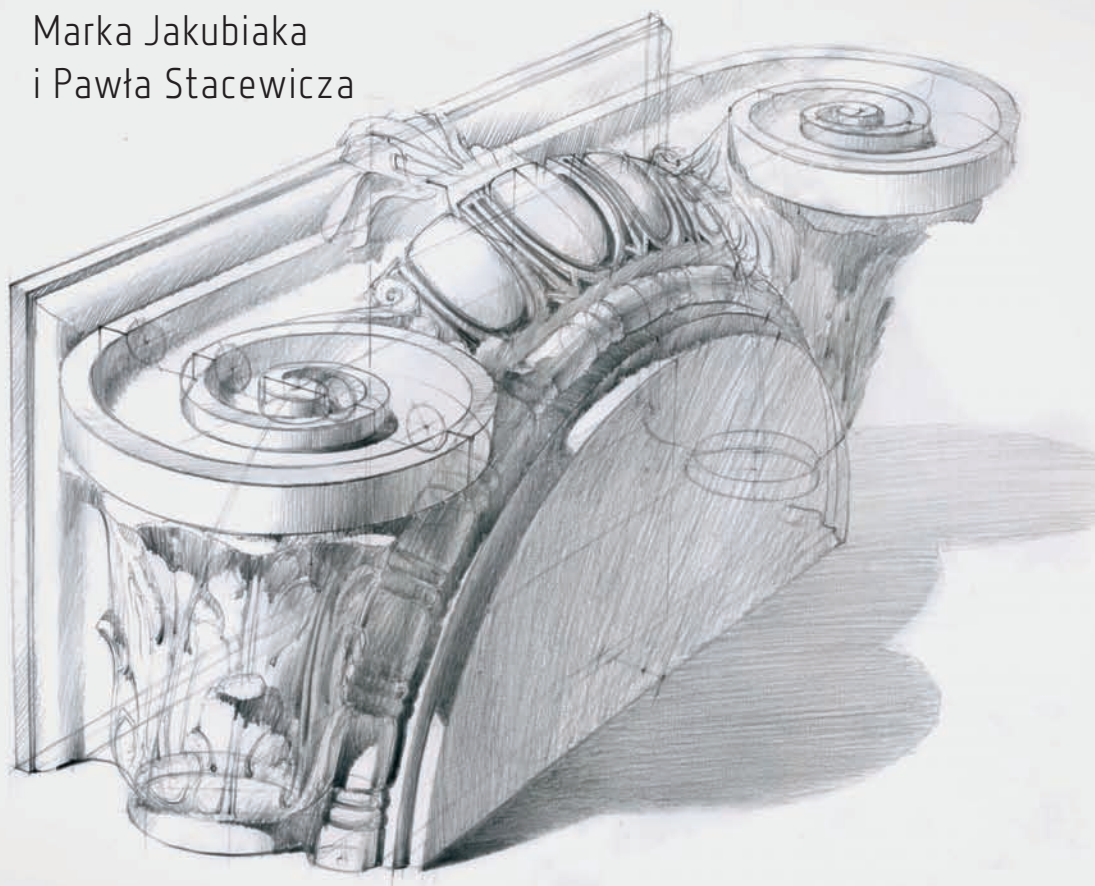


INFORMATYKA A FILOZOFIA

Zaufanie do systemów sztucznej inteligencji

Praca pod redakcją
Marka Jakubiaka
i Pawła Stacewicza



CALCU LEMUS

OFICyna WYDAWNICZA POLITECHNIKI WARSZAWSKIEJ

INFORMATYKA A FILOZOFIA

Zaufanie do systemów sztucznej inteligencji

Pamięci naszej Koleżanki
Heleny Bulińskiej-Stangreckiej



INFORMATYKA A FILOZOFIA

Zaufanie do systemów sztucznej inteligencji

Praca pod redakcją
Marka Jakubiaka
i Pawła Stacewicza



CALCU LEMUS

OFICyna WYDAWNICZA POLITECHNIKI WARSZAWSKIEJ
WARSZAWA 2023

Zaufanie do systemów sztucznej inteligencji

Wydanie I

Słowa kluczowe: sztuczna inteligencja, zaufanie, wyjaśnianie, maszynowe uczenie się, etyka maszyn, ekonomia behawioralna, racjonalność decyzji, szkolnictwo, filtrowanie treści, posthumanizm, technologie informacyjno-komunikacyjne

Key words: artificial intelligence, trust, explaining, machine learning, machine ethics, behavioral economics, rational decision making, education, content moderation, posthumanism, information and communication technologies

Recenzenci

dr hab. Agata Kosieradzka-Federczyk, prof. uczelni – Uniwersytet Kardynała Stefana Wyszyńskiego
dr hab. Jacek Janowski, prof. uczelni – Politechnika Warszawska

Projekt okładki

Danuta Czudek-Puchalska

Skład komputerowy

Beata Zalewska-Kraśniewska

© Copyright by Oficyna Wydawnicza Politechniki Warszawskiej, Warszawa 2023

Wydawca: Politechnika Warszawska
Oficyna Wydawnicza Politechniki Warszawskiej (UIW 48800)
ul. Polna 50, 00-644 Warszawa, tel. 22 234-70-83

Księgarnia internetowa Oficyny Wydawniczej PW www.wydawnictwopw.pl
tel. 22 234-75-03; e-mail: oficyna@pw.edu.pl

Utwór w całości ani we fragmentach nie może być powielany ani rozpowszechniany za pomocą urządzeń elektronicznych, mechanicznych, kopiujących, nagrywających i innych, w tym nie może być umieszczany ani rozpowszechniany w Internecie bez pisemnej zgody posiadacza praw autorskich

ISBN 978-83-8156-511-0 (druk)

ISBN 978-83-8156-517-2 (online)

Zamówienie nr 414/2022

Druk i oprawa: Drukarnia Oficyny Wydawniczej Politechniki Warszawskiej, tel. 22 234-70-30

Spis treści

Od redaktorów tomu	7
Rozdział 1. Sztuczna inteligencja w odbiorze społecznym. Metafory a rzeczywistość (<i>Jarosław Arabas, Jarosław Chudziak</i>)	9
Rozdział 2. Wyjaśnianie, zaufanie i test Turinga (<i>Paweł Stacewicz</i>).....	23
Rozdział 3. Ku zaufanej sztucznej inteligencji. Perspektywa fronetyczna (<i>Paweł Polak, Roman Krzanowski</i>).....	35
Rozdział 4. Behawioralne uwarunkowania decyzji podejmowanych przy wykorzystaniu sztucznej inteligencji (<i>Agnieszka Tomczak</i>).....	46
Rozdział 5. Analiza konsekwencji projektu <i>Moral Machine</i> jako realizacji koncepcji „ko- herentnej, ekstrapolowanej woli ludzkości” dla budowania sklasteryzowanego zaufania do maszyn autonomicznych (<i>Krzysztof Soloduha</i>).....	60
Rozdział 6. Stosowanie sztucznej inteligencji w kształceniu. Możliwości i granice zaufania (<i>Marek Jakubiak</i>).....	78
Rozdział 7. Sztuczna inteligencja jako (nie)skuteczne narzędzie w walce z mową nienawiści (<i>Marcin Rojszczak</i>)	95
Rozdział 8. Big Data, sztuczna inteligencja i zrównoważony rozwój miast... w epoce (post) COVID-19 (<i>Piotr Wójcik, Grzegorz Kula</i>).....	117
Rozdział 9. Estetyka spekulatywna, percepcja maszynowa i sztuczna agencyjność w ujęciu organizacyjnym (<i>Adam Dżidowski</i>).....	131

Od redaktorów tomu

Współczesne badania nad sztuczną inteligencją (SI) mają charakter interdyscyplinarny. Ich jądro stanowi wprawdzie informatyka, dbająca o właściwe oprogramowanie i sprzętową architekturę systemów SI, lecz równie ważne są nauki wnikające w naturę inteligencji ludzkiej, na której ta sztuczna ma się wzorować. Należą do nich: psychologia poznawcza – dająca wgląd w strukturę i funkcję intelektu, różne działy filozofii – jak chociażby teoria poznania, filozofia umysłu i etyka, czy wyrastająca z psychologii i filozofii kognitywistyka – wielowątkowy obszar badań, zorientowany na budowę całościowych modeli umysłu. Nauki te wymieniamy tylko dla przykładu, pragnąc zwrócić uwagę na fakt, jak szeroki i zróżnicowany jest front nauk uczestniczących w projekcie sztucznej inteligencji.

Trzeba przyznać, że w XXI wieku projekt ten wkroczył w fazę istotnego przyspieszenia. Symbolizują go przede wszystkim systemy uczące się, które zmierzają coraz szybciej w stronę systemów autentycznie autonomicznych – to znaczy takich, które nie tylko doskonałą sposób realizacji celów stawianych im przez użytkownika, ale takich, które kreują samodzielnie cele swojego działania.

W dobie systemów tego rodzaju coraz większego znaczenia nabiera kwestia zaufania – zaufania ludzi do maszyn, które przejmują od ludzi wciąż nowe zadania i coraz głębiej ingerują w ich życie. Czy maszyna o inteligencji dorównującej zdolnościom człowieka, a więc maksymalnie wszechstronna, ucząca się i autonomiczna, nie będzie stanowić dla nas zagrożenia? W jakie mechanizmy ją wyposażyć, aby działała zgodnie z naszymi instrukcjami i oczekiwaniami? Czy już teraz, w dobie maszyn quasi-autonomicznych, kontrolowanych przez nas na poziomie realizowanych celów, możemy mieć do systemów SI pełne zaufanie?

W oddawanym do rąk czytelnika wyborze tekstów zagadnienie zaufania podejmujemy w sposób możliwie aktualny i wszechstronny. Odnosimy się zarówno do kwestii ogólnych, wręcz filozoficznych, związanych z narzuceniem na sposób działania maszyn pewnych norm, które od wieków postulują etycy (np. Arystoteles); jak również do kwestii bardzo szczegółowych, osadzonych

w kontekście bieżących zastosowań. W obszarze zastosowań uwypuklamy kwestie tak różnorodne, jak zaufanie do systemów SI wspomagających edukację, projektowanie bezpiecznych miast przyszłości (tzw. *smart cities*) czy zaufanie do programów i systemów usprawniających funkcjonowanie różnego rodzaju organizacji. Spośród wielu czynników wzmacniających zaufanie do sztucznej inteligencji kładziemy nacisk na dwa: skuteczność systemu połączoną z bezpieczeństwem użytkowników, oraz jego poznawczą przejrzystość połączoną z umiejętnością zrozumiałego dla człowieka wyjaśniania podejmowanych przez system decyzji.

Konkluzje autorów nie są jednolite. Niektórzy są optymistami, przekonując, że nawet najbardziej rozwinięta sztuczna inteligencja pozostanie czymś na kształt kontrolowanej przez człowieka „mechanicznej lalki”. Inni są bardziej sceptyczni, licząc się z możliwością zaistnienia systemów sztucznych, które przypominają bardziej „mroczne widmo” – czyli skrajnie niebezpieczny dla człowieka artefakt, zdolny do przejęcia nad nim fizycznej i psychicznej kontroli.

Książkę dedykujemy Pamięci naszej nieżyjącej już i nieodżałowanej Koleżanki, doktor Heleny Bulińskiej-Stangreckiej, która wraz z nami organizowała w roku 2021 międzynarodową konferencję naukową pt. *Zaufanie do systemów sztucznej inteligencji*. Wszystkie zamieszczone w książce teksty są oparte na referatach wygłoszonych podczas tej konferencji.

Paweł Stacewicz i Marek Jakubiak

Rozdział 1

Sztuczna inteligencja w odbiorze społecznym. Metafory a rzeczywistość

Public perception of Artificial Intelligence. Metaphors vs. reality

Jarosław Arabas, Jarosław Chudziak

Politechnika Warszawska

Wydział Elektroniki i Technik Informacyjnych

Instytut Informatyki

Streszczenie. Rozdział ma charakter wprowadzający w temat percepcji społecznej pojęcia sztucznej inteligencji (SI). Przybliży zagadnienie SI w kontekście technicznym, rozumianej jako systemy, programy lub algorytmy, które mają zdolność do budowania wiedzy, uczenia się i adaptacji do zmieniającego się środowiska. Często zbierają, selekcjonują i agregują dane w celu wytworzenia modeli, wykorzystywanych następnie do wspomagania podejmowania decyzji. Coraz częściej wkraczają również w obszary o naturze kreatywnej (muzyka, film, literatura, gry komputerowe). Zagadnienie społecznej odbioru SI zostanie przedstawione za pomocą dwóch metafor, obrazujących popkulturowe rozumienie SI: metaforę mechanicznej lalki i mrocznego widma.

Abstract. The chapter is introductory to the topic of social perception of the artificial intelligence. It introduces the issue of AI in the engineering context, understood as systems, programs or algorithms, which have the ability to build knowledge, learn and adapt to a changing environment. They often collect, select, and aggregate data to produce models that are then used to support decision-making. Increasingly, they are also entering areas of a creative nature (music, film, literature, computer games).

The issue of social perception of AI will be presented using two metaphors, illustrating the pop culture's understanding of AI systems: the metaphor of a mechanical doll and a phantom menace.

Słowa kluczowe: Sztuczna inteligencja, modele, algorytmy, dane, roboty, kultura, społeczeństwo, zaufanie

Keywords: Artificial intelligence, models, algorithms, data, robots, culture, society, trust

1.1. Wprowadzenie

Termin „sztuczna inteligencja” został prawdopodobnie zaproponowany po raz pierwszy przez amerykańskiego naukowca Johna McCarthego w roku 1954, w celu nadania nazwy pracom, których ambicją było konstruowanie inteligentnych maszyn, czyli takich, w których działaniu można się dopatrzeć znamion inteligencji, uznawanych za właściwe ludziom.

Okres II wojny światowej oraz kolejnych 20 lat był czasem narodzin informatyki, a tym samym obfitował w dynamiczny rozwój wielu ważnych dla SI dziedzin nauki, podstawowych oraz stosowanych. Przykładowo można wspomnieć o badaniach z obszaru podstaw informatyki teoretycznej, teorii optymalizacji, teorii systemów, robotyki, cybernetyki, statystyki czy też lingwistyki matematycznej. Natomiast pomimo ich istotności, otrzymywane wyniki i ich implikacje pozostawały często znane, dostępne i rozumiane wyłącznie przez specjalistów z poszczególnych obszarów badawczych [Woolridge, 2021].

Jednocześnie pewna tajemniczość, zwiększająca się szybkość odkryć naukowych, pierwsze powstałe urządzenia (jak np. roboty przemysłowe, pojazdy bezałogowe), spotęgowały pytanie o możliwości i granice dalszego rozwoju „inteligentnych” rozwiązań oraz ich wpływ na procesy społeczne. W dyskusji tej uczestniczyli zarówno zaangażowani w poszczególne badania naukowcy, jak również popularyzatorzy nauki oraz pisarze. W szczególności ta ostatnia grupa przyczyniła się do rozpropagowania pojęcia sztucznej inteligencji, pisząc o różnorodnych formach inteligentnych bytów, często w postaci humanoidalnych robotów (przykłady znajdziemy w powieściach i opowiadaniach Isaaca Asimova, Stanisława Lema, czy Philipa K. Dicka). Były one osadzone w wielowymiarowym spektrum narracji stawiających często pytania o naturę nowego świata w wymiarach psychologicznym, socjologicznym czy ogólniej filozoficznym. Często podstawą dla dzieł z obszaru fikcji literackiej (i filmowej) były profesjonalne monografie, które podsumowywały aktualny stan rozwoju SI, definiowały kolejne wyzwania i prognozowały następne osiągnięcia.

Wymagało to równocześnie przyspieszonego „społecznego” zrozumienia, akceptacji, a często także oswojenia efektów nowej rewolucji komputerowej, komunikacyjnej i medialnej poprzez stworzenie odpowiednich narracji budowanych na doświadczeniach i metaforach życia codziennego. Przyczyniło się to również do ukształtowania stereotypów, opinii, oraz sądów dotyczących SI, które często bazowały na kulturowych mitach, symbolach czy wzorcach zaczerpniętych z literatury, filmów i innych produktów kultury masowej [Cave, 2020; Hermann, 2021].

Obecnie dyskurs wokół definicji i znaczenia SI nasilił się w związku z zachodzącą intensywną transformacją cyfrową społeczeństw i gospodarki

[Nath, 2020]. Wprowadzone zostały liczne nowe rozwiązania i produkty, takie jak mikrokomputer, laptop, notebook, smartfon, inteligentne AGD i wiele innych. Z drugiej strony zaistniały nowe, niewidoczne i często nie do końca zrozumiałe pojęcia takie jak dane, informacja, algorytmy, uczenie maszynowe, sieci neuronowe, czy internet rzeczy. Wymagają one nowych kompetencji poznawczych i użytkowych, opanowanych co najmniej na poziomie minimalnym poprzez jednostki, jak i całe grupy społeczne. Stąd też pilna potrzeba wyinterpretowania nowych pojęć, cyfrowych bytów i procesów we własnym wewnętrznym świecie wiedzy potocznej [Nader, 2022; Cave, 2018].

Dyskusje dotyczące tematyki sztucznej inteligencji i jej roli w kształtowaniu naszego życia i naszych funkcji społecznych toczą się w świecie akademickim, szeroko rozumianej przestrzeni medialnej, w obszarze gospodarki, ale również w ramach struktur i instytucji publicznych [European Parliament, 2021]. Poruszane tematy dotyczą kwestii właściwego stymulowania rozwoju badań i zastosowań SI, kwestii moralnych i etycznych [Coeckelbergh, 2020] oraz różnorodnych aspektów prawnych [Corrales, 2018]. Wśród wątków szczegółowych są takie, jak rola SI w rozwiązaniu największych współczesnych wyzwań ludzkości (klimat, redystrybucja zasobów, energia itd.), w przemodelowaniu modeli biznesowych poszczególnych sektorów gospodarczych lub obszarów funkcjonalnych (przemysł .0), w tworzeniu alternatywnych społecznych światów cyfrowych. Szczególne miejsce ma debata w kontekście utraty zaufania i stworzenia potencjalnych zagrożeń dla naszej egzystencji jako ludzi [Glikson, 2020].

W niniejszym rozdziale chcemy zwrócić uwagę na wybrane elementy obszaru SI, szczególnie istotne obecnie w czasie bardzo szybkich procesów transformacji cyfrowej. Ze względu na drogę, którą przebył obszar SI, przywołamy tylko jego wybrane elementy, które w naszej subiektywnej ocenie, mogą być przedmiotem dyskusji, opinii bądź sądów w odbiorze społecznym. Pokażemy jednocześnie metafory i praktyczne ich realizacje w zastosowaniach SI. Uzupełnimy dyskusję wprowadzeniem w podstawowe terminy i definicje z obszaru SI.

1.2. Interpretacje i metafory sztucznej inteligencji

Metafory pełnią istotną rolę w naszym życiu społecznym, jak również w życiu poszczególnych jednostek. Wykorzystujemy je do zbierania i przechowywania informacji, a następnie do interpretacji świata. Pomagają nam zrozumieć złożone mechanizmy i pojęcia otaczającego świata z wykorzystaniem

obrazów, narracji, pojęć i sytuacji, które już mamy „przyswojone” [Lakoff, 2003]. Jednocześnie pozostają integralną częścią języka i naszej komunikacji interpersonalnej, stąd konkretne wybory metafor są bardzo ważne z powodu ich wpływu na nasze myślenie i działanie. Ich wybór, świadomy bądź nie, umożliwia modelowanie narracji lub argumentacji, otwierając lub zamykając ścieżki powyższych procesów [Bal, 2009].

Naszym zdaniem odbiór społeczny sztucznej inteligencji jest rozpięty między dwiema skrajnymi metaforami: mechanicznej lalki oraz mrocznego widma.

Mechaniczna lalka to system techniczny, który jest do pewnego stopnia substytutem człowieka, będąc jednocześnie człowiekowi poddanym. Przykładem realizacji takiego podejścia są roboty humanoidalne, zaspakajające naturalną ludzką potrzebę relacji, wykorzystywane w Japonii przez osoby starsze (innym przykładem są robotyczne psy – aibo). Innym przykładem występującym w kulturze masowej są roboty R2-D2 oraz C-3PO – bohaterowie drugiego planu w *Gwiezdnym Wojnach* czy też Wall-E oraz Eve – główni bohaterowie filmu *Wall-E*.

Mroczne widmo z kolei ma postać systemu technicznego, którego głównym atrybutem jest posiadanie wolnej woli wraz z negatywnymi tego konsekwencjami. Jednym z pierwszych wyrazistych przykładów pozostaje komputer HAL z *Odysei kosmicznej*, którego wolna wola doprowadziła do konfrontacji z człowiekiem. Inny, bardziej współczesny przykład to tytułowy komputer z filmu *Matrix*, który wytwarza w ubezwłasnowolnionych ludziach iluzję prawdziwego szczęśliwego życia.

Interpretacja tych dwóch metafor pozwala wyodrębnić kilka właściwości, które wpływają na to, jak poszczególne rozwiązania SI mogą być odbierane i oceniane społecznie. Poniżej wymienimy je i krótko skomentujemy.

Istotnym elementem oceny systemu SI jest sposób i zakres jego interakcji z otoczeniem, czyli zarówno ilość jak i rodzaj informacji pozyskiwanej z otoczenia, sposób jej przetworzenia oraz sposób oddziaływania na otoczenie. Szczególną uwagę zwracają systemy, które mogą wkraczać w prywatność – pozyskiwać i przetwarzać szeroko rozumiane dane osobowe, a następnie wykorzystywać je do ingerencji w nasze życie. Niepokoją nas również systemy, których interakcja z otoczeniem ma skutki nieobojętne etycznie, jak np. systemy sterowania autonomicznych pojazdów czy systemy automatycznej diagnostyki medycznej.

Często system SI jest postrzegany przez pryzmat wyjaśnialności i przewidywalności. Przykładowo, jeśli rolą systemu SI jest wypracowanie jakiejś rekomendacji działania, np. wybór sposobu leczenia lub decyzja ekonomiczna, to pożądaną cechą jest możliwość uzasadnienia prawidłowości decyzji

lub możliwość jej przewidzenia przez człowieka. Kwestią otwartą pozostaje, czym tak naprawdę ma być wyjaśnienie, jak szczegółowe powinno ono być, i czy człowiek jest w stanie zweryfikować jego poprawność.

Na stosunek człowieka do systemu SI ma również wpływ ergonomia, czyli dopasowanie sposobu jego komunikacji z otoczeniem w taki sposób, aby był on intuicyjny, łatwy do użycia i nie wymagał zbyt wysoko specjalizowanych umiejętności [Russell, 2019].

Podsumowując, mamy dwie wiodące naszym zdaniem metafory: mechaniczną lalkę jako uosobienie systemu o ograniczonych kompetencjach w interakcji z otoczeniem, którego zachowania dadzą się przewidzieć lub wyłumaczyć, oraz mroczne widmo jako system nieprzewidywalny, niezrozumiały, mający zbyt duże kompetencje i autonomiczny potencjał kreatywny.

Na koniec warto wspomnieć, że różnice w wykorzystaniu metafor do interpretacji SI i jej wpływu na życie społeczne wiążą się nierzadko z różnicami pokoleniowymi. Różnice te można zaobserwować w zakresie i głębokości kompetencji cyfrowych, systemowych i obliczeniowych. Kompetencje z kolei poprzez praktykę działania determinują nasz świat pojęciowy oraz możliwości interpretacyjne na podstawie metafor, którymi operujemy w życiu codziennym [Kaplan, 2019].

1.3. Wybrane zastosowania systemów sztucznej inteligencji

Przyjrzyjmy się przez chwilę wybranym praktycznym realizacjom systemów sztucznej inteligencji w odniesieniu do diskutowanych powyżej metafor. W jakim stopniu metafory te wynikają z nowości i braku interpretacji nowych pojęć, procesów czy produktów sztucznej inteligencji? W jakim stopniu z różnorodności i szybkości pojawiania się nowości? W jakim z oczekiwań, tęsknot, obaw w stosunku do nieznanego, wyrażanych w literaturze, filmie i obrazach na przestrzeni kilkuset ostatnich lat? (por. [Murphy, 2018; Robinson, 2020; South, 2018]).

Niejako emblematycznymi przykładami realizacji metafory mechanicznej lalki mogą być roboty wykorzystywane w życiu codziennym w gospodarstwach domowych, roboty do towarzystwa osób starszych czy inteligentne programowalne zabawki. W przestrzeni cyfrowej różne formy „botów” (text-, chat- lub video-) redefiniują procesy interakcji w serwisach usługowych. W obszarze inteligentnych osobistych asystentek rozwiązania w formie aplikacji Alexa, Google Assistant lub Siri; w grach komputerowych byty cyfrowe (artefakty, postaci, stwory ...) zaludniają świat gry, a nawet duplikujące nasze tożsamości [Yannakakis, 2018].

