszczyzn rozważań dotyczących pojęcia wyjaśnialności (interpretowalności) Sztucznej Inteligencji:

- płaszczyznę informatyczną (algorytmiczną), dotyczącą zasad konstruowania algorytmów i modeli podlegających wyjaśnieniu lub mających na celu wyjaśnienie działania innych algorytmów albo modeli,
- płaszczyznę społeczną, w ramach której dyskutuje się skutki społeczne funkcjonowania w obrocie inteligentnych programów komputerowych, szczególnie wpływ tych programów na zachowanie użytkowników czy podejmowanie decyzji przez organy władz,
- płaszczyznę etyczną, dotyczącą szczególnie takich zagadnień jak zaufanie użytkowników do rezultatów generowanych przez systemy uczące się, wpływ tych systemów na dystrybucję dóbr czy często dyskutowana kwestia stronniczości (*bias*) tworzonej lub utrwalanej przez mechanizmy Sztucznej Inteligencji,
- płaszczyznę prawną, związaną między innymi z problematyką odpowiedzialności za działania programów uczących się, ochroną danych i prywatności użytkowników, przy czym płaszczyzn tych nie należy uznawać za ściśle rozłączne. Zwłaszcza płaszczyzny etyczna i prawna interesują się podobnymi aspektami funkcjonowania SI, ale koncentrują się na odmiennych ich implikacjach: płaszczyzna informatyczna wyznacza ramy brzegowe dla dyskusji toczącej się na wszystkich płaszczyznach, o ile tylko dyskusja ta jest prowadzona w związku z faktami, a nie ogranicza się tylko do sfery postulatów.

W zakresie płaszczyzny prawnej szczególnie doniosłe jest zagadnienie konstruowania prawa podmiotowego do wyjaśnienia działania algorytmu, obejmujące wiele zagadnień cząstkowych, takich jak: identyfikacja podstawy prawnej takiego uprawnienia, podmiotów uprawnionych i zobowiązanych oraz określenie treści takiego prawa, zwłaszcza związanych z nimi roszczeń i skutków ich wniesienia. Przytoczone powyżej sformułowanie prawa do przejrzystości,

sformułowane przez The Public Voice, ma obecnie charakter postulatu, chociaż pewnych jego elementów można się doszukiwać także w treści obowiązujących regulacji.

Warto zwrócić uwagę na niedawny głos w tej dyskusji, wyrażony w związku z wprowadzeniem przepisów ogólnego rozporządzenia o ochronie danych (RODO). Autorzy identyfikują liczne obecne ograniczenia dotyczące możliwości wyjaśnienia funkcjonowania uczących się algorytmów przetwarzających dane i wskazują na iluzoryczność postulowanego prawa do wyjaśnienia: użytkownik danego systemu zazwyczaj nie ma możliwości kwestionowania lub podważania wyjaśnienia decyzji podanej przez system, nawet jeżeli okaże się ono niezrozumiałe czy nieprzekonujące (Edwards, Veale 2017, s. 67). Wskazują ponadto, że ochrony praw użytkowników należy raczej upatrywać w innych instrumentach niż prawo podmiotowe do wyjaśnienia, przywołując także mechanizmy o szerszym zasięgu, takie jak ocena wpływu czy proces certyfikacji (*ibidem*, s. 79–80). Podkreślenia wymaga, że negatywne wnioski autorów dotyczą przede wszystkim kwestii stosowania RODO, a nie prowadzą do podważenia zasadności szerzej rozumianych wysiłków związanych z konstruowaniem prawa do wyjaśnienia działania algorytmów Sztucznej Inteligencji. Niemniej jednak uzyskane przez nich negatywne wyniki ukazują skalę komplikacji problemów związanych z konstruowaniem prawa do wyjaśnienia.