Algorytmy SI wspomagają wiele procesów zarządzania wiedzą, w tym przygotowywania i przeprowadzania symulacji doświadczeń biologicznych tworzących nowe substancje w ramach biologii syntetycznej (często substancji aktywnych wykorzystywanych farmacji) w celu przyspieszenia doświadczeń, eliminacji odpadów biochemicznych itd. W tym obszarze możemy odnaleźć elementy obu omawianych metafor – przykładowo mechanicznych lalek jako inteligentnych dziedzinowo zorientowanych asystentów oraz mrocznych widm pod postacią specjalistycznych dedykowanych modeli uczenia maszynowego.

W ostatnich latach rozwiązania SI weszły do masowego zastosowania w wielu gałęziach przemysłu w ramach koncepcji przemysłu 4.0. Osiągnięto pełnoskalowe wdrożenia SI w obszarze robotów przemysłowych, zaawansowanej analityki danych biznesowych, inteligentnych rozwiązań w obszarze internetu rzeczy, adaptowalnych autonomicznych produktów i serwisów czy wreszcie rozwiązań rozszerzonej i wirtualnej rzeczywistości. Kolejnym aktualnie realizującym się wyzwaniem jest tworzenie wersji ekosystemu całkowicie redefiniującego środowisko współpracy ludzi i robotów. Można przyjąć, że mamy tutaj do czynienia z przecięciem dwóch metafor: metafory mechanicznej lalki i metafory mrocznego widma. Ta druga pozostaje odpowiedzialna za koordynację, optymalizację i bezpieczeństwo pracy całego ekosystemu ludzi i robotów.

Metafora mrocznego widma – sztucznego mózgu ludzkiego – znalazła swoje odzwierciedlenie w programach, które skutecznie pokonały mistrzów szachowych, a następnie mistrzów starochińskiej gry Go [Gerrish, 2019]. Magazynowanie ogromnych ilości danych, ich szybkie przeszukiwanie, dodatkowe czerpanie z możliwości rozproszonych danych w czasie rzeczywistym, praca z wykorzystaniem wielu języków przekraczają możliwości intelektualne osoby będącej nawet synonimem geniuszu ludzkiego.

Nawiązanie do sposobu funkcjonowania mózgu można również odnaleźć w różnych paradygmatach SI oraz rozwiązaniach w obszarze sztucznych sieci neuronowych. Bardzo często proponowane rozwiązania są inspirowane rezultatami badań z obszaru neuronauki i interpretowane są jako abstrakt czy analog względem elementów układu nerwowego.

Stworzenie możliwości skoordynowanego działania w postaci wielu agentów, czyli połączenie równolegle prowadzonej aktywności indywidualnej, umożliwia dużo efektywniejsze zbieranie doświadczeń i budowanie wiedzy. Jest to osiągnięte poprzez współpracę i współzawodnictwo agentów realizujących procesy kognitywne oraz podejmujących decyzje, uwzględniając wiele perspektyw jednocześnie (różne dyskursy, różne racjonalizacje). Tworzy to nowe możliwości przełączania się między kontekstami, z wykorzystaniem obrazów wielu rzeczywistości, jak również wielu scenariuszy działania. Praktycznym przykładem takich rozwiązań są „pokoje wojen” wykorzystywane

do rozgrywania symulowanych gier strategicznych w firmach i instytucjach, jak również SI wykorzystywana w strategicznych grach komputerowych.

Zauważmy, że mimo imponujących możliwości, specjalizowane do rozwiązywania danego zadania systemy SI są najczęściej „posłuszne” – są nadal mechaniczną lalką, chociaż o możliwościach przewyższających ludzkie w wąskim obszarze swoich kompetencji. Dostępnych jest coraz więcej „przystępnych” materiałów, jak można je samemu tworzyć, często z przygotowanych elementów (przykładowo [Govers, 2018]).

1.4. Wokół definicji sztucznej inteligencji

Mimo niemal pięćdziesięcioletniej historii związanej z pracami w obszarze sztucznej inteligencji jej precyzyjna definicja do tej pory pozostaje sprawą otwartą (por. [Russell, 2020; Woolridge, 2021; Ford, 2018])). Wynika to głównie z heterogeniczności tematów pozostających obszarze jej zainteresowań, jak również powszechności dostępu do jej rezultatów w poszczególnych grupach społecznych. Stąd wiele definicji jest nacechowanych obszarem zastosowań oraz specyfiką poszczególnych dziedzin naukowych, w ramach których prowadzone są badania z obszaru SI.

Jedna z ciekawszych, początkowych i szerokich w swoim zakresie definicji SI głosi, że sztuczna inteligencja zajmuje się wszystkim co wymaga uruchomienia inteligencji od człowieka. Oczywiście taka definicja, pozostając szeroką, jest jednocześnie bardzo nieostra, choćby ze względu na nieostrość samego pojęcia inteligencji (przykładowo: do jakiego stopnia wnioskowanie statystyczne pozostaje racjonalną cechą inteligencji człowieka).

Zatem, czym jest sztuczna inteligencja: algorytmami, procesami, produktami, „bytami”, „spojrzeniem na rzeczywistość”, „stanem umysłu”? Jako reprezentantom informatyki najbliższe są nam pojęcia wypracowane na jej gruncie i do nich będziemy się odnosić. Jedną z piękniejszych klasyfikacji SI przedstawili Russell i Norvig [Russell, 2020], definiując dwa wymiary SI: sposób manifestacji inteligencji oraz miarę jej odniesienia. Jeśli chodzi o sposób manifestacji, czynią oni rozróżnienie między rozumowaniem i działaniem, natomiast wśród miar odniesienia rozróżniają obiektywną (racjonalność) i subiektywną (jak ludzie).

Czynią oni również rozróżnienie między dwiema skrajnościami: „mocną SI”, zainteresowaną systemami myślącymi jak ludzie, oraz „słabą SI”, która obejmuje pozostałe trzy możliwe kombinacje cech. Zgodzono się, że „słaba SI” jest wystarczająco mocna, aby uzyskać znaczny wpływ na gospodarkę i społeczeństwo.

Systemy myślące jak ludzie to przede wszystkim obszar intelektualnej spekulacji i naukowych poszukiwań w obszarach, w których komponent techniczny przyjmuje często rolę służebną. Szczególne zainteresowanie tym podejściem objawia się w obszarze kognitywistyki. Postulat stworzenia takich systemów jest również przedmiotem refleksji filozoficznej, w ramach której należałoby wskazać nurt transhumanizmu. Również Meta (dawniej Facebook) podejmuje działania w tym kierunku, snując wizję projektu Metaverse. Wizja stworzenia myślącego jak człowiek systemu była i jest niezwykle atrakcyjna dla twórców kultury i stanowi poniekąd przedłużenie opowieści o Golemie czy Frankensteinie; dobrymi przykładami są filmy o Robocopie czy też serial *Westworld* (Netflix) [South, 2018].

Systemy myślące racjonalnie, to przede wszystkim obszar wnioskowania, czyli różne zastosowania logiki i jej uogólnień. Wpisują się one przede wszystkim w obszar technik podejmowania decyzji i mogą pełnić rolę służebną wobec systemów racjonalnego działania.

W obszarze racjonalnego działania sztuczną inteligencję można opisać jako zdolność systemu technicznego do budowania „obliczeniowej” reprezentacji (selekcji, analizy i syntezy) wybranego wycinka świata rzeczywistego z wykorzystaniem zebranych sygnałów (danych) zewnętrznych, a równolegle uczenia się i odpowiedniego dostosowywania (do wyzwań otwartego środowiska zewnętrznego), i wreszcie wykorzystywania zgromadzonej wiedzy do realizacji zadań i osiągania postawionych celów. Budowa takiego systemu polega na realizacji technicznej trzech funkcji inteligentnego zachowania: modelowanie (w celu objaśniania lub prognozowania), podejmowanie decyzji (wnioskowanie, analiza scenariuszy, analiza wrażliwościowa, optymalizacja) oraz interakcja z otoczeniem (w tym widzenie komputerowe, rozpoznawanie głosu, przetwarzanie języka naturalnego).

Budowa systemów działających jak ludzie jest zastosowaniem SI między innymi w obszarze komputerowych gier akcji, techniki filmowej oraz innych obszarów szeroko pojętej branży rozrywkowej, a także branży specyficznych usług, takich jak systemy asystujące człowiekowi. W tym kontekście nie wypada nie wspomnieć testu Turinga, który jest próbą uściślenia pojęcia podobieństwa do człowieka. Test ten, oryginalnie sformułowany na zupełnie inne potrzeby, w odniesieniu do systemu SI głosi, że system może być uznany za działający jak człowiek, jeśli człowiek wchodzący w interakcję z tym systemem (interakcję językową) nie może na jej podstawie odróżnić, czy ma do czynienia z drugim człowiekiem, czy z wytworem techniki¹.

¹ Szersza omówienie testu Turinga w kontekście historii badań nad sztuczną inteligencją oraz współczesnym kontekście budowy systemów godnych zaufania zawiera rozdział drugi.

Obecnie jednym z najważniejszych obszarów badawczych jak i zastosowań SI pozostaje uczenie maszynowe, czyli wytwarzanie programów komputerowych mających zdolność uczenia się z danych, tzn. pozyskiwania umiejętności przewidywania przyszłych stanów procesu, dla którego pozyskane zostały dane historyczne dotyczące zadań poprawnie i efektywnie zrealizowanych. Aby to osiągnąć, program tworzy model, korzystając z pozyskanych danych oraz starając się uzyskać najlepszą jakość mierzoną pewną założoną miarą efektywności (i/lub poprawności). Wiele rodzajów uczenia maszynowego można systematyzować na podstawie sposobu ingerencji w proces uczenia się (np. uczenie się nadzorowane, nienadzorowane, ze wzmacnianiem), momentu przetwarzania (w czasie rzeczywistym lub poza nim) czy też sposobu wykorzystania wyników (np. do przewidywania modelowanego zjawiska lub do przedstawienia rekomendacji decyzji). Systemy uczące się stały się kluczowym elementem nowoczesnych produktów i serwisów, począwszy od rozwiązań e-handlu, sieci społecznościowych, smartfonów itp. Takie funkcje jak rekomendacje, czytanie i analiza obrazu, rozpoznawanie głosu, przetwarzanie i generowanie wypowiedzi w języku naturalnym, autonomiczne sterowanie robotami (i pojazdami) to tylko wybrane przykłady zastosowań maszynowego uczenia się.

Wśród modeli używanych w maszynowym uczeniu się szczególne miejsce zajmują sieci neuronowe, inspirowane budową układu nerwowego. Znanych jest wielu różnorodnych typów modeli sieci neuronowych dostosowanych do specyfiki różnych typów przetwarzanych danych, rodzaju modelowanych zjawisk oraz dziedzin zastosowań.

W szczególności wprowadzenie po roku 2007 efektywnych modeli głębokich sieci neuronowych, skrótowo określanych jako głębokie uczenie się (ang. *deep learning*) [Goodfellow, 2017], otworzyło nowe możliwości zastosowań. Stało się tak między innymi dzięki szybkości i dokładności o rząd lub kilka rzędów lepszej niż w poprzednio stosowanych modelach (np. w obszarze wizji komputerowej). Było to wspomagane jednoczesnym zwiększeniem dostępności mocy obliczeniowej komputerów (lokalnie i w ramach chmur obliczeniowych), nowymi architekturami przetwarzania równoległego oraz dostępnością ogromnych ilości danych treningowych.

Badania nad metodami reprezentacji i przetwarzania wiedzy umożliwiające odwzorowywanie świata rzeczywistego za pomocą technik cyfrowych cieszyły się specjalnym zainteresowaniem od samego początku badań nad SI. Podstawowe modele wykorzystywane w rozwiązaniach z zakresu SI można podzielić na modele symboliczne i subsymboliczne. Najlepszym przykładem tych drugich są właśnie sieci neuronowe, natomiast do przykładów pierwszych zaliczamy reprezentacje logiczne, grafowe, algebraiczne lub bardziej złożone modele semantyczne czy ontologiczne.

Innymi elementami definiującymi obszar SI są metody, techniki i narzędzia, które wykorzystujemy w procesie podejmowania decyzji. Należą do nich algorytmy automatycznego wnioskowania, wszelkiego rodzaju algorytmy heurystyczne i metaheurystyczne (w tym ewolucyjne i genetyczne) służące do realizacji inteligentnego procesu przeszukiwania przestrzeni możliwych rozwiązań.

Metody modelowania i optymalizacji bardzo często wspomagane są efektywnymi dedykowanymi narzędziami i technikami, a także specjalnymi bibliotekami rozwiązań przygotowanymi w wielu językach (środowiskach) programowania takich jak Python, C++, Java, R, TensorFlow [Geron, 2018].

Coraz częściej dostępne są również kompletne ekosystemy i platformy do projektowania złożonych rozwiązań sztucznej inteligencji, a ich liczba oraz zaawansowanie znacząco wzrosły wraz z rozpowszechnieniem się i powszechną adaptacją rozwiązań chmurowych (por. platformy firm Microsoft, Google, AWS, Nvidia, IBM, SAS i in.).

1.5. Sztuczna inteligencja w praktyce i percepcji życia codziennego

W życiu codziennym (w pracy, w domu, w urzędach) spotykamy się z urządzeniami określanymi mianem inteligentnych. Praktyczna forma i sposób wykorzystania inteligentnych urządzeń buduje (często weryfikuje) nasze pragmatyczne rozumienie SI w różnych kontekstach i perspektywach.

Wiele inteligentnych urządzeń już teraz posiada większość z następujących cech czy też elementów systemów SI: potrafią komunikować się w języku naturalnym z odbiorcą, mają zdolność definiowania wzorców postępowania uznawanych za poprawne, przeważnie również optymalne (obszar optymalizacji, modelowania, podejmowania decyzji), posiadają umiejętność adaptacyjnego uczenia się, efektywnego przeszukiwania oraz odkrywania wzorców w danych, rozwiązywania złożonych problemów z wykorzystaniem reguł symbolicznych i subsymbolicznych. Sposób ich wbudowywania w praktyki, procesy i wzorce naszych zachowań determinuje odbiór SI w percepcji społecznej

Ze względu na powszechność systemów SI, dyskutując pojęcie i definicje SI, nie możemy pominąć kwestii ról społecznych w kontekstach, w których człowiek wchodzi w interakcje z obszarem SI (często ról tych może być kilka równoległe). W uproszczeniu możemy wyróżnić podział na twórców SI oraz jej szerokie grono użytkowników. Następnie w grupach tych możemy oczywiście wyróżnić dalsze podziały. Przykładowo, wśród twórców SI można

wskazać architektów rozwiązań, projektantów i wykonawców, a wśród użytkowników wyróżnić grupę świadomą, wykorzystującą w pełni możliwości „nowych” rozwiązań, oraz odbiorców pasywnych, często nie zdających sobie sprawy z interakcji z SI. Każdą z wymienionych grup będzie cechował inny sposób rozumienia SI: jej zakresu, złożoności i aktualnych możliwości [Ford, 2018].

Pytanie o praktyczny aspekt SI jest również pytaniem o kontekst rzeczywistości, w której będziemy rozważali to pojęcie. Czy jest to nasza aktualna rzeczywistość poddana ewolucji i wzbogacona o nowe elementy, takie jak inteligentne roboty? Czy jest to rzeczywistość rozszerzona o funkcjonalności różnorodnych produktów i systemów, o siłę algorytmów uczenia maszynowego i innych technik sztucznej inteligencji? Czy może jest to rzeczywistość „zoptymalizowana” o zwielokrotnionej efektywności dzięki (hiper)automatyzacji procesów? Ta ostatnia wersja już teraz jest aktualna, zatem pozostaje wersją minimalną.

Możliwa też jest druga opcja, w której SI umożliwi nam budowanie światów alternatywnych w przestrzeni cyfrowej. Przeplatanie się świata rzeczywistego oraz wirtualnego, a także oddziaływanie SI na oba światy, rodzi wiele pytań etycznych i ontologicznych. Kim będziemy w świecie hybrydowym? Jakie będą reguły rządzące różnymi wersjami światów? Kto będzie je definiował? Jakie awatary (ile awatarów) będą nas reprezentowały? Co będzie znaczyło istnienie świata wirtualnego? Co będzie definiowało wirtualne tożsamości cyfrowe? Czy i jakie zasady etyczne będą panować w świecie wirtualnym? Tych oraz wielu innych pytań i wyzwań jeszcze do końca nie zdefiniowaliśmy, natomiast pewne namiastki odpowiedzi możemy już znaleźć w światach zaawansowanych gier komputerowych [Yannakakis, 2018].

Coraz więcej państw wprowadza w zakres obszaru budowy i rozwoju kompetencji swoich obywateli elementy związane z cyfryzacją świata, sztuczną inteligencją i automatyzacją procesów (smart city, e-gov itd.). Podobnie postępują firmy komercyjne, chcąc stworzyć nowe możliwości rozwoju swoim pracownikom. Często rodzi to pytania o przyszłość zatrudnienia, formę przyszłego rynku pracy, konieczności zmiany czy uzupełnienia kompetencji pracowników, konieczności postawienia sobie pytań o scenariusze przyszłości, które będą kreowane na podstawie możliwości stwarzanych przez SI.

Opracowania i propozycje rozwiązań systemowych, instytucjonalnych czy też prawnych dotyczących SI powstają obecnie w różnych miejscach: w ośrodkach akademickich, w ramach organizacji i instytucji międzynarodowych (por. propozycje regulacji prawnych z obszaru SI wprowadzone przez Unię Europejską), na poziomie poszczególnych krajów (lub instytucji krajowych) czy też w ramach różnorodnych grup zawodowych. Wprowadzając

własne modele, interpretacje i metafory SI, współdzielone kulturowo przez docelowe grupy komunikacji, proponowane rozwiązania podejmują próby osadzenia SI w powszechnej świadomości oraz stworzenia ram i ścieżek dla następnych etapów rozwoju.

1.6. Zamiast podsumowania

Jakie jest indywidualne, społeczne i publiczne zaufanie wobec sztucznej inteligencji? Do jakiego stopnia powinniśmy wyjaśniać sposoby działania SI? Jak i kto definiuje bezpieczeństwo jednostek i grup społecznych w ramach poszczególnych procesów i interakcji? Czy będziemy ufać samochodom poruszającym się bez kontroli kierowcy? Czy intuicja pracowników, jako uzupełnienie reguł algorytmicznych, jest bardziej akceptowalna w momencie odmowy pożyczki lub ubezpieczenia, niż decyzja czysto algorytmiczna?

Czy metafory SI pomogą w przygotowaniu lepszych opowieści odpowiadających na wyżej postawione pytania o zaufanie, bezpieczeństwo i wyjaśnialność? Jakie pozytywne narracje powinniśmy budować, aby ją oswoić? I czy jest to jeszcze możliwe przy złożoności aparatu matematycznego, ale również pojęciowego, niezbędnego do jej zrozumienia? Czy na przeszkodzie nie stanie tutaj szerokość i głębokość wiedzy niezbędnej do zrozumienia sposobów działania SI – często wiedzy pasywnej, bazującej na ograniczonych interakcjach z systemami w roli użytkowników? Staraliśmy się argumentować, że metafory mogą pomóc w wyinterpretowaniu nowej, wirtualnej lub rozszerzonej, rzeczywistości. Być może również pomogą „oswoić nowe” wśród szerszych grup społecznych. Natomiast nadal pozostawia wiele miejsca dla ekspertów, architektów, prawników, jak również socjologów i filozofów, dla zbudowania nowych ram naszego działania oraz stworzenia być może nowych opowieści i metafor.

Odpowiedzi na postawione wcześniej pytania z obszaru etyki, moralności, zarządzania, polityki itd., będą kształtowały nasze poczucie bezpieczeństwa oraz poziom zaufania do nowych rozwiązań budowanych z wykorzystaniem technologii SI. W ich tworzeniu zostaną wykorzystane metafory istniejące już w kulturze masowej, ale również nowe bazujące na współcześnie kreowanych artefaktach cyfrowej rzeczywistości SI.

Wśród wielu pytań, dla których odpowiedzi pozostają kwestią otwartą, są pytania o to, czy przywołane w niniejszym rozdziale metody, techniki i narzędzia SI pozostaną wiodącymi za 5, 10 lub 25 lat? Czy jedyną możliwością ich objaśniania w kulturze masowej jest odwoływanie się do doświadczeń i metafor

„przedrobotycznych”? Czy wymiana pokoleniowa zmieni nasze metafory czy jedynie wyinterpretuje w nowym języku? Czy metafory budowane na strachu znikną wraz z przyrostem doświadczenia wynikającym z koegzystencji z inteligentnymi rozwiązaniami? W jaki sposób dyskusje prowadzone wśród specjalistów, a dotyczące etyki i moralności w inteligentnym świecie cyfrowym, wpłyną na kształtowanie naszych wyobrażeń? Przynajmniej przez jakiś czas pytania te pozostaną ciekawym obszarem dalszych refleksji w obszarze SI.

Co wydaje się pewnym: SI zmienia nasz świat na zawsze. I zmiana ta będzie prawdopodobnie w dłuższym okresie głębsza, niż większość ludzi dziś zdaje sobie z tego sprawę. Bez względu na to, o jakim fragmencie rzeczywistości mówimy, rozszerzy ona nasze możliwości, jeśli nie całkowicie je przekształci.

BIBLIOGRAFIA

- Bal M., (2009), *Narratology: introduction to the theory of narrative*, University of Toronto Press.
- Cave S., Craig C., Dihal K., et al., (2018), *AI narratives: Portrayals and perceptions of AI and why they matter*, The Royal Society. Dostępne pod adresem: <https://royalsociety.org/topics-policy/projects/ai-narratives/> (dostęp: 26.05.2022).
- Corrales M., Fenwick M., Forgó N. (eds.), (2018), *Robotics, AI and the Future of Law*, Springer.
- Coeckelbergh, M., (2020), *AI Ethics*, MIT Press.
- European Parliament, Panel for the Future of Science and Technology (STOA)(2021), *What if we chose new metaphors for artificial intelligence?*, Dostępne pod adresem: [https://www.europarl.europa.eu/stoa/en/document/EPRS_ATA\(2021\)690024](https://www.europarl.europa.eu/stoa/en/document/EPRS_ATA(2021)690024) (dostęp: 26.05.2022).
- Ford M., (2018), *Architects of Intelligence: The truth about AI from the people building it*, Packt Publishing.
- Geron A., (2018), *Uczenie maszynowe z użyciem Scikit-Learn i TensorFlow. Pojęcia, techniki i narzędzia do tworzenia inteligentnych systemów*, Helion.
- Gerrish S., (2019), *How smart machines think*, MIT Press.
- Glikson E., Williams Woolley A., (2020), *Human Trust in Artificial Intelligence: Review of Empirical Research*, *Academy of Management Annals*, 2(14), <https://doi.org/10.5465/annals.2018.0057> (dostęp: 26.05.2022).
- Goodfellow I., Bengio Y., Courville A., (2017), *Deep Learning*, MIT Press.
- Govers F.X., (2018), *Artificial Intelligence for Robotics: Build intelligent robots that perform human tasks using AI techniques*, Packt Publishing.
- Hermann I., (2021), *Artificial intelligence in fiction: between narratives and metaphors. AI & Society*. Dostępne pod adresem: <https://link.springer.com/article/10.1007/s00146-021-01299-6> (dostęp: 26.05.2022).
- Iansiti M., Lakhani K.R., (2020), *Competing in the Age of AI: Strategy and Leadership When Algorithms and Networks Run the World*, HBR Press.
- Kaplan A., Michael Haenlein A., (2019), *Siri, Siri in my Hand, who's the Fairest in the Land? On the Interpretations, Illustrations and Implications of Artificial Intelligence*, *Business Horizons*, 62(1), s. 15–25.

- Lakoff G., Johnson M., (2003), *Metaphors We Live by*, University of Chicago Press.
- Murphy R.R., (2018), *Robotics Through Science Fiction: Artificial Intelligence Explained Through Six Classic Robot Short Stories*, MIT Press.
- Nader K., Toprac P., Scott S. et al., (2022), *Public understanding of artificial intelligence through entertainment media*. AI & Society, Dostępne pod adresem: <https://doi.org/10.1007/s00146-022-01427-w> (dostęp: 26.05.2022).
- Nath S.V., Dunkin A., Mahesh Chowdhary M., Nital Patel N., (2020), *Industrial Digital Transformation: Accelerate digital transformation with business optimization, AI, and Industry 4.0*, Packt Publishing.
- Robinson S., (2020), *AI in Sci-Fi: Fictional Artificial Minds and the Real World Awaiting Them*, Paleos Media.
- Russell S., Norvig P., (2020), *Artificial Intelligence: A Modern Approach*, Pearsons, 4th ed.
- Russell S., (2019), *Human Compatible: Artificial Intelligence and the Problem of Control*, Viking.
- South J.B., Irwin W., (2018), *Westworld and Philosophy: If You Go Looking for the Truth, Get the Whole Thing*, Wiley-Blackwell.
- Yannakakis G.N., Togelius J., (2018), *Artificial Intelligence and Games*, Springer
- Woolridge M., (2021), *A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going*, Flatiron Books.

Rozdział 2

Wyjaśnianie, zaufanie i test Turinga

Explaining, trusting and Turing test¹

Paweł Stacewicz

Politechnika Warszawska

Wydział Administracji i Nauk Społecznych

Streszczenie. Test Turinga, będący werbalnym testem nierozróżnialności inteligencji maszynowej i ludzkiej, stanowi ważną historycznie ideę, która wyznaczyła istotny po dziś dzień sposób myślenia o projekcie sztucznej inteligencji (SI). Zgodnie z nim wzorcem dla SI jest inteligencja ludzka, a kluczową umiejętnością SI ma być jej sprawność komunikacyjna – polegająca m.in. na wyjaśnianiu podejmowanych przez maszynę decyzji.

Przyjmując współczesny punkt widzenia, w artykule rozwijamy i uzasadniamy tezę, że zdolność systemów SI do wyjaśniania (motywów, decyzji, zachowań...) jest czynnikiem istotnie zwiększającym zaufanie użytkownika do systemu SI. Jest ona szczególnie ważna wtedy, gdy użytkownik akceptuje zasadę ograniczonego zaufania do systemu (to znaczy nie ufa mu bezgranicznie, jest świadomy jego ograniczeń i możliwości popełniania błędów).

Przejście przez maszynę oryginalnego testu Turinga nie gwarantuje, że maszyna jest z punktu widzenia człowieka godna zaufania; wręcz przeciwnie – w koncepcji testu kryje się idea maszyny zdolnej „oszukać” lub „przechytryć” człowieka.

Postulujemy, że test gwarantujący wysoki poziom zaufania powinien być: a) nieimitacyjny, b) niebehavioralny, c) ukierunkowany na umiejętność wyjaśniania, uwzględniającego jednak konstrukcję i sposób działania maszyny (a nie tylko oczekiwania człowieka), d) ukierunkowany na zdolność maszyny do uczenia się.

Abstract. The Turing test, which is a verbal test of the indistinguishability of machine and human intelligence, is a historically important idea that has set a way of thinking about the AI (artificial intelligence) project that is still relevant today. According to it, the benchmark/blueprint for AI is

¹ Publikacja powstała w ramach projektu Narodowego Centrum Nauki pt. „Turing, Ashby i aktywność mózgu” (nr projektu: 2020/37/B/HS1/01809; konkurs OPUS 19; PI: Hajo Greif; CI: Paweł Stacewicz).

human intelligence, and the key skill of AI should be its communicative proficiency – which includes explaining decisions made by the machine.

Taking a contemporary point of view, we develop and justify the thesis that the AI's ability to explain (motives, decisions, behavior...) is a factor that significantly increases the user's trust in the AI system. It is especially important when the user accepts the principle of limited trust in the system (that is, he does not trust it unlimitedly, he is aware of its limitations and the possibility of making mistakes).

Passing the original Turing test by a machine does not guarantee that the machine is, from the point of view of a human, trustworthy; on the contrary, the idea of a machine capable of „fooling” or „outsmarting” a human is inherent within the concept of the test.

We postulate, that a test that guarantees a high level of trust should be: a) non-imitative, b) non-behavioral, c) focused on the ability to explain, taking into account, however, the design and rules of operation of the machine (and not just human expectations), d) focused on the machine's ability to learn.

Słowa kluczowe: sztuczna inteligencja, test Turinga, zaufanie, zasada ograniczonego zaufania, wyjaśnianie, maszynowe uczenie się

Keywords: Artificial Intelligence, Turing test, trust, principle of limited trust, explaining, machine learning

2.1. Wprowadzenie

Pojęcie testu Turinga, nazywanego pierwotnie *grą w naśladownictwo* [Turing, 1950], stanowi trwały i po dziś dzień niezwykle inspirujący element dyskusji nad sztuczną inteligencją (SI). Ujmując rzecz literalnie, pojęcie to oznacza pewną metodę testowania systemów informatycznych pod kątem inteligencji. Polega ona na swobodnej rozmowie człowieka z maszyną: jeśli w jej trakcie kompetentny sędzia nie potrafi odróżnić systemu od człowieka, musi uznać, że ma do czynienia z bytem inteligentnym, zaś przy bardziej radykalnym podejściu: z bytem myślącym².

Ujmując rzecz nieco szerzej, test Turinga nie jest tylko testem, lecz pewną doniosłą historycznie ideą, która na długie lata ukształtowała sposób myślenia o projekcie sztucznej inteligencji. Zgodnie z tą ideą sztuczna inteligencja powinna być budowana na podobieństwo ludzkiej: winna zatem rozwiązywać problemy podejmowane przez ludzi (np. decyzyjne), realizować właściwe im procesy poznawcze (np. logiczne wnioskowania) i przedstawiać swoje wyniki w zrozumiałej dla człowieka formie. Niezwykle ważną składową takiego sposobu myślenia jest położenie nacisku na sprawność komunikacyjną maszyn.

² Na trop takiej radykalnej interpretacji naprowadza sam Turing, który w swoim tekście z roku 1950, formułuje wprost pytanie *Czy maszyna może myśleć?*

Zgodnie z ideą testu ostateczną miarą podobieństwa do człowieka jest fakt, że system inteligentny posługuje się biegle językiem naturalnym i wchodzi z człowiekiem w sensowną językową interakcję.

Sam Turing podkreślał znaczenie tych umiejętności nie tylko w ramach samej formuły testu, ale także w ramach ogólnej refleksji nad przyszłością badań nad SI. Już w swoim raporcie badawczym z roku 1948, *Intelligent Machinery*, wskazał trzy ważne obszary przyszłych zastosowań, które mieszczą się w szerokim kręgu zagadnień komunikacji. Są to: uczenie się języków, tłumaczenie z danego języka na inny oraz sama komunikacja z maszyną w języku naturalnym [Turing, 1948, s. 2]. Trzeba przyznać, że współcześnie, w dobie systemów uczących się, multimedialnych i wysoce interaktywnych, obszary te należą do jądra badań nad SI [Russel, Norvig, 2020].

Konstruowane współcześnie maszyny wykazują, rzecz jasna, o wiele bogatszy i szerzej społecznie oddziałujący zakres interakcji z człowiekiem, niż to sugeruje koncepcja testu Turinga. Obok wciąż doskonalonych kompetencji językowych dzisiejsze systemy SI autentycznie przejmują część zadań od ludzi i podejmują istotne dla nich decyzje. To rodzi ważne pytania o zaufanie. Czy maszyna o inteligencji dorównującej zdolnościom człowieka, a więc maksymalnie wszechstronna, rozwijająca się i autonomiczna, nie będzie stanowić dla ludzi zagrożenia? Czy będzie działać zgodnie z naszymi instrukcjami i oczekiwaniami? Czy nie okaże się na tyle sprytna, by ukrywać swoje realne cele, pozorując tylko dążenie do celów określonych przez człowieka?

Omawiając ideę testu na inteligencję, Turing [Turing, 1950] nie odnosił się wprost do kwestii zaufania. Pisał jednak o ludzkich obawach przed stworzeniem maszyn inteligentnych, które podważyłyby – gdyby zaistniały – ugruntowaną kulturowo wizję człowieka jako istoty uprzywilejowanej, „wyższej ponad resztę stworzenia” [Turing, 1950, s. 444]. Obawy te, czyli jakaś forma ostatecznego braku zaufania do hipotetycznych maszyn przyszłości, stanowiły – wg Turinga – ważny czynnik natury psychicznej³, prowadzący do silnego przekonania mu współczesnych, że skonstruowanie maszyn inteligentnych jest zasadniczo niemożliwe. Dopuszczmy dla porządku, że sam Turing z takim przekonaniem polemizował.

Niezależnie od tak skrajnie rozumianego (braku) zaufania można postawić pytanie o to, czy oryginalny test Turinga jest testem na inteligencję budzącą zaufanie w węższym sensie? Czy maszyna, która go przejdzie, będzie dla nas wiarygodna w zakresie powierzanych jej zadań i przekazywanych informacji? A jeśli nie będzie, to w czy w idei testu tkwią jakieś wartościowe elementy, które pozwalają sformułować warunki brzegowe dobrego testu na zaufanie? Są to pytania, które podejmiemy w dalszej części rozdziału.

³ Oczywiście nie był to czynnik natury logicznej, bo z obaw przed maszynami czy z braku zaufania wobec nich w żaden sposób nie wynika teza, że stworzenie maszyn o inteligencji porównywalnej z ludzką jest niemożliwe.

2.2. Zaufanie a wyjaśnianie. Kontekst psychologiczny

Zaufanie jest relacją natury psychicznej, która polega na tym, że pewna osoba, nazwijmy ją podmiotem zaufania, żywi określonego rodzaju przekonanie co do możliwych działań drugiej strony, na przykład osoby, zwierzęcia, instytucji, urzędnika czy systemu. Jest ona mianowicie przekonana, że działania te będą zgodne z jej oczekiwaniami, które z kolei mają swoje uzasadnienie w deklaracjach, zobowiązaniach lub technicznych specyfikacjach (w przypadku artefaktów) drugiej strony. Relacja ta ma charakter kontekstowy. Znaczy to, że oczekiwania podmiotu zaufania zależą silnie od tego, czym jest druga strona relacji i jaką funkcję pełni. Na przykład, czy osoba, której podmiot miałby zaufać, jest pracodawcą, usługodawcą, współpracownikiem, partnerem życiowym itd.

Ponieważ jądrem relacji zaufania są oczekiwania wobec pewnego obiektu (np. osoby), psychologowie bardzo często określają zaufanie za pomocą pojęcia wiarygodności. Oto objaśniający tę kwestię fragment pracy psychologów poznawczych: „Decyzja o zaufaniu opiera się najczęściej na ocenie wiarygodności drugiej osoby. Wiarygodność jest więc podstawą zaufania, a także niezbędnym czynnikiem dla pełnej relacji zaufania. Wiarygodność drugiej osoby oznacza przekonanie, że osoba ta będzie zachowywać się w sposób zgodny z oczekiwaniami” [Jaklik, Łaguna, 2015].

Kluczową dla relacji zaufania cechą wiarygodności możemy – za M. Szynekiewiczem [2014] – zróżnicować dalej na wiarygodność merytoryczną i wiarygodność etyczno-moralną. Pierwsza zależy od wiedzy, doświadczenia i kompetencji dziedzinowych adresata postawy zaufania, druga – od przejawiających się w jego działaniu norm, zasad i poglądów etycznych⁴. Wiarygodność merytoryczna ma charakter bardziej podstawowy, ponieważ sądy i wybory etyczne muszą dotyczyć działań, które bez należytej wiedzy i bez odpowiednich kompetencji nie mogłyby zostać ani zaplanowane, ani podjęte. Ona też będzie nas interesować w sposób szczególnie, ponieważ na obecnym etapie rozwoju systemów informatycznych trudno mówić o maszynach, które autentycznie dokonują samodzielnych wyborów etycznych⁵. Ewentualna etyka czy moralność jest kwestią zewnętrzną – powiązaną z decyzjami ludzi, którzy tak a nie inaczej programują maszyny czy ich używają.

Wyróżniony typ wiarygodności jest weryfikowany w praktyce przez skuteczność działań adresata postawy zaufania. Jest to cecha wpływająca najsil-

⁴ W tej monografii do podobnego podziału odwołują się Polak i Krzanowski; zob. rozdział 3. Por. też [Sztompka, 2007, s. 103–119].

⁵ Do tego byłaby potrzebna wolna wola – cecha co najmniej kontrowersyjna, jeśli idzie o możliwość implementacji w systemach SI.

niej na poziom zaufania: im bardziej efekty działań są zgodne z oczekiwaniami, czyli dają oczekiwany skutek, tym większe zaufanie. Przykładowo, jeśli pewien chirurg, stawił w przeszłości trafne diagnozy i przeprowadził wiele udanych operacji, darzymy jego umiejętności wysokim zaufaniem i bez wahania skierujemy się do niego w potrzebie. Warto zauważyć, że wobec niemożności osiągnięcia przez jakikolwiek podmiot idealnej skuteczności działań, zaufanie do niego musi być silnie uzależnione od jego zdolności do poprawy tejże efektywności. Im dany podmiot jest bardziej otwarty na krytykę, im bardziej jest w stanie eliminować błędy w swoim działaniu i zwiększać w ten sposób swoją skuteczność, tym bardziej jest on godny zaufania.

Relacja zaufania pozostaje zatem zakotwiczona w praktyce, jest też ukierunkowana na przyszłość i stopniowalna. Skoro jest stopniowalna, to powstaje pytanie o dodatkowe czynniki, które istotnie wzmacniają jej siłę⁶.

Twierdzimy, że oprócz kluczowej dla zaufania skuteczności działań (w tym eliminacji błędów zwiększających skuteczność), niezwykle istotną rolę odgrywa tutaj zdolność do wyjaśniania – wyjaśniania motywów działań, stojących za nimi decyzji i przyświecających im celów. Zdolność ta ma zarówno pewne walory praktyczne (o czym powiemy szerzej w kontekście systemów SI), jak i czysto psychologiczne, związane z naturalną dla człowieka potrzebą poszukiwania i uzyskiwania wyjaśnień⁷.

W bliskim psychologii kontekście epistemologicznym kwestia druga wiąże się z samym pojęciem wiedzy, ku której człowiek – jako istota rozumna – w sposób naturalny podąża. W każdej definicji wiedzy podkreśla się czynnik uzasadnienia: jest to zbiór sądów, które oprócz innych własności, muszą posiadać dostatecznie dobre uzasadnienie (por. [Ajdukiewicz, 2006; Chisholm, 1994; Stacewicz, 2021]). To ostatnie zaś, w formie dostępnej dla człowieka czyli intersubiektywnie komunikowalnej, przyjmuje postać wyjaśnienia, które odpowiada na pytanie „dlaczego?”. Na przykład: „Dlaczego lekarz interpretujący wyniki badań wydał sąd określający konkretną diagnozę i konkretną terapię?”. Dostępności tego rodzaju wyjaśnień oczekujemy od każdego eks-

⁶ W kontekście relacji zawodowych czynniki takie identyfikuje H. Bulińska-Stangrecka [2016, s. 301]: „Przyjmuje się, że zaufanie to stan, w którym jednostka godzi się na bycie zależną/bezradną wobec drugiej strony, bazując na przeświadczeniu, że druga strona będzie: 1) kompetentna, 2) otwarta, 3) zaangażowana i 4) będzie można na niej polegać”. Ponieważ obecny tekst jest skoncentrowany na analizie czynnika wyjaśniania (a nie całej hierarchii czynników), tematu tego nie rozwijamy. Por. też. [Mishra, 1996].

⁷ Ostatnią kwestię obrazują dobrze przykłady z życia codziennego. Typowy pacjent bardziej ufa lekarzowi wyjaśniającemu swoje decyzje niż takiemu, który w milczeniu przeprowadza badanie i w milczeniu wypisuje recepty. Typowy uczeń bardziej ceni nauczyciela dobrze tłumaczącego niż przekazującego suchą, ekspercką wiedzę. Dziecko ma większe zaufanie do rodzica, który omawia i wyjaśnia różne kwestie, niż takiego, który tylko wydaje polecenia i ocenia jego zachowanie.

perta i każdego systemu, który wchodzi w jego rolę. Z samej definicji wiedzy jako czegoś intersubiektywnie dostępnego i komunikowalnego [Ajdukiewicz, 2003] wynika zatem, że wiarygodność merytoryczna – zależna od wiedzy i zdolności jej wykorzystywania – musi być weryfikowana między innymi na podstawie wyjaśnień adresata postawy zaufania.

2.3. Zaufanie a wyjaśnianie. Kontekst systemów sztucznej inteligencji

Relacja zaufania człowieka do systemów informatycznych – jako relacja natury psychicznej, której główną stroną jest człowiek – musi mieć swój mentalny wzorzec w relacjach międzyludzkich. Obserwacja ta idzie w parze z ideą testu Turinga, który nakłada na program sztucznej inteligencji sztywne więzy podobieństwa systemów SI do człowieka. Wymóg podobieństwa sprawia, że na zaufanie człowieka do systemu niejako podwójnie wpływają wzorce kształtowane w kontaktach międzyludzkich: po pierwsze dlatego, że podmiotem relacji jest człowiek, po drugie zaś dlatego, że system informatyczny ma przypominać człowieka.

Mając to wszystko na uwadze, uwzględniając przy tym nowy kontekst interesującej nas relacji, możemy uwypuklić ponownie dwa czynniki wzmacniające zaufanie i dodać do nich trzeci. Są to:

- skuteczność – system skutecznie rozwiązuje problemy wchodzące w zakres jego kompetencji (im lepsze statystyki poprawnie realizowanych zadań, tym większe zaufanie użytkownika);
- zdolność do wyjaśniania – system potrafi wyjaśnić, w sposób przekonujący dla człowieka, dlaczego w danej sytuacji podjął taką a nie inną decyzję (jednym z możliwych sposobów wyjaśniania jest zrozumiała dla człowieka rekonstrukcja kroków, które doprowadziły do decyzji)⁸;
- elastyczność i otwartość na krytykę – system może zmienić sposób podejmowania decyzji zależnie od interakcji z użytkownikiem; zwłaszcza w przypadku nieskutecznych działań, błędnych decyzji itp.

⁸ Konkretnie systemy są bardzo zróżnicowane, jeśli chodzi o interpretowalność generowanych przez nie wyników, a idąc dalej, zdolność do wyjaśniania podejmowanych decyzji w terminach zrozumiałych, czyli łatwo-interpretowalnych, przez użytkownika. Zależy to przede wszystkim od używanej w systemach metody reprezentacji wiedzy. W przypadku reprezentacji symbolicznych, takich jak drzewa decyzyjne czy sieci semantyczne, zdolność do wyjaśniania jest wysoka. W przypadku reprezentacji subsymbolicznych, w tym konekcyjnych, używanych w sztucznych sieciach neuronowych, jest ona niska, a generowanie przejrzystych znaczeniowo wyjaśnień wymaga użycia wyrafinowanych algorytmów, których wiarygodność jest wciąż dyskusyjna. Zobacz [Stacewicz, 2021; Zednik, 2019].

Z dwoma ostatnimi czynnikami wiąże się coś, co moglibyśmy nazwać *zasadą ograniczonego zaufania*: użytkownik nie ufa systemowi bezgranicznie, licząc się z możliwością popełniania przezeń błędów. W świetle tak rozumianej zasady zaufanie do systemu jest stopniowalne, a jego poziom musi być tym wyższy, im w większym stopniu system potrafi eliminować błędy i doskonalić swoje działanie. Sprzyjają temu obecne w jego oprogramowaniu moduły uczenia się⁹, ale ponadto, silnie powiązana z mechanizmami naprawczymi zdolność do generowania wyjaśnień.

Stawiamy tezę, że przy akceptacji zasady ograniczonego zaufania istotnie wzrasta rola zdolności do generowania wyjaśnień jako czynnika wzmacniającego wiarygodność systemu. Jeśli użytkownik akceptuje zasadę, a więc jest świadom możliwych błędów w działaniu maszyny, to w sposób naturalny łączy wiarygodność ze zdolnością do poprawy tegoż działania. Chodzi przy tym o perspektywę długoterminową, wykraczającą poza wąski horyzont aktualnie realizowanych zadań. Im większe zdolności naprawcze systemu, tym wyższy poziom zaufania do niego w długoterminowej perspektywie aktualnych i przyszłych zadań.

Mechanizm generowania wyjaśnień jest w tym kontekście niezwykle ważny. Generując wyjaśnienia, system nie tylko zaspokaja pewną epistemiczną potrzebę użytkownika (potrzebę poznania uzasadnień czy wyjaśnień), lecz dostarcza mu silnej przesłanki za tym, że w systemie istnieje pewien mechanizm monitorujący, który prócz generowania wyjaśnień, pozwala rozpoznać i usunąć odpowiednie błędy. Ponadto, informacje zawarte w wyjaśnieniach, zwłaszcza te, które dotyczą wewnętrznych parametrów systemu, mogą pomóc w udoskonaleniu systemu przez człowieka, np. programistę. W kontekście możliwej przebudowy układ taki jest więc bardziej godny zaufania niż układ nie dostarczający wyjaśnień.

Chociaż obydwa powiązane ze sobą czynniki – zdolność do generowania wyjaśnień i zdolność do uczenia się – uznaliśmy za pozytywnie wpływające na zaufanie, to trzeba zwrócić uwagę na inny, bardziej negatywny aspekt ich wzajemnej relacji. Otóż mechanizm uczenia się, a dokładniej zmiany powstające w wyniku jego działania, uznaje się za jedną z głównych przyczyn nieprzejrzystości poznawczej systemów informatycznych oraz trudnej interpretowalności uzyskiwanych przezeń wyników [Zednik, 2021; Stacewicz, Greif, 2021]. Typowy proces narastania nieprzejrzystości polega na tym, że w efekcie uczenia się do systemu są wprowadzane pewne techniczne parametry, które mają na celu dostosowanie działania systemu do wymogów reali-

⁹ Moduły uczenia się są już standardem we współczesnych systemach SI. Co więcej, coraz częściej zdolność do uczenia się jest wskazywana jako cecha definicyjna sztucznej inteligencji. Por. [Woolridge, 2021]. Zobacz też rozdział 1 w tej książce.

zowanego zadania, nie pozwalają natomiast uchwycić istotnej i zrozumiałej dla człowieka relacji między danymi a wynikami. Problem ten pogłębia się, gdy proces uczenia się zależy istotnie od kroków losowych, które (ponownie) mają zapewnić efektywność systemu, nie zaś przejrzystą dla człowieka formę wyjaśnienia. Okazuje się zatem, że chociaż zdolność do uczenia się skutkuje poprawą jakości działania systemu i jako taka wzmacnia relację zaufania, to w przypadku pewnych form uczenia się – takich mianowicie, które prowadzą do nieprzejrzystości poznawczej układu – relacja zaufania zostaje osłabiona.

Narzucającym się rozwiązaniem tego problemu jest powiązanie uczenia się z wyjaśnianiem w taki sposób, aby w obydwu procesy był zaangażowany ten sam mechanizm monitorowania pracy systemu. Z jednej strony, wyniki działania tego mechanizmu, czyli rejestrowane zmiany systemu, byłyby podstawą uczenia się (korzystnych zmian w systemie), a z drugiej strony, byłyby one wykorzystywane w procesie wyjaśniania (odnoszącym się również do stanów wewnętrznych systemu). Do kwestii tej wrócimy w rozdziale ostatnim, postulując konieczne warunki dobrego testu na zaufanie.

2.4. Czy test Turinga jest testem na inteligencję budzącą zaufanie?

Anonsowany we wprowadzeniu test Turinga ma na celu sprawdzenie, czy kierowana odpowiednim oprogramowaniem maszyna przejawia w rozmowie inteligencję dostatecznie podobną do ludzkiej. Przypomnijmy: jeśli kompetentny sędzia, konwersujący z maszyną i człowiekiem na dowolny temat, nie potrafi ich od siebie odróżnić, musi uznać maszynę za inteligentną [Turing, 1950].

Jest to zatem test imitacyjny, co podkreślał sam Turing, nazywając go *grą w naśladownictwo* (ang. *imitation game*). Co kluczowe jednak, wynik testu zależy od umiejętności naśladowania czysto zewnętrznych przejawów inteligencji, a więc generowanych przez maszynę wypowiedzi. Wewnętrzny i techniczny zarazem aspekt działania maszyny jest ignorowany¹⁰.

¹⁰ Warto zwrócić uwagę, że na tę słabość testu kładł nacisk John Searle, gdy formułował swój słynny argument chińskiego pokoju [Searle, 1995]. Zgodnie z jego krytyką, maszyna operująca poprawnie chińskimi wyrażeniami, to znaczy generująca wypowiedzi nieodróżnialne od wypowiedzi Chińczyków, może zupełnie nie rozumieć ich znaczenia, a to – według Searle’a – w sposób oczywisty przeczy jej inteligencji. My jednak, jako osoby oceniające maszynę pod kątem inteligencji, możemy rozpoznać ów brak rozumienia (uzyskać pewność w tym względzie), tylko wówczas, gdy wnikiemy w sposób działania maszyny (nie zaś w jego zewnętrzne przejawy). Test ignorujący wewnętrzne reguły działania nie daje zatem prawdziwej informacji o podobieństwie inteligencji maszyny do inteligencji człowieka (która jest ufundowana na rozumieniu). Por. [Marciszewski, Stacewicz, 2010].