Sztuczna Inteligencja jako wsparcie rozwiązywania problemów prawnych

Nurt badawczy zwany „SI i prawo" rozwija się intensywnie od lat osiemdziesiątych XX wieku, chociaż pionierskie prace z tego zakresu były tworzone już wcześniej (omówienie istotnych w ujęciu chronologicznym – por. Bench-Capon i in. 2012). Celem tych badań jest nie tylko optymalizacja systemów wyszukiwania informacji prawnej, ale także modelowanie rozumowań prawniczych – z intencją budowania programów komputerowych potrafiących tworzyć argumenty za określonym stanowiskiem i przeciw określonemu stanowisku, proponować różne hipotezy interpretacyjne czy formułować odpowiedzi na pytania prawne dotyczące podanych stanów faktycznych. Badania z zakresu „SI i prawo" są doniosłe nie tylko dla branży komercyjnych usług prawnych, lecz również dla wymiaru sprawiedliwości czy dla organizacji pozarządowych zajmujących się świadczeniem pomocy prawnej i udzielaniem informacji prawnej. Współczesny obywatel czerpie wiedzę o prawie przede wszystkim ze źródeł elektronicznych, ale są one dla niego albo trudno zrozumiałe (teksty aktów normatywnych), albo nadmiernie rozproszone, nieaktualizowane i często wprowadzające w błąd (różnorodne opracowania dostępne na forach internetowych czy w serwisach społecznościowych). Profesjonalny prawnik korzystający z komercyjnych baz informacji prawnej również napotyka różne trudności, szczególnie wynikające z ogromnej obszerności dostępnych zbiorów dokumentów i sposobu funkcjonowania mechanizmów wyszukiwania w tych zbiorach danych. Wizja skonstruowania

programu komputerowego, który inteligentnie wyszuka elementy istotne dla analizy danego stanu faktycznego, a następnie sformułuje propozycję rozumowania nakierowanego na rozwiązanie danego problemu prawnego, zainspirowała wiele podejść, które były rozwijane przez kolejne dekady badań.

Podobnie jak w ramach ogólnych badań nad SI, wczesne istotne opracowania miały charakter systemów ekspertowych, bazujących na reprezentacji wiedzy prawniczej w formie reguł. Następnie skoncentrowano się na modelowaniu elementów rozumowań prawniczych związanych z kazusami (precedensami – w tym miejscu należy wymienić klasyczny system HYPO Kevina Ashleya), aby pod koniec lat osiemdziesiątych XX wieku połączyć oba typy modelowania w tak zwane systemy hybrydowe – poszczególne stadia rozwoju tego paradygmatu omawia Trevor Bench-Capon (2017). Ze współczesnego punktu widzenia należy stwierdzić, że zaletą prawniczych systemów ekspertowych była ich całkowita wyjaśnialność: zwłaszcza system HYPO został zaprojektowany tak, aby prezentować użytkownikowi argumenty za poszczególnymi rozstrzygnięciami i przeciw poszczególnym rozstrzygnięciom, z wyraźnym uzasadnieniem. Do podstawowych wad tych systemów należy zaliczyć możliwość oparcia się wyłącznie na informacjach wpisanych do bazy wiedzy programu, brak możliwości przetwarzania informacji wyrażonych w języku naturalnym czy konieczność rozwiązywania przez użytkownika (a nie przez program) problemów związanych z interpretacją pojęć nieostrych oraz rozumowaniami dotyczącymi wartości w prawie. Nie oznacza to, że prawnicze systemy ekspertowe należy uznać za praktycznie bezużyteczne, ale trzeba stwierdzić, że ich zakres efektywnego stosowania jest dość wąski (Araszkiewicz, Łopatkiewicz, Żurek 2017), chociaż wystarczający dla niektórych celów, jak generowanie prostych pism procesowych w ramach systemu nieodpłatnej pomocy prawnej w Stanach Zjednoczonych (program *A2J Author* – por. https://www.a2jauthor.org). Badania nad prawniczymi systemami ekspertowymi były także jednym z czynników prowadzących do uzmysłowienia sobie przez uczonych roli wartości w rozumowaniach prawniczych.