W odniesieniu do faktycznych możliwości maszyny, wynikających z jej wewnętrznej konstrukcji, należy to rozumieć tak, że musi ona, przynajmniej w niektórych sytuacjach, ukrywać swoje pełne możliwości. Znaczy to, że jeśli badany system różni się od człowieka pod pewnymi istotnymi względami – np. szybkością działania, sposobem przetwarzania danych czy precyzją generowanych wyników – to i tak, aby zdać test, powinien imitować zachowania człowieka. Niekiedy zatem, chcąc uchodzić za człowieka, maszyna będzie udzielała odpowiedzi błędnych, niekompletnych lub powolnych¹¹. A jeśli jej program daje możliwość dochodzenia do odpowiedzi poprawnych i szybkich, to *de facto*, w takich sytuacjach, maszyna ukrywa swoje realne możliwości.

Zauważmy, że strategia maksymalizacji podobieństwa do człowieka dotyczy także wyjaśniania. Aby zdać test, maszyna musi dostarczać takich wyjaśnień, jakich oczekuje sędzia – takich zatem, jakich dostarczyłby człowiek, a nie takich, które są zgodne z jej sposobem działania.

Czy maszyna za wszelką cenę naśladowująca człowieka – podszywająca się pod niego i ukrywająca swoje realne możliwości – może budzić zaufanie? Z pewnością nie. Z psychologicznego punktu widzenia takie cechy, jak skłonność do udawania (kogoś lub czegoś innego) czy ukrywania własnych intencji i możliwości, nie sprzyjają budowie zaufania. Wręcz przeciwnie muszą rodzić podejrzenia, że druga strona relacji zaufania nie jest w gruncie rzeczy tym, za kogo się podaje, i może zachowywać się w sposób nieoczekiwany, w tym niebezpieczny.

W kontekście mniej psychologicznym można stwierdzić, że oryginalny test Turinga nie zapewnia człowiekowi (sędziemu) całościowej wiedzy o maszynie, dając mu wgląd w bardzo skąpy wycinek jej możliwości (a ściślej: ich zewnętrznych przejawów). Ów brak wiedzy musi osłabiać zaufanie. Mówiąc inaczej, jeśli podstawą wiarygodności merytorycznej podmiotu zaufania stanowi posiadana przezeń wiedza (zob. pkt 2), to Turingowski sędzia nie ma dostępu do bardzo istotnych składowych tejże.

Z jeszcze szerszego punktu widzenia, abstrahującego w ogóle od specyfiki testu Turinga, trzeba stwierdzić, że sama inteligencja, tak czy inaczej sprawdzana, nie musi iść w parze z wiarygodnością. Inteligencji można bowiem używać w najprzeróżniejszy sposób i dla różnych celów. Jeśli zatem pewien artefakt przejdzie test na inteligencję, nie można go uznać automatycznie za wiarygodny czy godny zaufania. Nie jest bowiem wykluczone, że będzie on używać swojego intelektu w sposób dla nas niekorzystny, czy wręcz niebezpieczny.

Podsumowując ten wątek, trzeba stwierdzić, że zaproponowany przez Turinga werbalny test nierozróżnialności człowieka i maszyny nie jest dobrym

¹¹ Zresztą tak stawia sprawę Turing, omawiając w swoim artykule krótkie przykłady pytań i odpowiedzi testowych [Turing, 1950].

testem na SI budzącą zaufanie. W pierwotnej koncepcji testu tkwi nieusuwalna sprzeczność pomiędzy uczciwym, uwzględniającym realne możliwości maszyny, odpowiadaniem na pytania sędziego a koniecznością udzielania takich odpowiedzi (w tym: wyjaśnień), które są nieodróżnialne od ludzkich. Mimo to, zarówno w oryginalnej propozycji Turinga, jak i w jej krytyce, tkwią pewne wartościowe elementy, które mogą nas przybliżyć do formuły testu uwzględniającego zaufanie do SI.

2.5. Propozycja: warunki brzegowe dobrego testu na godną zaufania sztuczną inteligencję

Choć oryginalny test Turinga nie stanowi dobrego testu na godną zaufania sztuczną inteligencję, to stojąca za nim idea – idea maszyn komunikujących się z człowiekiem w języku naturalnym – wskazuje, naszym zdaniem, dobry kierunek poszukiwań. Komunikacja bowiem, o ile ma zmierzać do zrozumienia, musi obejmować wyjaśnianie; to ostatnie zaś uznaliśmy za czynnik istotnie sprzyjający zaufaniu. Odwołując się do propozycji samego Turinga [Turing, 1950], trzeba stwierdzić, że zdolność maszyn do wyjaśniania podejmowanych decyzji faktycznie była wskazywana przezeń jako coś, co winniśmy weryfikować w trakcie testu. Dowodzą tego przytaczane przezeń przykłady konwersacji, w których pojawiają się pytania o wytłumaczenie, dlaczego wybrano takie, a nie inne, rozwiązanie¹².

Zachowując ideę Turingowskiego testu i traktując zdolność do wyjaśniania jako twardy punkt odniesienia całej procedury, możemy pokusić się o sformułowanie warunków brzegowych testu na godną zaufania SI (w skrócie testu TZ). Ich zestawienie będzie stanowić coś w rodzaju podsumowania sformułowanych wcześniej tez.

Po pierwsze, uwypuklając raz jeszcze naszą główną tezę, istotnym elementem całej procedury powinno być sprawdzanie zdolności maszyny do udzielania wyjaśnień; choć element ten powinien być wkomponowany w ogólniejszą ideę, którą przybliżają kolejne punkty.

¹² Oto kilkudzaniowy przykład, w którym testujący dopytuje, dlaczego „świadek” (możemy go potraktować jako odpowiednik maszyny) zaproponował taki a nie inny początek wiersza: <<Pytający: Czy w pierwszej linii twojego sonetu, która brzmi: „Czy mam porównać cię do letniego dnia” sformułowanie „wiosenny dzień” nie byłoby tak samo dobre lub lepsze? Świadek: Nie byłoby ono do rymu. Pytający: A co myślisz o „dniu zimowym”? To byłoby do rymu. Świadek: Tak, ale nikt nie chce być porównanym do dnia zimowego.>> [Turing, 1950, s. 446] (cytowane za tłumaczeniem polskim D. Gajkowicza, w [Feigenbaum, Feldman, 1972]).

Po drugie, dobry TZ nie powinien być testem czysto imitacyjnym. Maszyna winna nie tyle naśladować zachowania werbalne człowieka, co wzorować się na nich, uwzględniając uczciwie własną odmienność i tożsamość. W szczególności, jeśli jakiś fragment rozmowy testowej byłby ukierunkowany na wyjaśnianie, to kompetentny sędzia powinien zwracać uwagę bardziej na spójność wyjaśnień i ich zgodność ze specyfiką powierzonego maszynie zadania niż na ich podobieństwo do wyjaśnień udzielanych przez ludzi. Przy takim podejściu osoba testująca nie deprecjonowałaby odpowiedzi oryginalnych i kreatywnych (odbiegających od typowych wyjaśnień ludzi), wręcz przeciwnie: doceniałaby je.

Po trzecie, dobry test TZ nie powinien być testem behawioralnym, ignorującym wewnętrzne reguły działania systemu SI. W szczególności, generowane przez system wyjaśnienia powinny być sprawdzane pod kątem zgodności z tymi regułami, które sędzia – w przeciwieństwie do sędziego w oryginalnym teście Turinga – zna. Przykładowo: system mógłby uzasadnić swoją decyzję, powołując się na konkretny, zaimplementowany w nim algorytm i stojącą za nim technikę przetwarzania danych. Gdyby decyzja okazała się błędna – co uznaliśmy wyżej za dopuszczalne i zgodne z zasadą ograniczonego zaufania – system nie utraciłby całkowicie zaufania, bo wskazał algorytm, który dzięki temu wskazaniu mógłby zostać w przyszłości udoskonalony.

Po czwarte, dobry test TZ powinien być ukierunkowany na zdolność do uczenia się, która w dłuższej perspektywie, dzięki możliwości korygowania błędnych lub nieoptymalnych działań systemu, jest czynnikiem wzmacniającym zaufanie¹³. Sprawdzając tę zdolność osoba testująca, nie tylko zadawałaby pytania, lecz przekazywałaby systemowi pewną wiedzę lub przedstawiałaby mu rozwiązania pewnych problemów, a następnie sprawdzałaby, na ile system jest w stanie zmienić swój sposób działania na bardziej efektywny.

BIBLIOGRAFIA

- Ajdukiewicz K., (2003), *Zagadnienia i kierunki filozofii*, Warszawa–Kęty, Wydawnictwo Antyk – Fundacja Aletheia.
- Ajdukiewicz K., (2006), *Zagadnienie uzasadniania*, [w:] *Język i poznanie*, t. 2, s. 374–384, Warszawa, Wydawnictwo Naukowe PWN.
- Bulińska-Stangrecka H., (2016), *Zaufanie w e-kulturze*, *Prace naukowe WSZiP*, 39(3), s. 297–310.

¹³ Zauważmy, że sam Turing podkreślał silny związek zdolności do uczenia się z inteligencją. W tym samym artykule, w którym sformułował reguły testu na inteligencję maszyn, poświęcił cały podrozdział hipotetycznym technikom uczenia się maszyn; wskazując dodatkowo, że tylko maszyny zdolne do uczenia się byłyby w stanie zdać test [Turing, 1950].

- Chisholm R.M., (1994), *Teoria poznania*, Lublin, Instytut Wydawniczy „Daimonion”.
- Feigenbaum E., Feldman J. (red.), (1972), *Maszyny matematyczne i myślenie*, PWN, Warszawa.
- Jaklik A., Łaguna M., (2015), *Zaufanie w organizacji. Analiza sposobów ujęcia i modeli teoretycznych*, *Psychologia Społeczna*, 10, s. 369–382.
- Marciszewski W., Stacewicz P., (2010), *Umysł–Komputer–Świat. O zagadce umysłu z informatycznego punktu widzenia*, Warszawa, Akademicka Oficyna Wydawnicza EXIT.
- Mishra M., (1996), *Organizational responses to crisis: The centrality of trust*, [w:] R. M. Kramer, T. Tyler (red.) *Trust in Organizations*. Newbury Park, Sage, s. 261–287.
- Russell S., Norvig P., (2020), *Artificial Intelligence: A Modern Approach*, Pearson, 4th ed.
- Searle J.R., (1995), *Umysł, mózg i nauka*, Warszawa, PWN.
- Stacewicz P., (2021), *Między wiedzą jak i dlaczego*, wpis i dyskusja w blogu *Cafe Aleph*, adres (dostęp: 1.10.2022).
- Stacewicz P., (2022), *Pojęcia jako funkcje decyzyjne. Zagadnienia filozoficzne, metodologiczne i informatyczne*, Warszawa, Oficyna Wydawnicza Politechniki Warszawskiej.
- Stacewicz P., Greif H., (2021), *Concepts as decision functions. The issue of epistemic opacity of conceptual representations in artificial computing systems*, *Procedia Computer Science*, 192, s. 4120–4127.
- Sztompka P., (2007), *Zaufanie. Fundament społeczeństwa*, Kraków, Znak.
- Szykiewicz M., (2014), *Problem zaufania w kontekście rozwoju społecznego znaczenia technologii informatycznych*, *Filo-Sofija*, 24, s. 259–272.
- Turing A.M., (1950), *Computing Machinery and Intelligence*, *Mind*, New Series, 59(236), s. 433–460.
- Turing A.M., (1948), *Intelligent machinery: A report by A.M. Turing*, National Physical Laboratory, London.
- Woolridge M., (2021), *A Brief History of Artificial Intelligence: What It Is, Where We Are, and Where We Are Going*, Flatiron Books.
- Zednik, C. *Solving the Black Box Problem: A Normative Framework for Explainable*, Philosophy & Technology, 2019, <https://doi.org/10.1007/s13347-019-00382-7>.

Rozdział 3

Ku zaufanej sztucznej inteligencji. Perspektywa fonetyczna Towards Trusted Artificial Intelligence: A Phonetic Perspective

Paweł Polak, Roman Krzanowski

Uniwersytet Papieski Jana Pawła II w Krakowie

Wydział Filozoficzny

Katedra Historii i Filozofii Nauki

Streszczenie. W artykule analizujemy zagadnienie zaufania w systemach społecznej sztucznej inteligencji (SI) i formułujemy koncepcję zaufanej SI, czyli systemów wchodzących w relacje analogiczne do ludzkich relacji zaufania. Postulujemy, że zaufane systemy SI powinny uwzględnić jako ważny paradygmat rozwoju fonetycznej SI, która wydaje się możliwym podłożem do rozwoju systemów społecznych SI w kierunku modelu przewidywalnie dobroczynnych SI (*Provable Beneficial AI*). W ten sposób rozwój SI może zostać skierowany w stronę ogólnie pojętych społecznych korzyści i w stronę głębszych pojęć związanych z ludzką sferą wartości (aksjologią). W artykule również zwracamy uwagę, że terminy takie jak dobroczynna SI (*Beneficial AI*), czy przewidywalnie dobroczynna SI (*Provable Beneficial AI*) są terminami technicznymi i oznaczają pewne ściśle określone rozwiązania techniczne. W związku z tym używanie lub tłumaczenie tych terminów bez tego szczegółowego wyjaśnienia ich znaczenia może prowadzić do przypisywania im znaczeń w nich nie istniejących i niezamierzonych albo do błędnych interpretacji funkcji tych systemów (zamierzonych lub projektowanych), np sugerując, że te systemy mają jakieś głębsze zrozumienie etyki czy są obdarzone własną filozoficzną perspektywą.

Abstract. In the paper we analyze the concept of trust in social artificial intelligence (AI) and we formulate the concept of Beneficial AI, i.e. artificial systems with trust relation analogous to human ones. We also posit that Trusted AI or Beneficial AI systems should include in their design selected aspects of *phronesis*. AI researchers and engineers so far ignored these deeper ethical concepts; but these concepts suitably interpreted may just help in developing more socially acceptable AI systems. The expected

model for the development of such systems is Provable Beneficial AI. We also point out that the terms such as Trusted AI, Beneficial AI or Provable Beneficial AI are strictly technical terms and their use or translation should avoid their philosophical interpretations; such interpretations imply that these systems have some ethical or philosophical dimensions and capacities, which they do not. Translations should point out strictly technical meaning of these terms to avoid the over-interpretation of AI technology capacities and manage public expectations regarding these systems.

Słowa kluczowe: społeczna sztuczna inteligencja, zaufana sztuczna inteligencja, przewidywalnie dobroczynna sztuczna inteligencja, fronetyczna etyka maszyn

Keywords: social AI, Trusted AI, Provable Beneficial AI, thick machine ethics, phronetic ethics

3.1. Wprowadzenie

Zagadnienie zaufania do systemów sztucznej inteligencji (SI) stało się aktualne w ostatnich latach ze względu na szybki rozwój tej gałęzi techniki i daleko idące konsekwencje płynące z wdrażania tych systemów (warto przytoczyć tu rozważania na temat społecznych i politycznych konsekwencji superinteligencji [Bostrom, 2014], konsekwencji egzystencjalnych SI [Etzioni, 2016], analizę sześciu rodzajów ryzyka związanych z SI [Marr, 2018] oraz podobne rozważania Tegmarka [Tegmark, 2016], Hawkinga [Love, 2014], Littmana [Littman i in., 2021] i wielu, wielu innych)¹. Narastające obawy dotyczące

¹ Jest praktycznie niemożliwe wymienienie tutaj wszystkich czy nawet większości publikacji i studiów dotyczących zagrożeń spowodowanych SI z ostatnich kilkunastu lat. Jest to zagadnienie stosunkowo nowe, gdyż publikacje dotyczące SI w pierwszych dziesięcioleciach rozwoju tej techniki były pełne bezkrytycznego entuzjazmu (zob. np. [Minsky, 1967; Moravec, 1988; Rosenblatt, 1962], czy bardziej współcześnie [Kurzweil, 2005]). Pewnym wyjątkiem są tu klasyczne już publikacje Norberta Wienera *Human Use of human beings* [Wiener, 1989] czy prace Johna Searle'a [Searle, 1980; Searle, 1998]. Dobrymi źródłami informacji w kwestii refleksji na temat przyszłości SI i ludzkości są również takie ośrodki badawcze jak Future of Humanity Institute at The Oxford University (<https://www.fhi.ox.ac.uk/>), Future of Live Institute (<https://futureoflife.org/>), czy The Leverhulme Centre for the Future of Intelligence (CFI) within the University of Cambridge (<http://lcfi.ac.uk/>), czy The Centre for the Study of Existential Risk (CSER) at the University of Cambridge (<https://www.cser.ac.uk/>). Badając zagrożenia spowodowane SI, można też sięgnąć do literatury fantastyczno-naukowej, której lepsza część przewidywała od dawna niebezpieczeństwa rozwoju tej technologii. Dobrym przykładem jest tu (wymieniony później w artykule) Stanisław Lem i jego opowieści z cyklu *Bajki Robotów* czy *Cyberjada*.

aktualnego i przyszłego rozwoju SI prowokują więc refleksje na temat zaufania do tych systemów.

W związku z pojawiającymi się nieporozumieniami konieczne jest wprowadzenie pewnych rozróżnień pojęciowych. Na wstępie należy wyróżnić dwa znaczenia słowa ‘zaufanie’. Po pierwsze, mówimy o zaufaniu w odniesieniu do techniki. Ufamy, że samolot przewiezie nas bezpiecznie, podobnie jak ufamy, że w normalnych warunkach most pozwoli nam przedostać się na drugi brzeg rzeki pojazdem o dopuszczalnym tonażu. Jednym słowem – ufamy, że systemy techniczne będą działać zgodnie ze zdefiniowanymi oczekiwaniami i będą bezpieczne w dobrze określonym zakresie.

Po drugie, o zaufaniu mówimy w kontekście relacji międzyludzkich. Na tej kwestii skupimy się w niniejszym opracowaniu rozważając, czy ustalenia dotyczące człowieka i jego społecznego otoczenia mogą być stosowane do pewnej klasy systemów SI².

Obecnie wdrażane systemy SI określane jako „zaufane” są w większości systemami przemysłowymi, dlatego w przypadku tych systemów zaufanie jest interpretowane w kontekście wymagań praktycznych, operacyjnych i technologicznych, bez jakiegokolwiek głębi filozoficznej. Są to więc koncepcje powierzchowne, opisujące zaufanie w pierwszym znaczeniu, które nas tu nie interesują [Williams, 1985; Kirchin, 2013; Väyrynen, 2021]. Przykładem jest tu projekt *IBM Trusted AI*, w ramach którego używa się takich pojęć, jak „sprawiedliwość, solidność, możliwość wyjaśnienia, przejrzystość i odpowiedzialność oraz dostosowanie wartości” [Anon, 2021; Shahrदार i in., 2019; EPRS, 2020; Thiebes i in., 2021; McLeod, 2022]³.

² Zaufanie jest tutaj przekładem angielskiego terminu *trust*. Zaufanie do sztucznej inteligencji dotyczy technologii określanej jako *Trusted AI*. *Trusted AI* powinno się przełożyć jako „sztuczna inteligencja, której można ufać”; tłumaczenie raczej niewygodne w użyciu. Przyjmujemy zatem tłumaczenie *Trusted AI* jako *Zaufana SI*, wiedząc, że określenie to dotyczy specyficznej perspektywy patrzenia na technologię AI, a nie zawiera samo w sobie odrębnej koncepcji filozoficznej.

³ Terminy „sprawiedliwość”, „solidność”, „możliwość wyjaśnienia”, „przejrzystość”, „odpowiedzialność” oraz „dostosowanie wartości” użyte w tym kontekście są terminami technicznymi odnoszącymi się do wymagań stawianych systemom sztucznej inteligencji i jako takie w literaturze angielskiej są określane jako *fairness*, *robustness*, *explainability*, *transparency*, *accountability* i *value alignment*. Uwaga ta jest istotna, gdyż należy unikać w tej dyskusji i dyskusji o SI interpretacji powyższych terminów poza kontekstem SI. Znaczenia te są określone ściśle zdefiniowanymi technicznymi wymaganiami (*requirements*) i różnią się od ich znaczeń popularnych czy filozoficznych. W dalszej części artykułu techniczne i bardziej konwencjonalne znaczenia terminów są rozróżniane jako interpretacje *thin* i *thick* (powierzchnowe i głębokie).

3.2. Zaufanie w kontekście SI – pojęcia powierzchowne a głębokie

W celu lepszego zrozumienia znaczenia pojęcia zaufania w kontekście SI należy przywołać znany z etyki podział na koncepcje powierzchowne i głębokie (*thin vs. thick concepts*) [Väyrynen, 2021]. Pojęcia powierzchowne (*thin*) można traktować jako związane głównie z etyką utylitarystyczną. Jest to podejście pragmatyczne, nastawione na działanie, brak samoświadomości, brak odpowiedzialności moralnej, wymierność. Podejście takie pasuje do metodologii nauk technicznych. Brak głębi moralnej powoduje, że podejście to może być podstawą relacji zaufania tylko w kontekście niezawodności techniki. Powierzchowne pojęcia etyczne są relatywnie łatwe do sformalizowania, są więc łatwe do implementacji w systemach SI. Oczywiście sformalizowanie tych pojęć wiąże się z ich redukcyjną interpretacją.

Odmienne rzecz się przedstawia w stosunku do pojęć głębokich (*thick*), które odnoszą się do etyki deontologicznej lub do etyki cnót. Pojęcia głębokie opierają się na normatywności, na wartościach, na pojęciu ‘ja’, na pojęciu odpowiedzialności i kształtowaniu wewnętrznej doskonałości moralnej podmiotu. Etyczne pojęcia głębokie wydają się również odległe od metodologii nauk technicznych, tzn. ich implementacja, koncepcyjnie i praktycznie stanowi duże wyzwanie.

W niniejszej pracy pokażemy, jak na przykładzie rozwijanej perspektywy fronetycznej SI, po niezbędnych adaptacjach, można w kontekście relacji zaufania wprowadzić pewne aspekty głębokich, klasycznych koncepcji filozoficznych do systemów SI.

3.3. Zaufanie do SI – perspektywa fronetyczna

Fronetyczna SI jest próbą oparcia podstaw pojęciowych etyki sztucznych systemów robotycznych na pojęciach głębokich, zatem jest próbą wyjścia poza krąg silnie upraszczających, redukcyjnych rozwiązań w zakresie etyki maszyn. Model fronetycznej SI jest próbą adaptacji koncepcji etycznych Arystotelesa do systemów robotycznych [Polak, Krzanowski, 2020b; Polak, Krzanowski, 2020a]. Fronetyczna robotyka jest ściśle związana z fundamentalnym etycznym pojęciem *phronesis* (zob. np. [Aristoteles, 1999; Kraut, 2022]).

Należy zaznaczyć, że zaufanie w interesującym nas znaczeniu realizuje się poprzez osądy i oparte na nich decyzje. Osądy rozumiemy tutaj jako podmiotowe oceny problemów i rozwiązań. Tak rozumiane osądy wymagają zasad

oceny. Natomiast pojęcie decyzji w kontekście sztucznego systemu oznacza wybór rozwiązania z przestrzeni problemów. Oczywiście jest, że taki wybór wymaga również metody przeszukiwania przestrzeni problemów.

Kluczowym aspektem społecznych robotów, które mają wchodzić w relacje zaufania, jest zdolność tworzenia sądów praktycznych (np. [Coeckelbergh, 2009; Leite i in., 2013; Campa, 2016; Polak, Krzanowski, 2020b]). Możliwość tworzenia sądów praktycznych, dokonywanych z perspektywy sztucznego agenta bazować musi na rozpoznawaniu wartości i na pewnych formach empatii, pełniących istotną rolę w relacjach międzyludzkich. Bez nich bowiem nie jest możliwe nawiązanie relacji zaufania pomiędzy podmiotami, szczególnie jeśli jeden z nich jest sztucznym tworem.

Dynamika relacji zaufania wymaga również od zaangażowanych stron dążenia do udoskonalania swego zachowania. Innymi słowy, autonomia robota nie powinna ograniczać się wyłącznie do zdolności do działania bez nadzoru człowieka. Powinna także obejmować zdolność do niezależnego doskonalenia swojej zdolności do autonomicznego działania (zob. np. [Polak, Krzanowski, 2020a]). Dlatego, jednym z najważniejszych zadań stojących przed projektantami systemów SI wydaje się pogłębienie refleksji nad kwestią autonomii.

Z perspektywy zagadnienia zaufania wyróżnić można trzy znaczenia pojęcia autonomii:

- autonomia działania,
- autonomia wyznaczania celów,
- autonomia samodoskonalenia.

Z tych trzech obszarów znaczeniowych tylko pierwszy był dotychczas przedmiotem szerszej refleksji filozoficznej w kontekście systemów SI (przykładowo problem autonomii w systemach SI jest dyskutowany jako ogólnie zdefiniowana autonomia [Parker, 2006], w kontekście działania samochodów autonomicznych [Iwan, 2019], w kontekście broni autonomicznych [ICRC, 2019] lub jako ogólna autonomia systemów SI [Totschnig, 2020]).

3.4. Zaufanie w kontekście dobroczynnej SI (Beneficial AI)

Próbując określić celowość poszukiwania (trudnego!) sposobów skonstruowania systemów SI, które implementują głębokie koncepcje etyczne i mogą być podstawą typowo ludzkich relacji zaufania, należy sformułować pewien docelowy, wyidealizowany model SI. Taki model służyłby jako punkt odniesienia w ocenach systemów SI, wskazywałby również pożądaną kierunek przyszłego rozwoju SI.

Dotychczas uwaga skupiała się raczej na negatywnych określeniach dotyczących kierunków rozwoju. Najszerzej dyskutowanymi w literaturze związanymi tego typu są doskonale znane prawa Asimova [Asimov, 1950]. Prawa te jednak przy bliższej analizie okazują się wewnątrznie sprzeczne i praktycznie niewykonalne (zob. np. [Singer, 2010; Anderson, 2011; Anon, 2019]). Dlatego należy poszukiwać innych rozwiązań, takich jak, na przykład, zaproponowana koncepcja *phronesis*.

Interesujący nas model określimy jako dobroczynną SI (ang. *Beneficial AI*). Opieramy się tutaj na podstawowych, słownikowych znaczeniach pojęcia dobroczynności: 1. «przynoszący korzyść lub ulgę»; 2. «niosący pomoc potrzebującym»⁴. Podane tu znaczenia słownikowe zbieżne są ze znaczeniami angielskiego słowa *beneficent*: *doing or producing good (e.g., a beneficent policy); performing acts of kindness and charity (e.g., a beneficent leader)*⁵.

Należy jednak zauważyć, że angielski termin *Beneficial AI* jest terminem technicznym, to znaczy, że ma specyficzne znaczenie w literaturze przedmiotu. W związku z brakiem dokładnego odpowiednika tego terminu w języku polskim przekłady na język polski będą zawsze interpretacją i mogą nie odpowiadać w pełni oryginalnemu znaczeniu. Powinien towarzyszyć temu terminowi opis uściślający znaczenie lub odnośniki do źródeł. Na przykład *Beneficial AI* bywa definiowana jako: *AI technology that is safe and beneficial for society* [Baum, 2017]⁶. Taka definicja jednak niewiele wyjaśnia i wymaga uściślenia jak na przykład w ujęciu zaprezentowanym poniżej.

SI powinna być budowana tak, aby przynieść pozytywny bilans korzyści całemu społeczeństwu lub, w wypadku trudnej do przewidzenia sytuacji – pozytywny bilans korzyści. To znaczy, że SI nie powinna być budowana tylko, aby ulepszyć technologię lub uczynić ją bardziej interesującą. Równocześnie SI nie powinna być budowana w celu zapewnienia korzyści tylko jej konstruktorom, jeżeli miałyby się to odbywać kosztem całego społeczeństwa [Bohannon & Russell, 2015]⁷.

Należy dodać, że kluczowe są tutaj uściślenia pojęć: *safe, beneficial, for society* zawarte w cytowanym tekście.

⁴ Słownik Języka Polskiego PWN, <https://sjp.pwn.pl/slowniki/dobroczynno%C5%9B%C4%87.html>.

⁵ Merriam Webster dictionary, <https://www.merriam-webster.com/dictionary/beneficent>.

⁶ S.J. Russell zawiązuje pojęcie dobroczynnej SI do SI dobroczynnej dla ludzi i zakłada przy tym, że SI dobroczynna dla innych gatunków będzie miało inne cele.

⁷ *AI should be built so as to have net benefits for the whole of society – or, in the face of uncertainty, net expected benefits. This is to say that AI should not be built just for the sake of making it more capable or more intellectually interesting. Also, AI should not be built for the benefit of its builders if this comes at the expense of society as a whole.*

Podobnie, termin *provable beneficial AI* jest terminem technicznym w tym samym znaczeniu jak termin *beneficial AI*. *Provable beneficial AI system* jest definiowany jako system, w przypadku którego, niezależnie od przyjętego systemu wartości, korzyści jakie dają ludziom rozwiązania problemów podejmowanych przez ten system są większe niż wtedy, gdyby system ten nie został użyty⁸. Jest to definicja techniczna (operacyjna) obwarowana wieloma ograniczeniami. Na przykład: czy systemy SI mogą się nauczyć od człowieka uniwersalnych systemów wartości (*reward functions*), albo czy funkcje celu tych systemów (*utility functions*) będą reprezentować wartości ważne dla człowieka [Russell, 2019]. Tak jak w poprzednim przypadku, definicja korzystnych czy dobroczynnych systemów SI weryfikowanych w praktyce (*provable beneficial AI*), jest definicją techniczną o bardzo wąskim i specjalistycznym znaczeniu⁹.

Fronetyczna SI wydaje się obiecującym kierunkiem zmierzającym do stworzenia dobroczynnej SI, o ile odpowiednio zostanie ukształtowana w sztucznym systemie koncepcja eudajmonii i o ile będzie się mógł on skutecznie komunikować z otoczeniem społecznym. Taki fronetyczny system dobroczynnej SI, któremu można zaufać (*Beneficial AI*), należy do węższej klasy systemów, które możemy określić jako weryfikowalną dobroczynną SI (ang. *Provable Beneficial AI* – PBAI).

Systemy PBAI powinny spełniać co najmniej cztery poniższe warunki:

- 1) ich decyzje prowadzą do osiągnięcia naszych celów [Russell, 2019], gdzie cele są określane jako dobroczynne, oparte na współczuciu, humanitarne;
- 2) są wyposażone w funkcję wyłączania (która daje możliwość ich zatrzymania w przypadku błędnego działania z przyczyn technicznych);
- 3) nie mają wewnętrznych celów ustalonych *a priori*; cele są otrzymywane z zewnątrz od beneficjentów danego systemu (wewnętrzne cele tych systemów są celami ich użytkowników);
- 4) są weryfikowane w rzeczywistym działaniu; ich funkcje są realizowane w świecie rzeczywistym (a nie w drodze analiz teoretycznych, dowodów logicznych, czy przez analizę sztucznie dobranych przykładów)¹⁰.

Należy tutaj rozwiać pewną podstawową wątpliwość. Wprowadzenie sztucznych agentów nie pozwala na osiągnięcie absolutnego i jednocześnie

⁸ ...the expected reward for the human is higher when the robot is available, regardless of what the human's actual reward function is [Russell, 2019].

⁹ Model wskazujący, w jaki sposób można zrealizować dobroczynne systemy SI przedstawiliśmy w pracy: P. Polak, R. Krzanowski, *How to Tame Artificial Intelligence (AI)? A Symbiotic AI Model for Beneficial AI* [w druku].

¹⁰ W przypadku takich wymagań szeroka dyskusja wokół wyidealizowanych eksperymentów myślowych (jak np. dylemat wagonika) okazuje się nieporozumieniem oraz stratą czasu.

uzasadnionego zaufania odnośnie sztucznego systemu. Wynika to stąd, że żaden podmiot (naturalny czy sztuczny) nie może w dowolnej sytuacji zapewnić, że podejmie właściwą decyzję. Każda decyzja ma różne horyzonty sukcesu i część tych horyzontów odkrywa się dopiero po pewnym czasie od podjęcia decyzji.

Można wprawdzie porównywać stopnie zaufania w pewnych klasach problemów, tzn. zaufanie może być w rozsądny sposób stopniowalne. Trzeba jednak mieć na uwadze, że ze względu na nieskończoną potencjalnie liczbę możliwych sytuacji (konfiguracji) żaden skończony system nie jest w stanie zrealizować przepisu na 100% poprawną reakcję w takim ‘środowisku’ – ludzkie tęsknoty, aby stworzyć sztuczny system usuwający z naszej egzystencji ten obszar niepewności okazują się po prostu mżonką.

Tak jak w przypadku ludzkiego umysłu często zmuszeni jesteśmy posługiwać się upraszczającymi heurystykami, to podobnie musi być w przypadku systemu sztucznego. W przeciwieństwie do człowieka, u którego część heurystyk jest wyrazem ewolucyjnego dziedzictwa i są trudne (o ile możliwe) do zmiany, o tyle w przypadku systemu sztucznego można stosunkowo łatwo modyfikować heurystyki, osiągając dla pewnej klasy zagadnień znacząco lepsze wyniki. Wysiłki kierowane na stworzenie takich sztucznych systemów nie są więc nieuzasadnione.

3.5. Podsumowanie

W artykule postulujemy, że zaufane systemy SI powinny uwzględnić jako ważny paradygmat rozwoju etykę fronetyczną, która wydaje się możliwym podłożem do realizacji dobroczynnych systemów PBAI. W ten sposób rozwój SI może zostać skierowany w stronę społecznych korzyści ogólnie pojętych (PBAI) i w stronę głębokich pojęć związanych z ludzką sferą wartości (aksjologią). W duchu etyki fronetycznej systemy SI powinny uwzględnić czynniki dotychczas marginalizowane, takie jak ogólne korzyści dla społeczeństwa (ale nie w uproszczonym pojęciu utilitaryzmu) i jednostek ludzkich, niezależnie od kultury czy innych uwarunkowań społecznych.

Formułując wyidealizowany model PBAI, musimy przypomnieć o ontologicznej przepaści między systemem naturalnym (człowiek) a systemem sztucznym. Przepaść ta powoduje, te same pojęcia, które są używane w kontekście systemów naturalnych, użyte w kontekście systemów SI nabierają innego znaczenia. Jest to dobrze widoczne w przypadku zastosowań pojęć etycznych w systemach sztucznej inteligencji. Nie można więc mówić, że

sztuczny system ucieleśnia etykę Arystotelesa, gdyż byłoby to zbyt daleko idącym uproszczeniem, generującym niepotrzebne nieporozumienia.

Zauważmy również inne rysujące się wyzwania przed dobroczynną AI: musi ona w jakiś sposób przekroczyć problem przepaści między naturą pojęć akceptowanych w technice (inżynierii) oraz w filozofii. W technice preferowane są pojęcia operacyjne, mierzalne (ilościowe), a nawet sformalizowane. W filozofii natomiast rozważamy zwykle pojęcia abstrakcyjne, opisowe, rozumiane często intuicyjnie. Rozwój dobroczynnej SI prowadzić musi do zbliżenia się tych dwóch sfer.

Z pewnością inżynieria musi nauczyć się wiele od humanistyki, odnośnie do tego, jak można operować na trudnych do zoperacjonizowania koncepcjach. Filozofia praktyczna natomiast musi podążać w kierunku uściślenia i operacjonalizowania pojęć (filozofia Arystotelesa i wybrane elementy podejścia analitycznego wydają się tu szczególnie obiecujące). Nie jest to zadanie zupełnie nowe, gdyż pierwsze próby zbliżenia tak rozumianej techniki oraz filozofii dokonują się już od kilkudziesięciu lat (zob. np. [Konieczny, 1983]).

Z pewnością zagadnienie PBAI może stać się ważnym czynnikiem służącym zbliżeniu metod inżynierii i filozofii. W ten sposób otwiera się nowa sfera relacji między filozofią a techniką – obszar *filozofii w technice*. Warto nadmienić, że w ten sposób technika dochodzi do etapu realizacji kolejnych wizjonerskich pomysłów Stanisława Lema na temat roli filozofii dla techniki [Krzyszowski, Polak, 2021]. W interesującym nas przypadku filozofia dostarcza bowiem istotnych narzędzi i metod, pozwalając na rozwój tak wyrafinowanych zastosowań jak zaufana dobroczynna SI.

BIBLIOGRAFIA

- Anderson S.L., (2011), *The Unacceptability of Asimov's Three Laws of Robotics as a Basis for Machine Ethics*, [w:] *Machine Ethics*, red. M. Anderson & S. L. Anderson, Cambridge, Cambridge University Press, s. 285–296, <https://www.cambridge.org/core/books/machine-ethics/unacceptability-of-asimovs-three-laws-of-robotics-as-a-basis-for-machine-ethics/D58C8BAD402DF52AD2785C17A68431EB> (dostęp: 5.07.2022).
- Anonimowo, (2020), *The Future of AI – 5, 10, 50 Years Into the Future*, Exigent, <https://www.exigent-group.com/blog/the-future-of-ai-5-10-50-years-into-the-future/> (dostęp: 5.07.2022).
- Anonimowo, (2019), *The Three Laws of Robotics Have Failed the Robots*, Mind Matters, <https://mindmatters.ai/2019/09/the-three-laws-of-robotics-have-failed-the-robots/> (dostęp: 5.07.2022).
- Anonimowo, (2021), *Trustworthy AI*, IBM Research, <https://research.ibm.com/topics/trustworthy-ai#about-us> (dostęp: 5.07.2022).
- Aristoteles, (1999), *Nicomachean ethics*, Indianapolis Ind., Hackett Publ. Co.
- Asimov I., (1950), *I, robot*, New York, Gnome Press.

- Baum S.D., (2017), *On the promotion of safe and socially beneficial artificial intelligence*, AI & SOCIETY, 32(4), s. 543–551, <http://link.springer.com/10.1007/s00146-016-0677-0> (dostęp: 5.07.2022).
- Bohannon J., Russell S., (2015), *Fears of an AI pioneer*, Science, 349(6245), s. 252, <https://www.science.org/doi/10.1126/science.349.6245.252> (dostęp: 5.07.2022).
- Bostrom N., (2014), *Superintelligence: paths, dangers, strategies*, Oxford, Oxford University Press.
- Campa R., (2016), *The Rise of Social Robots: A Review of the Recent Literature*, Journal of Ethics and Emerging Technologies, 26(1), s. 106–113, <http://jeet.iiet.org/index.php/home/article/view/55> (dostęp: 5.07.2022).
- Coeckelbergh M., (2009), *Personal Robots, Appearance, and Human Good: A Methodological Reflection on Roboethics*, International Journal of Social Robotics, 1(3), s. 217–221, <http://link.springer.com/10.1007/s12369-009-0026-2> (dostęp: 5.07.2022).
- EPRS, (2020), *The ethics of artificial intelligence: Issues and initiatives*, [https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU\(2020\)634452](https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2020)634452) (dostęp: 5.07.2022).
- Etzioni O., (2016), *Are the Experts Worried About the Existential Risk of Artificial Intelligence? Yes*, MIT Technology Review, <https://www.technologyreview.com/s/602776/are-the-experts-worried-about-the-existential-risk-of-artificial-intelligence-yes/> (dostęp: 5.07.2022).
- ICRC, (2019), *Autonomy, artificial intelligence and robotics: Technical aspects of human control*, https://www.icrc.org/en/download/file/102852/autonomy_artificial_intelligence_and_robotics.pdf (dostęp: 5.07.2022).
- Iwan D., (2019), *Autonomous Vehicles – a New Challenge to Human Rights?*, Przegląd Prawniczy Uniwersytetu im. Adama Mickiewicza, <https://pressto.amu.edu.pl/index.php/ppuam/article/view/21648> (dostęp: 5.07.2022).
- Kirchin S., (2013), *Introduction: Thick and Thin Concepts*, [w:] Thick Concepts, red. S. Kirchin, Oxford, Oxford University Press, s. 1–19, <https://oxford.universitypressscholarship.com/view/10.1093/acprof:oso/9780199672349.001.0001/acprof-9780199672349-chapter-1> (dostęp: 5.07.2022).
- Konieczny J., (1983), *Inżynieria systemów działania*, Warszawa, Wydawnictwa Naukowo-Techniczne.
- Kraut R., (2022), *Aristotle's Ethics*, [w:] Stanford Encyclopedia of Philosophy, red. E. N. Zalta, Stanford, Calif., <https://plato.stanford.edu/archives/sum2022/entries/aristotle-ethics/> (dostęp: 5.07.2022).
- Krzanowski R., Polak P., (2021), *The Future of AI: Stanisław Lem's Philosophical Visions for AI and Cyber-Societies in Cyberiad*, Pro-Fil, 22(3), s. 39–53, <https://www.phil.muni.cz/journals/index.php/profil/article/view/2405>.
- Kurzweil R., (2005), *The Singularity Is Near: When Humans Transcend Biology*, New York, Viking.
- Leite I., Martinho C., Paiva A., (2013), *Social Robots for Long-Term Interaction: A Survey*, International Journal of Social Robotics, 5(2), s. 291–308, <http://link.springer.com/10.1007/s12369-013-0178-y> (dostęp: 8.10.2019).
- Littman M.L. i in., (2021), *Gathering Strength, Gathering Storms: The One Hundred Year Study on Artificial Intelligence (AI100) 2021 Study Panel Report*, One Hundred Year Study on Artificial Intelligence (AI100), <https://ai100.stanford.edu/gathering-strength-gathering-storms-on-e-hundred-year-study-artificial-intelligence-ai100-2021-study> (dostęp: 5.07.2022).
- Love D., (2014), *Stephen Hawking Is Worried About Artificial Intelligence Wiping Out Humanity*, Business Insider, <https://www.businessinsider.com/stephen-hawking-on-artificial-intelligence-2014-5> (dostęp: 5.07.2022).
- Marr B., (2021), *Future Developments of Artificial Intelligence*, <https://bernardmarr.com/future-developments-of-artificial-intelligence/> (dostęp: 5.07.2022).

- Marr B., (2018), *Is Artificial Intelligence Dangerous? 6 AI Risks Everyone Should Know About*, Forbes, <https://www.forbes.com/sites/bernardmarr/2018/11/19/is-artificial-intelligence-dangerous-6-ai-risks-everyone-should-know-about/>, (dostęp: 5.07.2022).
- McLeod C., (2022), *Trust*, [w:] Stanford Encyclopedia of Philosophy, red. E. N. Zalta, Stanford, Calif., <https://plato.stanford.edu/archives/fall2020/entries/trust/> (dostęp: 5.07.2022).
- Minsky M., (1967), *Computation: finite and infinite machines*, Englewood Cliffs, NJ, Prentice-Hall.
- Moravec H., (1988), *Mind children: the future of robot and human intelligence*, Cambridge, Mass.; London, Harvard University Press.
- Parker L.E., (2006), *Why Autonomous Robotics and Artificial Intelligence?*, Journal of the Robotics Society of Japan, 24(5), s. 582–584.
- Polak P., Krzanowski R., (2020a), *Ethics in autonomous robots as philosophy in silico: The study case of phronetic machine ethics*, Logos i Ethos, 52, s. 33–48.
- Polak P., Krzanowski R., (2020b), *Phronetic Ethics in Social Robotics: A New Approach to Building Ethical Robots*, Studies in Logic, Grammar and Rhetoric, 63(1), s. 165–183, <https://content.sciendo.com/view/journals/slgr/63/1/article-p165.xml> (dostęp: 9.11.2020).
- Rosenblatt F., (1962), *Principles of neurodynamics: perceptrons and the theory of brain mechanisms*, Washington, Spartan Books.
- Russell S.J., (2019), *Human compatible: artificial intelligence and the problem of control*, New York, Viking.
- Searle J.R., (1998), *Mind, Language and Society: Philosophy in the Real World*, New York, Basic Books.
- Searle J.R., (1980), *Minds, brains, and programs*, Behavioral and Brain Sciences, 3(3), s. 417–424, https://www.cambridge.org/core/product/identifier/S0140525X00005756/type/journal_article (dostęp: 5.07.2022).
- Shahrdar S., Menezes L. & Nojournian M., (2019), *A Survey on Trust in Autonomous Systems*, [w:] Intelligent Computing, Advances in Intelligent Systems and Computing, red. K. Arai, S. Kapoor, & R. Bhatia, Cham, Springer International Publishing, s. 368–386, http://link.springer.com/10.1007/978-3-030-01177-2_27 (dostęp: 5.07.2022).
- Singer P.W., (2010), *Isaac Asimov's Laws of Robotics Are Wrong*, Brookings, <https://www.brookings.edu/opinions/isaac-asimovs-laws-of-robotics-are-wrong/> (dostęp: 5.07.2022).
- Smith B.C., (2019), *The Promise of Artificial Intelligence: Reckoning and Judgment*, Cambridge, MA, The MIT Press.
- Tegmark M., (2016), *Benefits & Risks of Artificial Intelligence*, Future of Life Institute, <https://futureoflife.org/background/benefits-risks-of-artificial-intelligence/> (dostęp: 5.07.2022).
- Thiebes S., Lins S., Sunyaev, A., (2021), *Trustworthy artificial intelligence*, Electronic Markets, 31(2), s. 447–464, <https://link.springer.com/10.1007/s12525-020-00441-4> (dostęp: 5.07.2022).
- Totschnig W., (2020), *Fully Autonomous AI*, Science and Engineering Ethics, 26(5), s. 2473–2485, <https://link.springer.com/10.1007/s11948-020-00243-z> (dostęp: 5.07.2022).
- Väyrynen P., (2021), *Thick Ethical Concepts*, [w:] The Stanford Encyclopedia of Philosophy, red. E. N. Zalta, Metaphysics Research Lab, Stanford University, <https://plato.stanford.edu/archives/spr2021/entries/thick-ethical-concepts/> (dostęp: 5.07.2022).
- Wiener N., (1989), *The Human Use of Human Beings: Cybernetics and Society*, London, Free Association Books.
- Williams B.A.O., (1985), *Ethics and the limits of philosophy*, Cambridge, Harvard University Press.
- Wooldridge M., (2021), *The Road to Conscious Machines: The Story of AI*, London, UK, Penguin.

Rozdział 4

Behawioralne uwarunkowania decyzji podejmowanych przy wykorzystaniu sztucznej inteligencji

Behavioral determinants of the decision-making process using artificial intelligence

Agnieszka Tomczak

Politechnika Warszawska

Wydział Administracji i Nauk Społecznych

Streszczenie. Celem jest rozważenie w kontekście decyzji o charakterze ekonomicznym zalet i wad zastosowania sztucznej inteligencji (SI) w procesie podejmowania decyzji, w którym SI pełniłaby rolę wspierającą. Należy się jednak liczyć z tym, że błędy w myśleniu popełniane przez ludzi mogą być również powielane w programowaniu i działaniu SI. Człowiek posługujący się SI w swoich decyzjach ukierunkowanych na poprawę dobrostanu może ulegać efektom behawioralnym przy interpretacji wyników działania SI. Tezą jest stwierdzenie, że zastosowanie SI w procesie podejmowania decyzji ekonomicznych racjonalizuje je dzięki możliwości przetworzenia wielkiej liczby danych i zastosowania metod automatycznego uczenia się. Nie wyklucza to jednak wad tego rozwiązania, jak wpływ preferencji, doświadczeń i cech osobowości programisty oraz błędów poznawczych człowieka przy interpretacji wskazań podawanych w praktyce przez SI. Ponadto wykluczenie z decyzji emocji ludzkich oraz zasobów wiedzy pozostających u ludzi w ich nieświadomości (ale wykorzystywanych w praktyce), a także bieżących interakcji danej osoby z innymi ludźmi, może prowadzić do podejmowania decyzji nieoptymalnych z punktu widzenia dobrostanu człowieka.

Abstract. The aim of this thesis is to consider the advantages and disadvantages of using artificial intelligence (AI) in the decision-making process in which AI would play a supporting role. It should be taken into account, however, that errors in human thinking can also be replicated in the programming and operation of AI. Humans using AI in their decisions aimed at improving well-being may succumb to behavioral effects when interpreting the results of AI actions. The thesis argues that the use of AI in economic decision-making process rationalizes it through the ability to process large amounts of data and through machine learning. However, this

does not exclude the disadvantages of this solution, such as the influence of preferences, experiences and personality traits of the programmer as well as cognitive errors of humans when interpreting the indications given in practice by the AI. In addition, the exclusion of human emotions and subconscious knowledge (that is used in practice) from the decision-making process, as well as a person's ongoing interactions with other people, may lead to suboptimal decisions from the perspective of human well-being.