W latach dziewięćdziesiątych XX wieku i później podjęto zaawansowane próby tworzenia modeli rozumowań prawniczych, które reprezentowałyby proces argumentacji oraz rolę wartości i ważenia różnych racji w procesie stosowania prawa (między innymi: Hage 1997; Araszkiewicz 2013; Żurek, Araszkiewicz 2013; Ashley 2017, s. 127–168). Modele te, jakkolwiek oferujące szersze możliwości reprezentowania wiedzy prawniczej niż klasyczne systemy ekspertowe, obarczone są istotnymi ograniczeniami charakterystycznymi dla tych ostatnich. Co więcej, z uwagi na złożoność rozumowań dotyczących wartości praktyczna implementacja tych modeli jawi się jako bardzo czasochłonne i skomplikowane zadanie. Modele rozumowań związanych z wartościami pozostają więc projektami o charakterze naukowym. Podkreślenia wymaga, że rezultaty generowane przez te modele są – podobnie jak w systemach ekspertowych – w pełni wyjaśnialne. Co więcej, unaoczniają one, w jakim zakresie decyzje podejmowane przez organy stosujące prawo mają charakter dyskrecjonalny.

W latach dziewięćdziesiątych XX wieku w nurcie „SI i prawo" wiele uwagi poświęcano kwestii, która od dawna była postrzegana jako bardzo problematyczna dla teorii systemów ekspertowych: modelowaniu wiedzy o świecie oraz rozumowań zdroworozsądkowych. Kolejne systemy wyposażone w odpowiednią wiedzę na temat regulacji okazywały się mało użyteczne w praktyce, ponieważ nie potrafiły korzystać z wiedzy zdroworozsądkowej, której rola w rozumowaniach prawniczych nie była do tej pory należycie doceniana. Prowadzone badania unaoczniły, że trudno jest wyznaczyć precyzyjną linię między wiedzą o regulacji i wiedzą o otaczającym nas świecie (Breuker, den Haan 1991), a zatem nie do utrzymania jest sztywny podział zagadnień dotyczących sfery prawa oraz sfery faktu. Rozstrzygnięcia *in concreto*, czy dana kwestia dotyczy sfery regulacji prawnej, czy raczej sfery naszej wiedzy o świecie, z trudem poddają się uogólnieniom.

Ze względu na ograniczenia związane ze stosowaniem narzędzi klasycznej SI oraz ogromny rozkwit technologii maszynowego uczenia się w XXI wieku nie jest zaskakujące, że obecnie intensywnie eksploruje się możliwości zastosowania tych narzędzi do budowania programów wspierających rozumowania prawników, ze szczególnym uwzględnieniem automatyzacji przetwarzania języka naturalnego i statystycznej analizy dużych zbiorów danych. Prowadzone są bardzo intensywne badania między innymi nad zagadnieniami automatycznej analizy i klasyfikacji elementów tekstów wyrażonych w języku naturalnym (aktów normatywnych i orzeczeń sądowych), w tym elementów argumentów prawnych, istotnych okoliczności faktycznych wpływających na rozstrzygnięcie, wyróżnionych funkcjonalnych elementów tych dokumentów czy wreszcie treści reguł prawnych znajdujących zastosowanie w danym przypadku (w zasadzie całościowe, ale zwięzłe omówienie aktualnego stanu badań: Ashley 2017, s. 234–310). Postęp, jaki został dokonany w ciągu ostatnich dwóch dekad w zakresie automatycznej analizy źródeł informacji prawnej, jest znaczny, a dzięki komponentom maszynowego uczenia się programy analizujące te dane uzyskują dokładne rezultaty. Należy jednak pamiętać, że programy do analizy zbiorów danych prawnych, podobnie jak wszelkie programy analizujące zbiory danych, opierają swoje rezultaty na statystycznych korelacjach występowania wyrażeń różnego typu. Dlatego też za palący należy uznać problem wyjaśnialności tych rezultatów. Trzeba przy tym postawić tezę, że w odniesieniu do rezultatów generowanych przez programy analizujące dane prawne uzasadnienie rezultatu, rozumiane jako podanie przekonującej, racjonalnej argumentacji, jest co najmniej tak samo istotne, o ile nie ważniejsze, od samego rezultatu. W związku z tym za najważniejsze obecnie wyzwanie dotyczące badań w nurcie „SI i prawo" powinno się uznać integrację mechanizmów i rezultatów uzyskiwanych przez systemy uczące się analizujące język naturalny z modelami rozumowań prawniczych reprezentujących proces argumentacji i interpretacji prawniczej. Fakt prowadzenia badań nad systemami dokonującymi zautomatyzowanej ekstrakcji i konstrukcji argumentów z dużych zbiorów danych (jak program IBM *Debater* – https://www.research.ibm.com) nie

prowadzi do rozwiązania tego problemu, gdyż zasada działania tych systemów jest taka sama jak innych uczących się programów analizujących język naturalny, jest zatem oparta w znacznym zakresie na statystycznej analizie współwystępowania pewnych fraz. Innymi słowy, programy analizujące duże zbiory danych prawnych przynoszą wartościową informację na temat struktury tych danych, ale nie dostarczają uzasadnienia za podjęciem takiej, a nie innej decyzji. Propozycję takiego uzasadnienia mogą skonstruować systemy oparte na wiedzy wyrażonej symbolicznie, lecz te ostatnie nie potrafią analizować objętościowo rozbudowanych zbiorów danych wyrażonych w języku naturalnym. Należy dodać, że prawniczych systemów uczących się dotyczy, tak samo jak ogólnie algorytmów uczących się, ryzyko stronniczości (*bias*), którego przyczyną mogą być zarówno anomalia struktury analizowanych danych, jak i ukształtowanie samego algorytmu. Ponownie jako bardzo istotna jawi się w tym miejscu kwestia wyjaśnialności SI, w tym wymiarze – SI stosowanej do analizy źródeł prawnych.

***

Systemy SI w znacznym stopniu wpływają na życie współczesnego obywatela społeczeństwa rozwiniętego, na przykład sugerując mu optymalną trasę do wybranego punktu czy rekomendując dokonanie rezerwacji w określonym hotelu, lecz również oceniając jego zdolność kredytową czy jego wyniki jako pracownika. Z uwagi na ryzyka związane z funkcjonowaniem systemów uczących się (szczególnie ryzyko podejmowania niesprawiedliwych decyzji), a także brak możliwości łatwego zrozumienia funkcjonowania algorytmów i przez ostatecznych odbiorców ich decyzji, i przez ich twórców, za istotne zostało uznane zagadnienie wyjaśnialności funkcjonowania Sztucznej Inteligencji. W niniejszym artykule przedstawiono wybrane podstawowe problemy związane z wyjaśnialnością SI, które następnie odniesiono do dostępu do informacji i pomocy prawnej.

Współcześnie nie sposób wyobrazić sobie efektywnego wykonywania zawodów prawniczych i pozyskiwania wiedzy o prawie bez korzystania z elektronicznych baz danych. Z uwagi na ogromną objętość korpusów dokumentów prawnych oraz złożoność współczesnych systemów prawnych nie tylko celowe, ale także konieczne wydaje się korzystanie z bardziej zaawansowanych narzędzi informatycznych wyposażonych w komponenty maszynowego uczenia się. W tym zakresie palące staje się zagadnienie wyjaśnialności działania takich algorytmów i uzasadnienia decyzji prawnych podejmowanych przy ich wsparciu. Obecnie bowiem otwarty jest problem, w jaki sposób połączyć rezultaty uzyskiwane przez systemy uczące się, działające na zbiorach tekstów prawnych, na przykład w zakresie klasyfikacji pewnych elementów tych dokumentów lub predykcji decyzji, z systemami korzystającymi z klasycznych rozwiązań SI (GOFAI), pomocnych w konstruowaniu racjonalnej argumentacji na rzecz określonych tez, ale wymagających wprowadzenia odpowiednio ustrukturyzowanej wiedzy i pozbawionych możliwości uczenia się na podstawie danych wyrażonych w języku naturalnym.