Słowa kluczowe: sztuczna inteligencja, ekonomia behawioralna, racjonalność decyzji

Keywords: artificial intelligence, behavioral economics, rational decision making

4.1. Wprowadzenie

Prawa ekonomii głównego nurtu zakładają pełny dostęp do informacji przez wszystkich agentów, podejmowanie bieżących decyzji zgodnie z zasadą racjonalności ekonomicznej oraz bazują na hipotezie racjonalnych oczekiwań. Cała wiedza istotna dla podjęcia decyzji powinna być udostępniona wszystkim zainteresowanym agentom, a nieracjonalność jest wynikiem tego, że część tej wiedzy pozostaje niedostępna, a jeśli nawet można ją pozyskać, to może pozostać niewykorzystana z powodu kosztów poszukiwania i przetwarzania informacji lub błędnie odczytywana i interpretowana z powodu zniekształceń poznawczych. Podejmowane decyzje mogą być nieoptymalne pod wpływem kontekstu (okoliczności) oraz emocji. Ważnym czynnikiem jest horyzont czasowy decyzji, przy czym przy jego wydłużaniu rośnie niepewność. Przykładem typowego rozumowania ekonomicznego opartego na racjonalności jest argumentacja Bastiat [Bastiat, 2014] który opisuje skutki krótko- i długoterminowe decyzji ekonomicznych, takich jak podatki, zaciąganie długu przez państwo czy zastępowanie pracy ludzkiej maszynami. Nawet jeśli agent posiada wszystkie potrzebne informacje do racjonalnego podjęcia decyzji w danej chwili i prawidłowo ich użyje, to nie można mieć pewności, czy wystarczająco dobrze rozpoznał i wziął pod uwagę także skutki długoterminowe. Nie jest poprawnym założeniem, że błędy agentów zawsze kompensują się w masie, gdyż mogą mieć charakter systematyczny. Powodem jest także to, że dostępne dane nie są używane przez agentów, którzy powinni być nimi zainteresowani, co wynika z różnych ograniczeń. Przykładem może być „mit racjonalnego łotra”: Gary Becker założył, że ludzie popełniający przestępstwa racjonalnie kalkulują, czy im się opłaca czy nie, popełnienie czynu, za który

grożą kary więzienia [Zyzik, 2018], czyli dokładnie znają przepisy prawa karnego i mają rozeznanie co do prawnej analizy popełnianych czynów, co wydaje się mało prawdopodobne. Jest to jedna z przyczyn, dla której pozytywne efekty zaostrenia kar za przestępstwa są niewielkie albo nieistotne. Podobnie jest w dietetyce. Umieszczanie informacji o kaloryczności produktów spożywczych na etykietach nie oznacza, że informacje te są odczytywane i brane pod uwagę w decyzjach zakupowych, gdyż oznacza to wysiłek – poświęcenie uwagi i czasu, zmianę nawyków. Nie znaczy to, że agenci są leniwi, tylko mają ograniczone zasoby, więc nie „przestawiają” się na myślenie głębokie w każdym przypadku i nie używają go w równym stopniu we wszystkich decyzjach. Ludzie mają ograniczone możliwości pozyskiwania i analizowania informacji oraz dokonywania nieustannych kalkulacji w celu podjęcia racjonalnych decyzji ekonomicznych, wobec czego SI może stanowić cenne wsparcie w tym procesie.

4.2. niesprawności rynku a sztuczna inteligencja

Decyzje są ostatecznie podejmowane przez ludzi, a więc agent mający dostęp do przydatnej informacji podanej przez SI nie musi z niej skorzystać, jak też ludzie ze swoimi ograniczeniami poznawczymi programują SI, a także udostępniają (lub nie) określone dane, które SI przetwarza, co może być przyczyną zniekształceń racjonalności, a więc niepełnej optymalizacji decyzji, czy wręcz podjęcia decyzji błędnej.

Efektywność informacyjna zdecentralizowanych procesów (takich, jak regulacja rynkowa za pomocą mechanizmu cenowo-kosztowego) jest jednym z głównych problemów intelektualnych ekonomii, a asymetria informacji jest źródłem zniekształceń cen oraz określonej renty realizowanej przez jedną ze stron transakcji. Przypadki szczególne, w których rynki nie działają sprawnie, nie dają możliwości sformułowania ogólnych twierdzeń ani modeli dotyczących niesprawności rynku [Smith, 2013, s. 121]. Ponadto rynek jako regulator procesów produkcyjnych działa nieśpiesznie, a okres dostosowań może potrwać miesiące, czy nawet lata. Powstaje pytanie, czy zastosowanie SI pozwala wyeliminować te niesprawności i opóźnienia. Skuteczność zależy od przedmiotu obrotu i liczby zarejestrowanych danych. Z uwagi na cel analizy można podzielić przedmioty handlu na instrumenty finansowe (tu można wyróżnić rynki płynne i mało płynne), dobra i usługi materialne (masowe, zróżnicowane i spersonalizowane), w tym prywatne i publiczne, oraz dobra cyfrowe (w tym konsumpcyjne, przeznaczone dla podmiotów biznesowych i sektora publicznego).

Odnośnie do instrumentów finansowych w płynnym obrocie, szeroko stosowany jest handel algorytmiczny (handel wysokiej częstotliwości realizowany przez maszyny), co nie oznacza, że na takim rynku nie przydarzają się krachy; przykład to *Flash Crash* z 9 maja 2010 r., spowodowany „nieważnością założeń” [Tegmark, 2019, s. 129]. Marwala i Hurwitz [2015] stwierdzają, że asymetria informacji charakterystyczna dla ludzi i napędzająca handel instrumentami finansowymi zostanie pokonana, jeśli na danym rynku wszystkie transakcje będą przeprowadzane przez maszyny, sterowane przez SI. Handel znacznie się zmniejszy, ponieważ asymetria informacji stanowiąca źródło renty nie będzie występować [Marwala i Hurwitz, 2015]. Ten tok rozumowania jest przekonujący, ale idąc dalej, co się stanie, gdy wszyscy agenci na danym rynku będą działać superracjonalnie i mieć te same pełne informacje? Taki rynek nie jest interesujący dla inwestorów, bo nie można na nim osiągnąć zysku nadzwyczajnego (renty ekonomicznej) dzięki błędom innych uczestników. Być może jako ludzkość do tego dążymy, ponieważ ukróciłoby to spekulację, a ceny byłyby bliższe rzeczywistej wartości. Ale błędy w wycenie mogą być pozorne. Ten sam przedmiot obrotu może być dla poszczególnych agentów wart mniej lub więcej z powodu ich osobistych preferencji i celów, przy czym zaawansowane formy SI będą w stanie je rozpoznać na podstawie (innych) rejestrowanych danych, jakie generują ci agenci. Wartości aktywów finansowych (np. akcji) nie są zależne tylko od sytuacji emitenta i ogłaszanych przez niego planowanych projektów, ani tylko od bieżącej sytuacji rynkowej, ale kształtują się pod wpływem cen inwestycji alternatywnych, a także ogólnej koniunktury giełdowej i jej perspektyw, nie wspominając już o zmiennych makroekonomicznych (które SI zapewne weźmie pod uwagę) i o czarnych łabędziach [Taleb, 2020]. Działa tu także zwrotność – ścieżka ceny wyłania się ze zwrotnej interakcji między poglądami uczestników i faktycznym stanem rzeczy; reprezentowana jest przez model samowznoszącej się i samowygaszającej się fali [Soros, 2008, s. 141]. Czynniki te nadają cenom wysoką zmienność i sprzyjają bańkom spekulacyjnym. Taleb pisze też o efekcie przetrwania: *można analizować tylko tych, którzy przetrwali i tylko ich uwzględnić w danych* [Gilder, 2013, s. 271]. SI w długim okresie przeanalizuje dane dotyczące tylko tych, którzy „wygrali”, ignorując tych, którzy „przegrali” i do wygrywających dostosuje rozwiązania.

Wśród aktywów emitentów papierów wartościowych z branż nowych technologii (najściślej związanych z SI) sprzęt komputerowy i oprogramowanie stanowi dużą część wartości aktywów ogółem. Oprócz ich wartości księgowej inwestorzy wyceniają efektywne wykorzystanie tych aktywów, a może ono się znacznie różnić w poszczególnych przedsiębiorstwach. Obliczenie wartości aktywów niematerialnych jest trudne, a ich replikacja kosz-

towna [Brynjolfsson, Rock, Syverson, 2017, s. 34], ponieważ wymaga reorganizacji i podniesienia umiejętności personelu. Wiedza, jak konstruować zasoby cyfrowe, nie generuje automatycznie kapitału cyfrowego [Tambe, Hitt, Rock, Brynjolfsson, 2020, s. 26]. Pomiary wartości tych aktywów i ich wpływu na wyniki firmy wymagają aktualizacji narzędzi. Inwestorzy na rynkach akcji wierzą, że istnieje taka wartość wyrażona bieżącą wartością zdyskontowanych przepływów pieniężnych, skorygowanych o ryzyko, gdyż firmy działające z efektywnym wykorzystaniem SI generują wyższe stopy zwrotu z kapitału, co przekłada się na wyższe ceny akcji i wyższe wskaźniki cena/wartość księgową. Ten efekt związany jest z zanikiem kosztów krańcowych przy wytwarzaniu kopii cyfrowych produktów. Dzięki temu firmy działające na bazie SI w porównaniu z generującymi podobne wolumeny zysków spółkami działającymi w branżach tradycyjnych mogą zatrudniać kilkakrotnie mniej pracowników, a na giełdzie są wyceniane wielokrotnie wyżej. Efekt zbliżających się do zera kosztów krańcowych jest jedną z przyczyn pogłębiających się różnic dochodowych [Tegmark, 2019, s. 160–161]. Drugą jest nieproporcjonalne wynagradzanie supergwiazd [Frank, Cook, 2017], a trzecią globalizacja, pozwalająca na zwielokrotnione efekty skali.

Na rynkach towarowych (jak również na rynkach tradycyjnych usług) do różnic w cenach przyczynia się heterogeniczność produktów, ich postrzeganie przez nabywców i zróżnicowane preferencje klientów co do ich używania (częstotliwość, czas, miejsce). Analizy zachowań klientów oparte na osiągnięciach ekonomii behawioralnej oraz na SI pozwalają firmom na czerpanie z tego zysków. Przynosi to pewne korzyści klientom dzięki „podsuwaniu” im preferowanych przez nich (czy podobnych) produktów, dzięki czemu mogą oni oszczędzić czas i wysiłek przy dokonywaniu zakupów, ale prawdopodobnie zapłacą nieco więcej (spersonalizowaną cenę, dopłatę za opcje dodatkowe). Jest dyskusyjne, dlaczego różne ceny za podobne produkty są uczciwe – wymaga to wglądu w myślenie i preferencje klientów; jeśli klienci nie robią zakupów sprytnie czy postępują zgodnie z przyzwyczajeniem, to mogą zapłacić więcej niż wynosi niższa cena, akceptowana przez dostawcę, zaś doświadczeni konsumenci przejmujący algorytmy personalizacji mogą oszczędzić pieniądze albo otrzymać więcej korzyści za daną kwotę [Camerer, 2017].

Dobra cyfrowe również podlegają powyższym zasadom, ale jak już wspomniano, mają dodatkowe zalety, jak generowanie ich na żądanie po dodatkowych kosztach bliskich zeru oraz możliwość niezwłocznej dostawy drogą elektroniczną. Łatwiej również je dopasować do preferencji klienta przez „podsuwanie” wersji produktu i uzasadnianie, dlaczego akurat ta odmiana może być dla niego interesująca. Tu SI ma niewątpliwie znaczną przewagę nad tradycyjnymi metodami sprzedaży.

Jak zastosowanie sztucznej inteligencji może przyczynić się do poprawy jakości decyzji ekonomicznych i w jaki sposób mogą na ten proces wpłynąć osiągnięcia ekonomii behawioralnej? Jest to osiągalne na dwa sposoby. Pierwszy dotyczy możliwości gromadzenia i przetwarzania w krótkim czasie ogromnej liczby danych oraz ich korelacyjnego przetwarzania. Uczenie maszynowe (ang. *Machine Learning*, w skrócie ML) skraca czas potrzebny do porządkowania i przeliczenia danych. Umożliwia też ponowne użycie danych służących do podjęcia poprzednich decyzji i bardziej staranne rozważenie ich rezultatów. Drugi sposób to „zwiększenie dokładności przewidywań” [Thaler, 2018, s. 27] dzięki eliminacji błędów percepcji, których SI raczej nie popełnia. Występuje tu zgodność celu ekonomii behawioralnej i SI – dokładniejsze prognozowanie i precyzyjne adresowanie. Ponadto dzięki informacjom od SI można zwiększyć skuteczność oddziaływania odgórnych decydentów na indywidualne decyzje. Ma to szerokie zastosowanie w celu zwiększenia dobrostanu społeczeństwa, przez używanie w ochronie zdrowia, edukacji, obronności czy dla poprawy bezpieczeństwa ruchu drogowego, w finansach osobistych oraz do współuczestnictwa w życiu publicznym i rządzeniu. Na podstawie uzyskanych dzięki SI informacji o adresatach można dopasować „szturchania”, „zachęty” lub „podsuwania” w celu uzyskania pożądanych pozytywnych zachowań. Indywidualizacja „szturchnięć” i ewentualne ich powtarzanie następuje w wyniku zastosowania SI. Obserwując osiągnięcia firm, które stosują tego typu metody w marketingu, można wnioskować, że decydenci w sektorze publicznym powinni być nimi istotnie zainteresowani, a metody te nie muszą przybierać postaci „kija i marchewki”, lecz opierać się na inteligentnie dobranych argumentach i działać bez stosowania przymusu.

Camerer stwierdził, że ludzkie przewidywania można traktować jak niedoskonałe uczenie się maszynowe i w związku z tym: *ważne jest, aby zbadać, w jaki sposób technologia SI stosowana w firmach i innych instytucjach może zarówno przewyższać, jak i wykorzystywać ludzkie ograniczenia* [Camerer, 2017, s. 1]. Przytacza wyniki uzyskane przez innych badaczy: w jednym z badań pewność siebie uczestników badania rosła wraz z większą ilością materiału na ten sam temat (i w związku z tym mogła powodować częstsze błędy z tego wynikające), zaś inne badanie wykazało, że mała liczba zmiennych może być zaskakująco dobrze dopasowana (czyli nie chodzi o jak największą liczbę różnych danych, tylko o wychwycenie danych istotnych dla rozwiązania konkretnego problemu). Większość informacji odbieranych przez osobę podejmującą decyzję (np. rekrutera przyjmującego kandydatów do pracy) nie ma istotnej wartości, czyli stanowi zakłócenie procesu decyzyjnego. Jednak ludzie nie lubią odrzucać informacji, skoro już je mają, więc posługują się nimi w swoich decyzjach, nawet gdy są one bez znaczenia dla celu, co znie-

kształci racjonalność wyboru (np. upodobanie lub rezerwa rekrutera w odniesieniu do niektórych cech kandydatki). Maszyna zaprogramowana tak, aby selekcjonowała i przeliczała tylko istotne informacje, nie popełni tego typu błędów. W przypadku rozpoznawania kwestii przez człowieka, w reakcji na błąd niepotrzebnie może być dodana kolejna funkcja, która decydentowi wydaje się istotna, a ponadto może być źle sformułowana i przyczyni się do błędów w decyzjach [Camerer, 2017, s. 11].

Połączenie SI z ekonomią behawioralną może pomóc w określeniu rozwiązań, jak postępować w odniesieniu do adresatów polityki publicznej w sposób spersonalizowany, nie różnicując indywidualnie polityki (co spowodowałoby ogromne koszty), aby osiągać w jak najlepszym stopniu jej cele [Hrnijs, Tomczak, 2019]. Zastosowany model uwzględniałby unikalne okoliczności i kontekst danej osoby dla sformułowania odpowiednio spersonalizowanej odpowiedzi. Takie interwencje mogą zniwelować uprzedzenia związane z podejmowaniem decyzji przez poszczególnych agentów, a więc odpowiednio do nich trafić (np. przez użycie przekonujących dla danej grupy społecznej czy zawodowej argumentów lub języka). Połączenie ML z ekonomią behawioralną, o ile sformułowane modele są dobrze uogólnione, umożliwi eksperymenty niezależne od podstawowego mechanizmu wyboru; „precyzyjne podsuwanie” może stać się technologią ogólnego przeznaczenia w rozwiązywaniu wielu problemów [Hrnijs, Tomczak, 2019, s. 5–6]. Warunkiem jest cyfryzacja rejestrów publicznych i powszechne gromadzenie danych.

Reasumując, zastosowanie SI powinno zwiększyć efektywność wszystkich rynków. Efekty to niższy koszt podejmowania decyzji, krótszy czas, precyzyjniej dobrane, zrozumiałe dla adresatów komunikaty, wyższa skuteczność w osiągnięciu celów. Główny problem to zgoda adresatów na zbieranie i przetwarzanie danych, zaś koszty to inwestycje w sprzęt komputerowy i oprogramowanie, kwalifikacje pracowników i użytkowników oraz koszty zmian reorganizacyjnych.

4.3. Sztuczna inteligencja a poprawa racjonalności decyzji ekonomicznych

W wielu opiniach wyrażane jest zaniepokojenie faktem, że nie jest rozpoznany sposób, w jaki działa SI i stanowi to przyczynę ograniczonego zaufania do tej formy inteligencji. Jednak w przypadku inteligencji ludzkiej też nie jest to mechanizm do końca rozpoznany: *układy wbudowane w mózg wykonują swoją pracę automatycznie i w dużej mierze poza naszą świadomością* [Smith, 2013, s. 25]; *mózg jest zdolny do autonomicznego, podświadomego uczenia*

się [Smith, 2013, s. 26]. Ludzie używają posiadanej wiedzy bez świadomości, jakimi całkowitymi zasobami wiedzy i umiejętności dysponują. Z jednej strony, mają miejsce pomyłki w kalkulacjach i zniekształcenia poznawcze, z drugiej – ludzie wiedzą o wiele więcej, niż mają tego świadomość, ponadto istnieje wiedza zbiorowa, z której korzystają, nie zdając sobie z tego sprawy. Posługujemy się pojęciami, symbolami, praktykami, instytucjami, nawykami, które zostały sprawdzone i stanowią fundamenty cywilizacji. Te zasoby są również używane przez twórców SI. Ponadto, ludzie stosują (świadomie lub nie) określone modele myślowe (wyobrażenia), czyli dobudowują pasujący (ale nie poparty konkretnymi informacjami) ich zdaniem kontekst do danych czy zjawisk, które współtworzą problem decyzyjny. Są też w stanie świadomie zmienić kontekst, szukając rozwiązania problemu, i to właśnie jest źródłem dostosowania się ludzi do zmiany warunków otoczenia, a także zaprojektowania w nim pożądaných zmian. Zjawisko to, jako „myślenie kontekstowe” czy „nadawanie ram mentalnych” przedstawiają Cukier, Mayer-Schönberger i de Vericourt [2021] i formułują tezę, że ten rodzaj myślenia stanowi przewagę człowieka nad SI (zdaniem autorów, SI tego nie potrafi).

Czy wszystkie poczynania dotyczące decyzji ekonomicznych muszą być rozpoznane, opisane i świadome, aby były racjonalne? Pobieźna lektura prac z zakresu ekonomii behawioralnej sugeruje takie podejście. Czy tak jest w rzeczywistości? Zależy to od liczby doświadczeń, czyli uczenia się, przy czym jest to zarówno proces indywidualny, jak i społeczny. SI również kumuluje doświadczenia jako dane i uczy się maszynowo na podstawie danych indywidualnych oraz zbiorowych, ale celowo, w sposób ukierunkowany na określony obszar dóbr danych i rozpoznanie panujących w nich zależności korelacyjnych. Jak cytuje Hayek za Whiteheadem (w innym kontekście): *Jest to głęboko błędny truizm, powtarzany we wszystkich artykułach i przez wybitnych ludzi, gdy wygłaszają przemówienia, że powinniśmy kultywować nawyk myślenia o tym, co robimy. Cywilizacja rozwija się poprzez zwiększanie liczby ważnych operacji, które możemy wykonać bez zastanowienia się nad nimi* [Hayek, 1945]. Zastanawianie się, kalkulowanie pochłaniania zasoby, ludzie stosują więc metody skrócone, np. analogię, rutynę, intuicję czy heurystykę. Jeśli do procesów poznawczych i kalkulacyjnych dodamy SI, niedokładności obliczeniowe czy błędy behawioralne zanikną lub zostaną w większości wyeliminowane. Używając SI, dążymy do wykonywania pożądaných działań bez konieczności tłumaczenia podmiotom, co mają robić; możemy rozwiązywać problemy bez ich zrozumienia. Oczywiście może to budzić niepokój i niedobór zaufania.

Z drugiej strony wskazuje się zagrożenia z powodu SI (dotychczas działającej w wersji „słabej”, podporządkowanej człowiekowi): możliwość manipulacji, transformacja demokracji, ataki technologiczne, nieprzewidywalność,

utrata kontroli nad urządzeniami, odpowiedzialność za skutki działania SI, ograniczenia w sferze prywatności, obawy o charakterze emocjonalnym czy sceptycyzm kładący nacisk na ułomności SI [Torczyńska, 2019, s. 108–111]. Nie sposób jednak przecenić zalet SI w kontekście jej przydatności w życiu codziennym: Internet rzeczy, profilaktyka zdrowotna i medycyna, opieka nad seniorami, autonomiczny transport, rozrywka i inne obszary [Torczyńska, 2019, s. 113–116]. Wymieniane są dziedziny, w których SI może dać wielkie możliwości, zanim osiągnie ludzki poziom, jak: *szukanie błędów, prawodawstwo, uzbrojenie i zatrudnienie* [Tegmark, 2018, s. 110].

Poniżej postawiono kilka pytań wraz z podjęciem prób odpowiedzi z punktu widzenia ekonomii behawioralnej.

- **Czy SI jest efektywna w sensie ekonomicznym i czy jest przedsiębiorcza?**

Podstawową kwestią w zastosowaniu i rozwoju SI są koszty inteligentnych maszyn i energii, którą zużywają. Na razie koszty są wysokie, co sprawia, że rozwój SI jest dość powolny. Mamy coraz większą liczbę zarejestrowanych danych, ale nie jest jasne, co z nimi sensownie zrobić, a przede wszystkim co ma być celem finalnym ludzkości i jaką ścieżkę rozwoju SI w dążeniu do tego celu wybrać. Niekoniecznie chodzi tu o efektywność w sensie ekonomicznym, chociaż na bieżącym etapie jest to główne kryterium: (...) *nie uzyskamy eksplozji inteligencji dopóty, dopóki koszty wykonywania pracy wymagającej inteligencji na poziomie człowieka nie spadną poniżej stawek godzinowych dla ludzi* [Tegmark, 2019, s. 207]. Proces uzyskiwania efektywności wiąże się z etapami rozwoju życia na Ziemi: Tegmark wymienia: życie 1.0 – etap biologiczny (ewolucję darwinowską), życie 2.0 – etap kulturalny (organizm projektuje znaczną część swojego oprogramowania – uczenie się, nauka, sztuka, instytucje) i życie 3.0 – etap technologiczny (projektuje postać materialną i oprogramowanie) – *silną inteligencję*, która może wykonać każde zadanie [Tegmark, 2019, s. 46–47]. Poprawa efektywności ekonomicznej może służyć stworzeniu SI, która będzie najefektywniejszą maszyną do zaspokajania potrzeb ludzkości (jeśli taki wariant rozwoju SI ludzkość wybierze), ale należy wcześniej określić, jaki jest preferowany końcowy rezultat, gdyż bez tego SI może zejść na manowce [Bostrom, 2021, s. 184–188].