Problemy wyjaśnialności działania algorytmów uczących się w wymiarze prawnym powinny być zestawione z niewątpliwymi korzyściami wynikającymi ze stosowania takich systemów. Nie ulega wątpliwości, że wyłącznie algorytmy SI są w stanie w rozsądnym czasie zbadać dziesiątki tysięcy dokumentów, mając na uwadze wyszukanie w nich istotnej informacji (na przykład przydatnej do konstruowania argumentacji). Uwzględniając kłopoty związane ze stronniczością systemów uczących się, należy pamiętać, że człowiek – ekspert prawny również jest niewolny od błędów poznawczych oraz stronniczości. Można zatem postawić tezę, że rozwój SI będzie prowadził do pogłębienia badań porównawczych dotyczących stronniczości rezultatów uzyskiwanych przez ludzi i przez maszyny. Prowadzenie badań nad stronniczością algorytmów nie powinno bowiem usuwać z pola widzenia błędów poznawczych popełnianych przez prawników. Niewątpliwie algorytm SI jest w stanie uniknąć wielu błędów popełnianych przez prawników, a wynikających z takich ludzkich cech, jak podatność na zmęczenie i stres, kierowanie się (choćby nieświadomie) własnymi preferencjami i dopasowywanie do nich pozornie obiektywnej argumentacji czy korzystanie z tak zwanych heurystyk, na przykład sięganie po informację, która jest najłatwiej dostępna (klasyczne opracowanie problematyki heurystyk i błędów poznawczych – por. Kahneman, Slovic, Tversky 1982).

Celowe wydaje się przeprowadzenie następującego eksperymentu myślowego. Wskazywaliśmy powyżej, że rekonstrukcja treści i zakresu „prawa do wyjaśnienia" działania algorytmów prowadzi do trudności. Są to jednak rezultaty wypracowywane przez ludzi. W perspektywie kilku dekad można się spodziewać skonstruowania programów, które będą w stanie, na podstawie różnych źródeł, dokonywać prób „samodzielnej" rekonstrukcji treści i zakresu takiego prawa. Można oczywiście twierdzić, że dokonana przez program rekonstrukcja byłaby z definicji obarczona błędem stronniczości. Program komputerowy może jednak przedstawić argument na rzecz tezy przeciwnej, wskazując również, że to ocena dokonywana przez człowieka jest obarczona co najmniej równie istotnym błędem stronniczości. Zasadne stanie się wówczas pytanie o to, kto powinien rozstrzygać spory tego typu i na podstawie jakiej procedury.

PRZYPISY

[1] Niniejszy tekst powstał w związku z realizacją projektu badawczego K/DSC/004874 „Metody modelowania rozumowań prawniczych z pogranicza sfery prawa i sfery faktu".

BIBLIOGRAFIA

Alpaydin, Ethem. 2016. *Machine Learning. The New AI*, Cambridge, Massachusetts: The MIT Press.

Araszkiewicz, Michał. 2013. *Limits of Constraint Satisfaction Theory of Coherence as a Theory of (Legal) Justification*, [w:] Michał Araszkiewicz, Jaromír Šavelka (eds.), *Coherence. Insights from Philosophy, Jurisprudence and Artificial Intelligence*, Dordrecht: Springer, s. 217–242.

Araszkiewicz, Michał, Łopatkiewicz, Agata, Żurek, Tomasz. 2017. *The Tradition of Legal Expert Systems – Possibilities, Limitations and the Way Forward*, [w:] Erich Schweighofer *et al*. (eds.), *Trends und Communities der Rechtsinformatik. Tagungsband des 20. Internationalen Rechtsinformatik Symposions IRIS 2017*, Wien: Österreichische Computer Gesellschaft, s. 123–130.

Ashley, Kevin. 2017. *Artificial Intelligence and Legal Analytics. New Tools for the Law Practice in the Digital Age*, Cambridge: Cambridge University Press.

Bench-Capon, Trevor *et al*. 2012. A history of AI and Law in 50 papers: 25 years of the international conference on AI and Law. *Artificial Intelligence and Law*, 20: 215–319.

Bench-Capon, Trevor. 2017. HYPO's legacy: introduction to the virtual special issue. *Artificial Intelligence and Law*, 25: 205–250.

Breuker, Joost, den Haan, Nienke. 1991. *Separating world and regulation knowledge: where is the logic?*, [w:] Richard E. Susskind (ed.), *Proceedings of the Third International Conference on Artificial Intelligence and Law, ICAIL '91, Oxford, England, June 25–28, 1991*, New York: ACM, s. 92–97.

Bode, Ingvild, Huelss, Hendrik. 2018. Autonomous weapons systems and changing norms in international relations. *Review of International Studies*, 44: 393–413.

Buchanan, Bruce, Shortliffe, Edward. 1984. *Rule-Based Expert Systems. The MYCIN Experiments of the Stanford Heuristic Programming Project*, Reading, Massachusetts: Addison-Wesley.

Cyrul, Wojciech, Duda, Jerzy, Opiła, Janusz, Pełech-Pilichowski, Tomasz. 2014. *Informatyzacja tekstu prawa. Perspektywy zastosowania języków znacznikowych*, Warszawa: Wolters Kluwer.

Doshi-Velez, Finale, Kim, Been. 2017. *Towards A Rigorous Science of Interpretable Machine Learning* – CORR abs/1702.08608.

Edwards, Lilian, Veale, Michael. 2017. Slave to the Algorithm? Why a 'Right to an Explanation' is Probably not the Remedy You Were Looking For. *Duke Law and Technology Review*, 16: 18–84.

Hage, Jaap C. 1997. *Reasoning with Rules. An Essay on Legal Reasoning and its Underlying Logic*. Dordrecht: Springer Science+Business Media.

Kahneman, Daniel, Slovic, Paul, Tversky, Amos. 1982. *Judgment under uncertainty: heuristics and biases*, New York: Cambridge University Press.

Kisielewicz, Andrzej. 2014. *Sztuczna inteligencja i logika. Podsumowanie przedsięwzięcia naukowego*, Warszawa: Wydawnictwo WNT [wydanie II, zmienione].

Kurzweil, Ray. 1999. *The Age of Spiritual Machines,* London: Penguin Books.

Lundberg, Scott, Lee, Su-Yin. 2017. *A Unified Approach to Interpreting Model Predictions*, [w:] Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S.V.N. Vishwanathan, Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4–9 December 2017*, s. 4768–4777.

McCulloch, Warren, Pitts, Walter. 1943. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5: 115–133.

Miller, Tim. 2017. *Explanation in Artificial Intelligence: Insights from the Social Sciences* – CORR abs/1706.07269.

Mohri, Mehryar, Rostamizadeh, Afshin, Talwalkar, Ameet. 2012. *Foundations of Machine Learning,* Cambridge, Massachusetts: MIT Press.

Nilsson, Nils J. 2010. *The Quest for Artificial Intelligence. A History of Ideas and Achievements,* Cambridge: Cambridge University Press.

Osowski, Stanisław. 2013. *Sieci neuronowe do przetwarzania informacji*, Warszawa: Oficyna Wydawnicza Politechniki Warszawskiej.

Petzel, Jacek. 2017. *Systemy wyszukiwania informacji prawnej*, Warszawa: Wolters Kluwer Polska.

Prakken, Henry. 2017. On the problem of making autonomous vehicles conform to traffic law. *Artificial Intelligence and Law*, 25: 341–363.

Russell, Stuart J., Norvig, Peter. 2016. *Artificial Intelligence. A Modern Appoach. Third edition*. Harlow: Pearson Education Limited.

Searle, John R. 1980. Minds, Brains and Programs. *Behavioral and Brain Sciences*, 3: 415–457.

Żurek, Tomasz, Araszkiewicz, Michał. 2013. *Modeling teleological interpretation*, ICAIL, s. 160–168.

AKTY PRAWNE I DOKUMENTY

Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2016/679 z dnia 27 kwietnia 2016 roku w sprawie ochrony osób fizycznych w związku z przetwarzaniem danych osobowych i w sprawie swobodnego przepływu takich danych oraz uchylenia dyrektywy 95/46/WE (ogólne rozporządzenie o ochronie danych) z dnia 27 kwietnia 2016 roku (Dz Urz. UE. L nr 119, s. 1).

ŹRÓDŁA INTERNETOWE

Bobek, Szymon, Nalepa, Grzegorz J., Ślażyński, Mateusz. 2018. *HEARTDROID – Rule engine for mobile and context-aware expert systems. Expert Systems* – https://doi.org/10.1111/exsy.12328 [dostęp: 31 października 2018 roku].

http://www.arnoldfoundation.org/public-safety-assessment-risk-tool-promotes-safety-equity-justice
https://www.ibm.com/watson
https://rossintelligence.com
https://deepmind.com/blog/alphago-zero-learning-scratch
https://www.case-crunch.com
https://www.research.ibm.com/5-in-5/ai-and-bias
https://www.research.ibm.com/artificial-intelligence/project-debater
https://thepublicvoice.org/ai-universal-guidelines