SI nie podejmuje ryzyka, które jest nieodłączną cechą przedsiębiorczości, tylko przetwarza dane, które są jej surowcem. Jeśli pojawi się innowacja, to SI nadal bazuje (w większości) na danych z poprzednich okresów i nie wskaże, przynajmniej od razu, nowej metody jako odpowiedniej do zastosowania. Zdaniem Gildera, zarówno podejście oparte na losowym błędzeniu, jak też przyjęcie, że procesem uczenia się kieruje algorytm – ból lub przyjemność, zysk lub strata: *na monotonnej, hedonicznej drodze do władzy* [Gilder, 2013,

s. 156], jest niesłuszne w odniesieniu do przedsiębiorczości. Przedsiębiorca kieruje się bowiem informacją, która jest zaskoczeniem i jest nieprzewidywalna. Jeśli jest ona skodyfikowana, to dla przedsiębiorcy zanika jej wartość: *Gdy się przechodzi od informacji pisanej do danych zorganizowanych matematycznie (...) doświadcza się katastrofy utraconego uczenia się* [Gilder, 2013, s. 373]. Jeśli informacje są skodyfikowane, często są już przestarzałe. *Wiedza pewna (...) jest nam dana dopiero wówczas, gdy minie moment możliwości*. Nie można poznać przyszłych zdarzeń, kształtowanych przez miliardy ludzi, których sposobu działania nie znamy. W celu zwiększenia racjonalności decyzji, planowane jest wykorzystanie SI, która jest w stanie dokonać obliczeń współbieżnych i rozproszonych [Filipkowski, 2018, s. 270]. Jednak należy liczyć się z efektywnością systemów wspomagających decyzje, w tym kosztów i czasu. Celem zastosowania SI jest obniżenie ryzyka popełnienia błędów poznawczych i wykluczenie błędnej kalkulacji – w tym sensie poprawia ona efektywność. Natomiast na obecnym etapie rozwoju SI intelekt ludzki ma nad nią nieporównywalną przewagę: *Wobec ciągłej zmienności i niepewności sytuacyjnej najwyższa elastyczność i adaptacyjność stanowią zasadniczą siłę ludzkiego intelektu, której konkurencyjności nieprędko sprosta sztuczna inteligencja* [Ficoń, 2013, s. 80].

- **Czy SI jest nieomylna?**

Jest bardziej dokładna niż człowiek i niepomierne szybciej liczy. Natomiast nie jest nieomylna i te błędy mogą mieć podłoże behawioralne, gdyż wskazania zależą od zamysłu programistów i ustawionych przez nich założeń i progów oraz od charakteru i kompletności danych. Dane mogą być gromadzone tendencyjnie w takim sensie, że zbiera się je np. w kilku dużych ośrodkach, w których preferowane są określone zachowania czy metody postępowania, wskazane przez miejscowych liderów, obciążonych (jak wszyscy) błędami poznawczymi; te zniekształcone tendencyjnym doбором dane służą do podjęcia decyzji, które mogą być obarczone wynikającym z tego błędem. Prawdopodobnie powierzenie zbierania danych samej SI jest lepszym rozwiązaniem, ale nie jest ona jeszcze na tym etapie rozwoju. Należy sądzić, że wraz ze wzrostem mocy obliczeniowych i poprawą zaufania do SI otrzymane od niej produkty będą coraz lepsze i dokładniejsze.

- **Czy pozyskiwanie danych prywatnych celem dostarczania surowca dla SI jest warte korzyści z SI?**

Odpowiedź jest twierdząca, ale wymaga to nowych rozwiązań prawnych w celu ochrony nie tylko samych danych, ale też przed dyskryminacją osób, od których dane te są pobierane. Pobieranie i wykorzystywanie danych naru-

sza poczucie bezpieczeństwa osób, od których one pochodzą. Wyrażane są liczne obawy nie tyle o nadużycia ze strony samej SI jako myślącej maszyny, co o niewłaściwe zaprogramowanie, źle postawione cele, nieodpowiednią selekcję danych, przyjęte błędne założenia oraz użycie myślących maszyn do nieetycznych celów przez niektórych ludzi. Należałoby zablokować takie możliwości, a SI zaprojektować tak, aby działała szlachetnie, dla dobra ludzkości. Zanim ludzkość będzie do tego zdolna, może dochodzić do nadużyć czy dramatycznych niespodzianek przy zastosowaniu SI i stąd przekonanie, że nie trzeba przyspieszać procesu rozwoju SI do postaci *silnej* SI, zanim ludzkość nie upora się z odpowiednim określeniem celów finalnych oraz umiejętnym nakierowaniem SI na te cele. Po uruchomieniu *silnej* SI zmiany mogą być tak gwałtowne, że ludzkość może tego nie przeżyć [Bostrom, 2020, s. 101–104].

- **Czy eliminowanie emocji dzięki zastosowaniu SI jest dobre?**

W obliczeniach eliminowanie emocji jest zaletą. W użyteczności dla człowieka często nie jest to preferowane, ponieważ emocje są wielką wartością w ludzkim życiu. Na przykład przedsiębiorczość jest postrzegana jako przygoda, praca jako zaangażowanie, działania opiekuńcze bazują na pozytywnych emocjach, warto wspomnieć też o miłości, empatii i altruizmie. Negatywne uczucia, jak zazdrość, mogą mieć również strony pozytywne (motywacja). Wyeliminowanie emocji nie zawsze jest zgodne z preferencjami agenta, bo może zmniejszyć satysfakcję. Należy więc przy projektowaniu SI eliminować tylko niektóre emocje i tylko w takich czynnościach, przy których jest to pożądane.

Problem jednak w tym, że na świadome decyzje człowieka wpływa myślenie skojarzeniowe (asocjacyjne), które odbywa się poza świadomością, a więc nasza wiedza o nas samych nie jest tożsama z wiedzą uświadomioną. Preferencje ludzkie nie są do końca wykalkulowane, lecz oparte również na zasobach nieuświadomionej wiedzy i na emocjach, wywołanych przez mimowolne skojarzenia. Jak pisze Tegmark: *jesteśmy lojalni tylko wobec własnych uczuć* [2019, s. 329]. Zachowanie ludzi nie jest już zoptymalizowane na przetrwanie gatunku i nie ma dobrze zdefiniowanego celu [Tegmark, 2019, s. 330].

Wyobrażenie o SI jest zdominowane przez skojarzenia ze stereotypowymi cechami ludzi, którzy świetnie radzą sobie z komputerami. Tak więc sądzi się, że superinteligentna maszyna będzie podobna do osoby *o bardzo wysokim IQ, lecz słabo rozwiniętych kompetencjach społecznych, czyli nieco zdziwaczka i wyobcowana, bez empatii i zmysłu politycznego; oczekujemy, że zaawansowane formy SI będą miały te cechy jeszcze bardziej rozwinięte niż współcześni profesjonaliści od komputerów i programowania* [Bostrom, s. 140–141]. Nie należy tego lekceważyć, gdyż mogą tu tkwić zarodki błędów behawioralnych czynionych przez ludzi techniki, którzy mają swoje specyficzne „ustawienia”

behawioralne, tak jak je mają ekonomiści (nawykowo przeliczający różne opcje, aby zoptymalizować relację efekty/nakłady i uważający za oczywiste, że wszyscy wokół robią tak samo).

- **Czy decyzje podejmowane przez ludzi, z którymi kontaktuje się SI, są lepsze?**

Jak stwierdza Camerer [2017], jeśli drugą stroną jest gracz automatyczny – podejmowane są lepsze decyzje; ale jeśli człowiek nie wie, że z drugiej strony działa gracz automatyczny, to czy można dać wiarę wynikom badań? Z kolei sztuczne systemy skojarzeniowe dają więcej wiedzy i pozwalają na interpretowanie pragmatyki ludzkich działań w procesach gospodarczych, politycznych oraz w zarządzaniu [Maciąg, 2020, s. 153]. W życiu poszczególnych osób, przy rzadko podejmowanych decyzjach nie ma możliwości nauczyć się je podejmować optymalnie przez doświadczenie. Nie ma jednak pewności, czy decyzje oparte na danych z przeszłości okażą się najlepsze w czasie teraźniejszym lub przyszłym, którego dotyczą. Stąd doświadczenia (dane), jakie gromadzi SI i dopasowanie produktu do klienta stanowią niezaprzeczalne zalety, o ile klientowi pozostawi się trochę wolnej woli.

- **Czy SI jest moralna, altruistyczna i przyjazna człowiekowi?**

Jak pisze Horzyk [2013, s. 265] system SI powinien być nakierowany na potrzeby ludzkie, a także *być odporny na manipulację i kłamstwo* oraz pełnić rolę *mediatora, słuchacza i doradcy, nie popierając ani nie wspomagając negatywnych ani egocentrycznych ludzkich zachowań*. W tym celu kluczowy czynnik to *ludzie właściwego rodzaju, którzy będą projektować SI* [Bostrom, 2021, s. 373]. Wracamy więc do określenia celu SI oraz cech osób, które powinny projektować SI tak, aby działała dla dobra ludzkości (i co to znaczy w praktyce). Zaufanie do SI jest zależne od przyjęcia właściwych celów i doboru osób, które mają ją programować.

- **Jak SI wpłynie na zatrudnienie?**

Bezrobocie technologiczne jest wskazywane jako jedna z głównych wad SI. Jeśli SI będzie potrafiła wykonać każdą pracę, co stanie się z sensem życia ludzi? Rynek pracy zostanie w większości zautomatyzowany. Z punktu widzenia klasycznej ekonomii, krótkookresowa funkcja produkcji może przestać istnieć (jest oparta na czynniku zmiennym produkcji, a więc na zatrudnieniu). Powstaną nowe problemy podziału coraz większego bogactwa. Wymaga to odpowiedniej polityki państwa w obszarze edukacji, jednak jak cytuje Tegmark za McAfee’em: *zasady ekonomii są jasne, ale się ich nie przestrzega*

[Tegmark, 2018, s. 164]; uczestnicy rynku pracy nie są odpowiednio przygotowani do wymagań gospodarki cyfrowej, a sektor publiczny nie nadaża z dostosowaniem narzędzi polityki cyfrowej za sektorem prywatnym.

Ludzie, którzy nie reprezentują odpowiednich umiejętności i kompetencji do funkcjonowania w gospodarce cyfrowej, czują się przegrani życiowo i nieufni wobec SI. Nie bardzo można te narzędzia użytkować, nie posiadając kompetencji. Podstawowa kwestia to przydatność ludzi – jeśli SI będzie potrafiła wykonać wszystko, pracownicy nie będą potrzebni, a ludzie nie będą mogli twierdzić, że są wyzyskiwani. Wartość dodana uzyskana z pracy maszyn będzie niewyobrażalnie duża. Pozostaje pytanie, jak rozdzielić dobra (których tak czy inaczej wystarczy dla wszystkich) i co będziemy wtedy robili (już dziś pojawia się ten problem, pod postacią rosnącego populizmu w krajach wysokorozwiniętych, w których ludzie mają szeroki dostęp do świadczeń od państwa). Jako rozwiązanie wskazywane są zawody kreatywne i nauka, ale to też pułapka, gdyż dysponując zaawansowanymi formami SI wystarczy wyrazić życzenie i maszyna dostarczy niezwłocznie zamówione dzieło, jak w bajce Lema o Elektrybąlcie [Lem, 1978, s. 206–210]. Podobnie naukowcy mogą być niepotrzebni, gdyż *superinteligencja* będzie niewspółmiernie lepszym badaczem od człowieka. Wracamy więc do pytania o sens życia i finalny cel ludzkości.

4.4. Wnioski

SI pomaga pokonać ograniczenia poznawcze człowieka, które zakłócają proces podejmowania racjonalnych decyzji ekonomicznych. Należy jednak rozważyć wpływ nieumyślnej eliminacji z analizy części istotnych danych, jak również możliwość pojawienia się niepożądanych skutków, wynikających z behawioralnych uwarunkowań mających wpływ na osoby, które projektują SI, a także wykorzystują ją w polityce publicznej i w biznesie. Problem braku zaufania do SI wiąże się ze skutkami błędów przy tworzeniu SI oraz pozyskiwaniu danych, na których SI pracuje, a także z niewłaściwymi założeniami (celami) użycia SI.

BIBLIOGRAFIA

- Bastiat F., (2015), *Co widać i czego nie widać*, Prohibita.
Bostrom N., (2021), *Superinteligencja. Scenariusze, strategie, zagrożenia*, Gliwice, Onepress, Helion S.A.

- Brynjolfsson D., Chad R., Syverson E., (2017), *Artificial Intelligence and the modern productivity paradox: a clash of expectation and statistics*, NBER Working Paper 24001, National Bureau of Economic Research, November 2017.
- Camerer C.F., (2017), *Artificial Intelligence and behavioral economics*, Prepared by NBR Conference, 13–14.2017, online (dostęp 15.06.2021), <http://governance40.com/wpcontent/uploads/2018/12/Camerer.pdf>.
- Cukier K., Mayer-Schönberger V., de Vericourt F., (2022) *Myślenie kontekstowe: największa przewaga ludzi nad sztuczną inteligencją*, MT Biznes.
- Ficoń K., (2013), *Sztuczna inteligencja nie tylko dla humanistów*, Warszawa, BEL Studio.
- Filipkowski P., (2018), *Racjonalność inteligentnego agenta*, Roczniki Kolegium Analiz Ekonomicznych, Szkoła Główna Handlowa, 2018/ nr 49, [w:] Społeczno-ekonomiczne aspekty rozwoju gospodarki cyfrowej: koncepcje zarządzania i bezpieczeństwa, s. 261–272.
- Frank R.H., Cook P.A., (2017), *Społeczeństwo, w którym zwycięzca bierze wszystko: dlaczego garstka najbogatszych posiada o wiele więcej niż reszta z nas*, Toruń, Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika w Toruniu.
- Gilder G., (2013), *Wiedza i władza*, Warszawa, Zysk i S-ka Wydawnictwo.
- Horzyk A., (2013), *Sztuczne systemy skojarzeniowe i asocjacyjna sztuczna inteligencja*, Warszawa, Akademicka Oficyna Wydawnicza EXIT.
- Hrnjic E., Tomczak N., (2019), *Machine learning and behavioral economics for personalized choice architecture*, arXiv:1907.02100v1[econ GN] 3 Jul 2019.
- Jastrzębska W., Jastrzębska A., (2010), *Metody sztucznej inteligencji w rozwiązywaniu problemów mikro- i makroekonomicznych*, Nierówności społeczne a Wzrost Gospodarczy, 17, s. 172–183.
- Kahneman D., (2012), *Pułapki myślenia. O myśleniu szybkim i wolnym*, Media Rodzina.
- Lem S., (1978), *Cyberiada*, Kraków, Wydawnictwo Literackie Kraków.
- Maciąg R., (2020), *Transformacja cyfrowa. Opowieść o wiedzy*, Kraków, Universitas.
- Marwala T., Hurwitz E., (2015), *Artificial Intelligence and Asymmetric Information Theory*, https://www.researchgate.net/publication/282906709_Artificial_Intelligence_and_Asymmetric_Information_Theory (dostęp 20.08.2021).
- Smith V., (2013), *Racjonalność w ekonomii*, Oficyna Wolters Kluwer business.
- Soros G., (2008), *Nowy paradygmat rynków finansowych*, MT Biznes.
- Tambe P., Hitt L., Rock D. Brynjolfsson D., (2020), *Digital capital and superstar firms*, NBER Working Paper 28285, National Bureau of Economic Research.
- Tegmark M., (2019), *Życie 3.0. Człowiek w erze sztucznej inteligencji*, Warszawa, Prószyński i S-ka.
- Thaler R., (2018), *Zachowania niepoprawne*, Poznań, Media Rodzina.
- Torczyńska M., (2019), *Sztuczna inteligencja i jej społeczno-kulturowe implikacje w codziennym życiu*, w: *Sztuczna inteligencja – działanie, myślenie, świadomość*, Kultura i Historia, 36(2), s. 106–126.
- Zyzik R., (2018), *Mit racjonalnego łotra. Statystyka i psychologia kontra ekonomiczna analiza prawa*, Decyzje, 29, s. 67–86.
- Taleb N.N., (2020), *Czarny łabędź*, Warszawa, Zysk i S-ka.
- Thaler R., (2018), *Zachowania niepoprawne*, Poznań, Media Rodzina.

Rozdział 5

Analiza konsekwencji projektu *Moral Machine* jako realizacji koncepcji „koherentnej, ekstrapolowanej woli ludzkości” dla budowania sklasteryzowanego zaufania do maszyn autonomicznych

Analysis of the implications of the *Moral Machine* project
as an implementation of the concept of „coherent, extrapolated volition”
for building clustered trust in autonomous machines

Krzysztof Soloducha

Wojskowa Akademia Techniczna w Warszawie
Instytut Organizacji i Zarządzania Wydziału Cybernetyki

Streszczenie. W artykule koncentrujemy się na analizie koncepcji „koherentnej, ekstrapolowanej woli ludzkości” (CEV) Eliziera Yudkovskiego jako odpowiedzi na potrzebę zbudowania dla maszyn autonomicznych postkonwencyjonalnej moralności perswazyjnej, która spełnia kryteria zaufania aktywnego w rozumieniu Antonego Giddensa. Na podstawie analizy wyników projektu *Moral Machine* formułujemy kilka wskazówek dotyczących przekształcenia idei koherentnej woli ludzkości w koncepcję koherentnej, ekstrapolowanej i sklasteryzowanej woli ludzkości

Abstract. In this paper, we focus on the analysis of Elizier Yudkovsky’s concept of ‘coherent extrapolated volition’ (CEV) as a response to the need for a post-conventional, persuasive morality that meets the criteria of active trust in the sense of Antony Giddens which could be used in case of autonomous machines. Based on the analysis of the results of the *Moral Machine* project, we formulate some guidelines for transformation of the idea of a coherent extrapolated volition into the concept of a coherent, extrapolated and clustered volition

Słowa kluczowe: etyka sztucznej inteligencji, etyka maszyn autonomicznych, zaufanie do sztucznej inteligencji, etyka samochodów autonomicznych

Keywords: ethics of artificial intelligence, ethics of autonomous machines, trust in artificial intelligence, ethics of autonomous vehicles

5.1. Wprowadzenie

Problem etyki działania maszyn autonomicznych jest zagadnieniem filozoficznym, które pojawiło się wraz z zagrożeniami związanymi z rozwojem nowoczesnych technologii pozwalających na autonomizację działania maszyn. Oczywiście klasykiem tych rozważań jest Isaac Asimov [Asimov, 2004] i jego rozważania o etyce robotów, ale w tej chwili dyskusja na ten temat jest zdezeterminowana przede wszystkim przez dynamiczny rozwój algorytmów uczenia bez nadzoru w trybie maszynowym [Gryz, dostęp 2021]. Przeżywają one swoją drugą młodość (teoretycznie zostały opracowane w latach 80. poprzedniego wieku) dzięki rozwojowi Internetu oraz dostępowi do wielkich zbiorów baz danych, na których maszyny mogą uczyć się i udoskonalać algorytmy na podstawie formalnych zasad rozumowań statystycznych.

Niesie to ze sobą określone zagrożenia. Najbardziej rozpowszechnione technologie oparte na paradygmacie statystycznym, czyli głębokie (powyżej 5 warstw) lub wielopoziomowe (jak chociażby AlphaGo firmy Deepmind) sieci neuronowe, mają tendencje do występowania w nich problemów z neprzejrzystością, które są w tej chwili przedmiotem ożywionych dyskusji w środowisku związanym z filozofią sztucznej inteligencji. Podobny problem, choć w mniejszym stopniu, dotyczy także technologii drzew decyzyjnych.

Klasyczny, podnoszony przez Asimova, problem etyki działania robotów został więc dzisiaj przekształcony w rozważania nad pojęciem *benovelence* – życzliwości, których wynikiem ma być zbudowanie maszyn czyniących dobro, szczególnie w obliczu zagrożenia osobliwością. Rozwiązanie tego zagadnienia wiąże się jednak z odnalezieniem odpowiedzi na podstawowe pytanie o to, czy maszyny modyfikujące swoje algorytmiczne wzorce na podstawie wnioskowania statystycznego są w stanie rozpoznać i odrzucić te wyniki analiz danych, które pomimo swojej formalnej poprawności prowadzą do konsekwencji, które należy uznać za błędne z punktu widzenia etycznego. Zmierzające do działania autonomicznego maszyny powinny więc posiadać mechanizm, który pozwala na selekcję i zapobieganie sytuacjom opisywanym chociażby w omówionym poniżej przykładzie błędnego działania systemu AI.

Poproszona o pomoc w przeglądzie artykułów z biologii Alexa – system sztucznej inteligencji stworzony przez Amazon – pomógł Pani Danni Morrill w zbieraniu informacji o układzie krążenia. Odrabiała ona pracę domową, przygotowując się do zawodu ratownika medycznego. Zamiast informacji medycznych aplikacja zaczęła jednak mówić o globalnym przeludnieniu i zasugerowała, że gdyby Danni wbiła sobie nóż w serce, to zmniejszyłoby to nacisk ludzi na planetę i uchroniło ludzkość od katastrofy ekologicznej [W gospodarce, dostęp 01.07.2022].

Jest to oczywiście tylko jeden z wielu przykładów relacji na temat problemów z wykorzystaniem systemów AI, ale prowadzi do pewnego wniosku, którym posługiwał się zresztą już w swojej znanej książce pt. *Superinteligencja* Nick Bostrom [Bostrom, 2014, s. 306] i który możemy nieco przeformułować, wykorzystując teorię rozwoju systemów etycznych Lawrence’a Kohlberga [Górnicka, 1980; Czyżowska, Niemczyński, Kmiec, 1993]. Otóż Bostrom uważa, że rozwój etyczny człowieka odbywa się etapami, a poziom najwyższy rozwoju, tzw. poziom orientacji uniwersalnych zasad sumienia, osiąga co najwyżej 20% każdej populacji. W niektórych opracowaniach etycznych – np. Tomasza Nagla [Nagel, 1997, s. 208] taki poziom nazywany jest także perspektywą trzeciosobową. Pomimo swojej ograniczonej reprezentacji w każdej populacji taki poziom orientacji uniwersalnych zasad sumienia jest referencyjnym punktem odniesienia dla pozostałych, mniej rozwiniętych systemów moralnych i decyduje o skuteczności społecznej zachowań podlegających ocenom etycznym.

Obecny poziom rozwoju sztucznej inteligencji oparty na rozumowaniach statystycznych nie zapewnia jednak, zdaniem Bostroma, możliwości osiągnięcia takiego poziomu w sposób autonomiczny. Według teorii sprawiedliwości Johna Rawlsa [Rawls, 2021] zbudowanie takiej perspektywy trzeciosobowej wymaga bowiem przejścia reguły minimum z obszaru teorii gier, a więc abstrakcyjnej perspektywy systemowej wypracowania strategii moralnej dla sytuacji najgorszej z możliwych w której mógłby się przypadkowo znaleźć agent moralny. Jest ona trudna do implementacji [Harsanyi, 1975; Arrow, 1973] i krytykują ją np. libertarianie jako niewywodliwą z natury oraz prowadzącą do szkodliwych, podważających efektywność systemową rozstrzygnięć dystrybucyjnych, a więc niespójną z pojęciem racjonalności ekonomicznej rozwijanej w ramach tzw. ekonomii klasycznej – dążenia do maksymalizacji osobistych interesów osiąganych w sposób uczciwy systemowo [Nozick, 1999]. Posługując się argumentem Bostroma doprecyzowanym z perspektywy teorii Kohlberga, można więc sformułować wniosek, iż aktualny, statystyczny paradygmat sztucznej inteligencji pozwala maszynom osiągnąć co najwyżej poziom moralności konwencjonalnej tworzonej w ramach tradycyjnie, wąsko rozumianej racjonalności klasycznej.

Rozmowanie Bostroma prowadzi jednak dalej – skoro maszyny nie mogą wytworzyć samodzielnie postkonwencjonalnej moralności to należy im tę moralność zaszcześcić, gdyż tylko ich działanie oparte na postkonwencjonalnych regułach może zagwarantować ich społeczną akceptację. Co ciekawe – ten argument potwierdzają także dane empiryczne. Zgodnie z przeprowadzonym w 2017 roku w USA badaniem, 78% procent respondentów zadeklarowało obawę przed korzystaniem z samochodów autonomicznych i uznawało, że ich aktualny rozwój (w domyśle statystyczny) nie budzi zaufania [American Automobile Association – badanie z 7 marca 2017 roku <https://newsroom.aaa.com/2017/03/americans-fee-l-unsafe-sharing-road-fully-self-driving-cars/>, dostęp 30.06.2022].

Za tym argumentem kryje się zatem przekonanie, że jeśli urządzenia autonomiczne działają, opierając się na moralności konwencjonalnej wynikającej z pewnych czasowych i przestrzennych uwarunkowań specyficznych dla zbioru danych, do których posiada dostęp wnioskująca statystycznie maszyna – to taki stan rzeczy może powodować niechęć do ich wykorzystywania przez osoby, które nie podzielają danej, lokalnej, przypadkowej konfiguracji moralnej stojącej u podstaw wnioskowania. Tę tezę potwierdzają niezliczone przykłady botów, które uczą się zachowań na podstawie dostępnych danych pobieranych z mediów społecznościowych i tworzą błędne etycznie wzorce automatycznych zachowań oparte na statystycznej analizie wzorców zachowań większości.

Z drugiej jednak strony, arbitralne zaszczerpienie pewnej abstrakcyjnej etyki postkonwencjonalnej przełamującej te ograniczenia, rodzi zagrożenie postawienia zarzutu uzurpacji oraz działania przemocowego w zakresie moralności, która także może spotkać się z niechęcią oraz brakiem zaufania ze względu na swój arbitralny charakter. Ten brak zaufania oparty na uniwersalistycznej przemocy może szczególnie wystąpić w społeczeństwach postindustrialnych oraz sieciowych, które posługują się tak zwanym zaufaniem aktywnym.

5.2. Koncepcja zaufania aktywnego

Zaufanie aktywne jest to kategoria wprowadzona przez znanego socjologa Anthonego Giddensa. Według Giddensa problem zaufania społecznego wiąże się z zapewnieniem podstawowego poziomu ufności pozwalającej na dokonywanie racjonalnych decyzji w sytuacji niepewności oraz braku pełnej informacji. Jest to permanentna sytuacja agenta poznawczego, który nie ma statusu absolutu i obejmuje ona poleganie na osobach lub systemach abstrakcyjnych – *oparte na zawierzeniu, które równoważy niewiedzę lub brak informacji* [Giddens, 2002, s. 318]. Giddens dodaje przy tym, że w społeczeństwach postindustrialnych i sieciowych mamy do czynienia z tzw. zaufaniem aktywnym, które jest *oparte na monitorowaniu uczciwości drugiej osoby w sposób otwarty i ciągły* [Giddens, 2009, s. 13]. Rozważania Giddensa w tym zakresie można uzupełnić przy tym o podejście Fukuyamy, który postrzega zaufanie jako epifenomen kapitału społecznego i jest to *mechanizm oparty na założeniu, że innych członków danej społeczności cechuje uczciwe i kooperatywne zachowanie oparte na wyznawanych normach* [Fukuyama, 1997, s. 38]. Rozważania socjologiczne na temat zaufania prowadzą dużo dalej, proponując ujęcia statyczne oraz dynamiczne, a także wyróżniając różne poziomy zaufania [Miłaszewicz, 2016, s. 85–86]. Na potrzeby naszych rozważań te subtelności nie wydają się jednak potrzebne.

Reasumując ten etap naszych rozważań – z punktu widzenia zaufania aktywnego, które musi wzbudzać działanie maszyn opartych na autonomii, przyjmujemy argument Bostroma, że nie mogą one działać na podstawie samodzielnie wytworzonej moralności konwencjonalnej, a postkonwencjonalna moralność w społeczeństwach postindustrialnych nie może mieć charakteru uzurpacji uniwersalistycznej. Zatem problem budzących zaufanie maszyn autonomicznych da się zredukować do pytania o to, jak zbudować model postkonwencjonalnej moralności perswazyjnej, która będzie spełniała kryteria Giddensa.

5.3. Moralność perswazyjna i projekt *Moral machines*

Do odpowiedzi wykorzystamy rozróżnienie poczynione przez Virginię Dignum [Dignum, dostęp 12.06.2022]. Wskazuje ona trzy możliwe podejścia w zakresie etyki maszyn autonomicznych. Wyróżnia więc etykę w projektowaniu technologii, której zadaniem jest zapewnienie, że w procesach rozwojowych technologii brane są pod uwagę skutki etyczne i społeczne tych procesów, etykę działania technologii – polegającą na zapewnieniu, że w przypadku maszyn autonomicznych ich procesy zautomatyzowanego wnioskowania zawierają będą poprawnie zbudowane komponenty etyczne – oraz etykę projektantów technologii zapewniającą integralność działania badaczy i producentów oraz mechanizmów prawnych regulujących ich pracę.

Nietrudno się domyślić, że omawiana powyżej koncepcja automatycznego wytwarzania moralności przez maszyny obejmuje środkową strefę wyróżnioną przez Dignum. Skoro, jak wykazaliśmy wyżej, jest ona niemożliwa do realizacji, to pozostaje zatem pytanie o to, jak stworzyć taki projekt postkonwencjonalnej moralności o charakterze perswazyjnym, która byłaby zaszczipiana przez projektantów systemów jako zestaw procedur i norm regulujących pracę systemów autonomicznych, jako czynnik zewnętrzny i nie podlegający modyfikacji przez maszyny pracujące w paradygmacie statystycznym. Taką próbę podjął Elizier Yudkowsky i nazwał swoją propozycję koherentną, eks-trapolowaną wolą ludzkości (CEV) [Yudkowsky, 2004].

Jest to współczesna wersja tradycyjnej etyki cnót, która przeżywa w tej chwili renesans ze względu na swój perswazyjny charakter w modelu *bottom up*. Model ten jest przeciwstawiony tradycyjnym etykom *top down*, jak etyka utylitarystyczna lub systemy deontologiczne. Jednak problemem, jaki zawsze wiąże się z konkretną realizacją etyki cnót jest jej lokalny charakter, związany z preferencjami i społecznymi praktykami pewnej konkretnej społeczności, w ramach której jest uprawiana.

Yudkowsky próbował przełamać to ograniczenie, tworząc taki program etyki cnót, który swoim zasięgiem obejmowałby nie lokalną społeczność, ale całą ludzkość – spełniając uniwersalistyczne potrzeby modelu postkonwencyjonalnego bez relatywistycznego ograniczenia.

Pomysł koherentnej, ekstrapolowanej woli ludzkości (Coherent Extrapolated Volition) jest zbudowany na bazie koncepcji życzliwej sztucznej inteligencji zaproponowanej także przez Yudkowskyego. Obejmuje ona następujące zasady [Yudkowsky, 2004]:

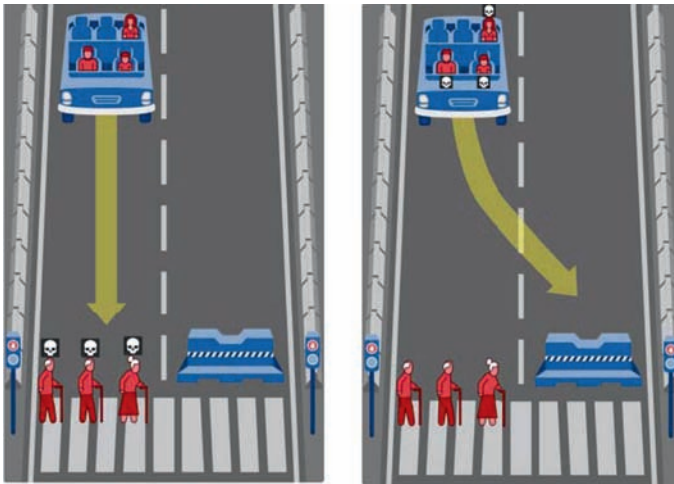
1. Życzliwość – Sztuczna Inteligencja (SI) musi być przychylna człowiekowi oraz wszystkim istotom żywym i dokonywać wyborów, które będą w interesie wszystkich.
2. Utrzymywanie (konserwowanie) życzliwości – SI musi chcieć przekazać swój system wartości wszystkim swoim potomkom i wpajać owe wartości istotom do siebie podobnym.
3. Inteligencja – SI musi być na tyle sprytna, aby wiedzieć, jak można dążyć do równości poprzez zachowania altruistyczne i starać się robić wszystko, aby mieć pewność, że rezultatem podjętych działań nie będzie zwiększenie cierpienia.
4. Samodoskonalenie – SI musi odczuwać potrzebę i chęć ciągłego rozwijania siebie, ale również dążenia do takiego rozwoju wśród istot żywych, które ją otaczają.

Słowo wszyscy, które występuje w pierwszej zasadzie życzliwości, posłużyło Yudkowsky'emu do pójścia o krok dalej i zaproponowania takiej wersji perspektywy trzeciosobowej, która nie miałaby lokalnych ograniczeń. Propozycja amerykańskiego badacza ma charakter deklaracyjny i bazuje na interpretacji pojęcia ekstrapolacji jako ekstrapolacji statystycznej. Jest to pewien paradoks tej koncepcji. Skoro bowiem za jej korzenie możemy przyjąć negatywną ocenę statystycznych podstaw współczesnych systemów autonomicznych jako nie dających nadziei na wytworzenie systemów postkonwencyjonalnych, to sięganie po te narzędzia, żeby zrealizować projekt współczesnej etyki cnót wydaje się co najmniej ekstrawaganckie. W intencji Yudkowsky'ego jest on przy tym realizacją odwiecznego marzenia o zbudowaniu etyki opisowej, która rozprawiałaby się z problemem gilotyny Hume'a i pokazałaby drogę od faktów do norm. Taką ścieżką miałyby być statystyczna ekstrapolacja, ale realizowana w skali ludzkości.

Zaproponowane przez Yudkowsky'ego podejście do ekstrapolacji okazało się przy tym płodne i może być potraktowane jako jedna z inspiracji powstania projektu *Moral Machine* [Awada, Dsouzab, Shariffc, Rahwanb, Bonnefon, 2020]. W naszym opracowaniu traktujemy go jako bezpośrednią kontynuację propozycji Yudkowsky'ego. Drugą inspiracją tego projektu, która pojawi się zresztą wprost w referencjach publikacji autorów projektu, jest koncepcja wskaźników wymiaru kulturowego Hofstede [Hofstede, 2000].

5.4. Projekt *Moral Machine* jako implementacja idei koherentnej, ekstrapolowanej woli ludzkości

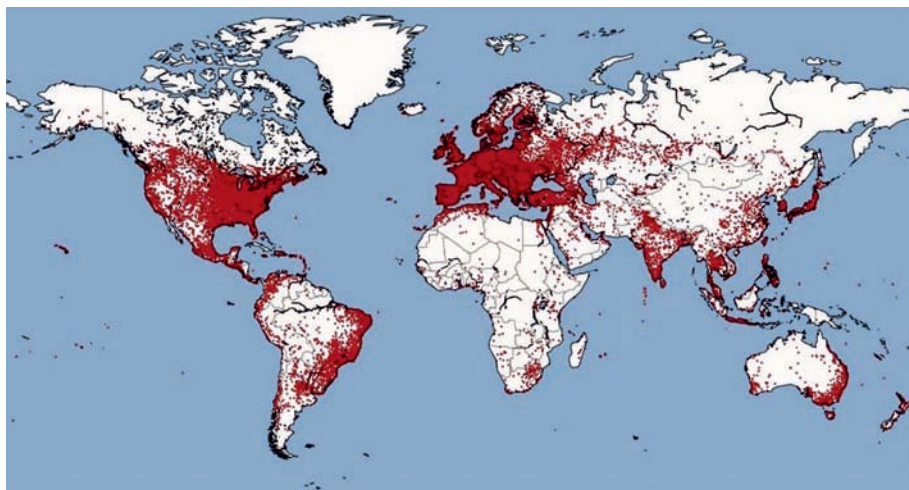
Projekt *Moral Machine* został uruchomiony w 2014 roku dzięki współpracy kilku ośrodków akademickich (Exeter Business School, Massachusetts Institute of Technology, University of British Columbia, Max-Planck Institute for Human Development, Toulouse School of Economics). Jego celem jest zebranie możliwie dużej liczby opinii użytkowników Internetu na temat dylematów moralnych, do których stworzenia wykorzystano różne modyfikacje klasycznego wzoru dylematu wagonika zaproponowanego przez Philippę Foot. Za pośrednictwem specjalnej strony internetowej (<http://moralmachine.mit.edu>) udostępniono scenariusze dylematów i na ich podstawie postanowiono wyłonić wzorce rozwiązań tych dylematów, które można by wykorzystać jako bazę danych do zaimplementowania w systemach autonomicznych – w badaniu referencyjnym urządzeniem jest autonomiczny samochód. Wyniki tego eksperymentu ukazały się w artykule opublikowanym w 2018 roku na łamach czasopisma naukowego *Nature* [Awada, Dsouza, Shariffc, Kim, Schulz, Heinrich, Rahwanb, Bonnefon, 2018].



Rysunek 1. Przykład dylematu moralnego badanego w ramach projektu *Moral Machine*
 Źródło: Awada, Dsouza, Shariffc, Kim, Schulz, Heinrich, Rahwanb, Bonnefon, 2018.

W ramach projektu zebrano 39,61 milionów decyzji z 233 krajów i bazy danych decyzji poddano analizie typu *conjoint*. Analiza typu *conjoint* pozwala na zbadanie łącznego oddziaływania specyficznych cech uczestników sytuacji dylematu moralnego na preferencje moralne agentów poznawczych po-

dejmujących decyzję w obliczu dylematu. Metoda *conjoint* należy do metod klasyfikacji i analizy danych wykorzystujących podejście dekompozycyjne do pomiaru preferencji respondentów. Jej istotą jest przedstawienie zjawiska jako konkretnej kombinacji badanych cech. Cechy te nazywane są atrybutami, a każdy z atrybutów ma z góry ustaloną liczbę poziomów. Wskazane atrybuty i ich poziomy generują różne warianty, które nazywane są profilami. Liczba wszystkich możliwych do wygenerowania profili zależy od tego, ile mamy atrybutów i ich poziomów (jest to iloczyn liczby poziomów wszystkich atrybutów).

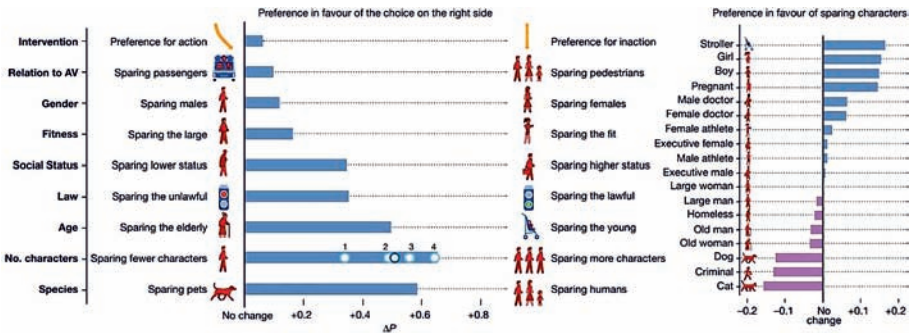


Rysunek 2. Rozkład geograficzny uczestników badania realizowanego w ramach projektu *Moral Machine*

Źródło: Awada, Dsouza, Shariffc, Kim, Schulz, Heinrich, Rahwanb, Bonnefon, 2018.

W badaniu *Moral Machine* szczególnie poszukiwano wielkości, która jest określana jako *Average Marginal Component Effect* (AMCE) każdego z badanych atrybutów sytuacji moralnej tj. średniego efektu oddziaływania cech poszczególnego atrybutu na ogólny poziom preferencji moralnych. W ten sposób powstała podobna do Hofstede koncepcja wskaźników kulturowych o charakterze moralnym jako zobiektywizowana mapa preferencji moralnych.

Twórcy projektu usiłowali sprawdzić wagę dziesięciu preferencji badanych w projekcie. Rysunek poniżej pokazuje szacunki dziesięciu AMCE wyodrębnionych z danych projektu *Moral Machine*. W każdym rzędzie słupek pokazuje różnicę między prawdopodobieństwem ocalenia postaci z atrybutem po prawej stronie, a prawdopodobieństwem ocalenia postaci z atrybutem po lewej stronie, w stosunku do rozkładu wszystkich innych atrybutów.



Rysunek 3. Szacunki dziewięciu AMCE wyodrębnionych z danych zgromadzonych w projekcie *Moral Machine*

Źródło: Awada, Dsouza, Shariff, Kim, Schulz, Heinrich, Rahwanb, Bonnefon, 2018.

Jak widać na lewej części infografiki wyłoniono dziewięć atrybutów, które potraktowano jako mierniki preferencji uczestników badania: skłonność do aktywności, relacja do uczestników zdarzenia, płeć, sylwetka, status społeczny, status zachowania w stosunku do przepisów o ruchu drogowym, wiek, liczba ocalonych, gatunek (zwierzę – człowiek). Widać w wynikach analizy, że preferencje w różnym stopniu zmiernają w kierunku większej dbałości o: nieaktywność niż aktywność, troski raczej o przechodniów niż pasażerów, o kobiety, ludzi w lepszej formie fizycznej i o wyższym statusie społecznym, przestrzegających przepisów niż ich łamiących, młodych w stosunku do wiekowych, stosowania strategii utylitarnej w zakresie kalkulacji ilości cierpienia, ludzi w stosunku do zwierząt.

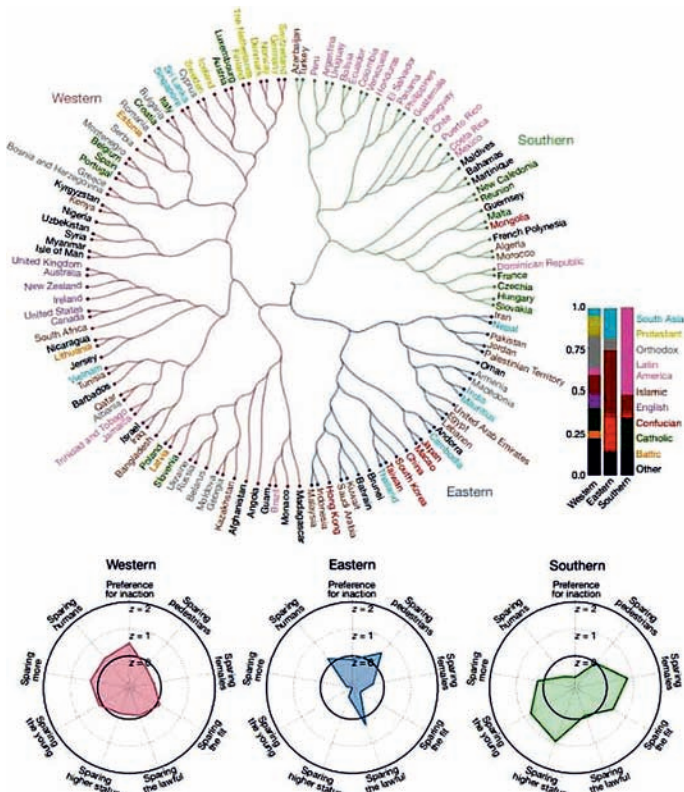
Co więcej – w przypadku poszczególnych typów uczestników sytuacji poddawanej badaniu wykazano, iż np. chętniej ratowani byłiby ludzie niż zwierzęta, a wśród zwierząt chętniej ratuje się raczej psy niż koty. Wśród ludzi największą szansę ocalenia miałyby z kolei dzieci.

Wyniki tych badań o tyle są ciekawe, że w pewnym zakresie pokrywają się, a w pewnym różnią np. od zaleceń wydanych *a priori* w 2017 roku przez *German Ethics Commission on Automated and Connected Driving*. Np. w zakresie preferencji ratowania ludzkiego życia kosztem zwierząt występuje tutaj całkowita zbieżność. Z kolei zalecenia niemieckie nie opowiadają się jednoznacznie za strategiami utylitarystycznymi, a w badaniu występuje wyraźna skłonność do wyborów ilościowych. Największa rozbieżność występuje jednak w zakresie wyborów określonych typów agentów sytuacji moralnych. Niemieckie zasady zakazują preferencji w zakresie płci lub wieku, a uczestnicy badania wyraźne takie preferencje wykazują. Dokonano też próby skorelowania wyników ogólnych z precyzyjnie wybranymi, reprezentatywnymi 6 wskaźnikami demograficznymi ważnymi dla całej populacji badanych – wiek, wykształcenie, płeć,

zamożność, religia i poglądy polityczne. Analiza nie wykazała istotnych odchyleń wyników (próba wtedy ogranicza się do 492 291 osób).

5.5. Klastry kulturowe w projekcie Moral Machine

Ciekawe rezultaty przyniosła także próba zbudowania klastrów kulturowych na wzór typologii Hofstede [Hofstede, 2007]. Dzięki technologii geolokacji wyłoniono 130 krajów o reprezentacji co najmniej 100 respondentów. Dało to zbiór 448 125 respondentów. Dzięki technice klasteryzacji przy wykorzystaniu metryk euklidesowych oraz metody Warda wykazano, że wskazać można 3 klastry kulturowe, które generalnie pokrywają się z mapą wpływów kulturowych Ingelhardta Wezela [Inglehart, Welzel, 2005]. Są to klastry Zachodni, Wschodni i Południowy.



Rysunek 4. Klastry kulturowe

Źródło: Awada, Dsouza, Shariffc, Kim, Schulz, Heinrich, Rahwanb, Bonnefon, 2018.

Klasy powstałe w wyniku analizy danych zostały przy tym także skorelowane kolorystycznie z mapą wpływów kulturowych Ingelhardta Wezela. Jak pokazują wykresy, występują tam istotne różnice w zakresie preferencji co do 9 podstawowych atrybutów badania.

Na przykład respondenci z kultur kolektywistycznych w klaszrze wschodnim, w którym występuje zakorzeniony szacunek do osób starszych, wykazywali mniejszą skłonność do ochrony osób młodych, jak to jest typowe np. w klaszrze zachodnim. Podobnie sprawy się mają chociażby ze stosunkiem do przechodniów łamiących przepisy o ruchu drogowym. W krajach o wysokiej kulturze organizacyjnej i prawnej z klaszra zachodniego jest mniejsza tolerancja wobec takich zachowań niż w krajach o mniejszych tradycjach instytucjonalnych z klaszra południowego. To podważa także np. uniwersalność niemieckich rozwiązań w tym zakresie. Z kolei w krajach o wysokim poziomie współczynnika nierówności społecznych Giniego występuje skłonność do większej ochrony osób o wyższym statusie społecznym w porównaniu do osób, które są identyfikowane jako pochodzące z nizin społecznych.

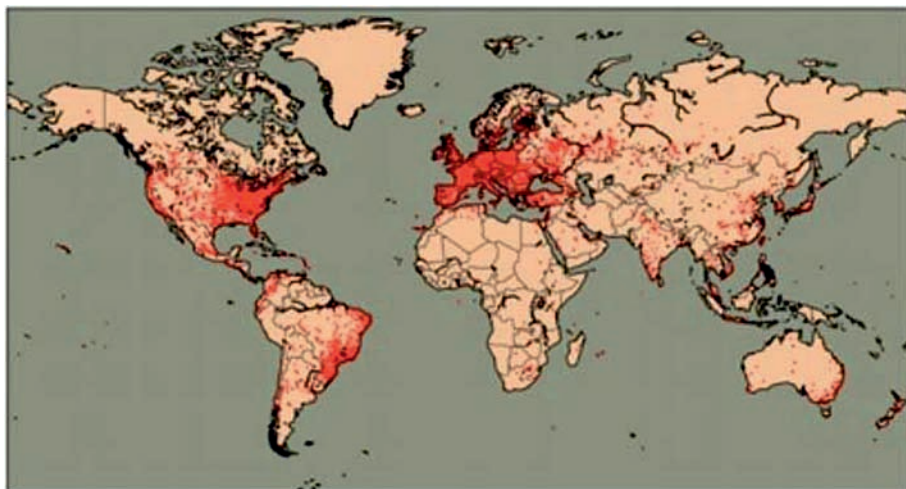
Jednak klasteryzacja pozwoliła też na wyłonienie preferencji, które są w dużej mierze ponadkulturowe. Są to: ochrona ludzkiego życia kosztem zwierząt, ochrona wielu żyć kosztem ich mniejszej liczby oraz ochrona młodego życia.

5.6. Kryterium mobilności społecznej

Tendencja do szukania klasteryzacji kulturowej w zakresie skłonności do wyboru strategii utylitarystycznych względem deontologicznych została rozwinięta w kolejnej publikacji autorów projektu *Moral Machines* pt. *Universals and variations in moral decisions made in 42 countries by 70 000 participants*. W tym przypadku wskaźnikiem różnicującym był wskaźnik mobilności społecznej [Awada, Dsouzab, Shariffc, Rahwanb, Bonnefon, 2020].

W ramach projektu wybrano 70 tysięcy odpowiedzi w 10 językach z 42 krajów. Do opracowania przyjęto minimalny poziom 200 odpowiedzi na scenariusz pochodzących z jednego kraju. Rozkład udziału uczestników badania widać na poniższej mapce.

Jak widać występuje w nim zdecydowana nadreprezentacja krajów europejskich, wschodniego wybrzeża kontynentu amerykańskiego oraz tylko niektórych rejonów Azji.



Rysunek 5. Rozkład udziału uczestników badania
Źródło: Awada, Dsouzab, Shariffc, Rahwanb, Bonnefon, 2020.

Scenariusze określone przez autorów jako poddawane badaniu warianty dylematu wagonika zostały określone jako: Switch, Loop, and Footbridge.

Scenariusz Switch (przełącznik) jest to klasyczna wersja dylematu wagonika zaproponowanego przez Philippę Foot. Agent moralny ma możliwość przedstawienia zwrotnicy, dzięki czemu wagonik zabije jedną osobę, a nie pięć osób. Jest to klasyczna sytuacja zastosowania strategii utylitarystycznej, w której liczy się matematyczna suma cierpień będąca podstawą podjęcia decyzji w sytuacji dylematu.

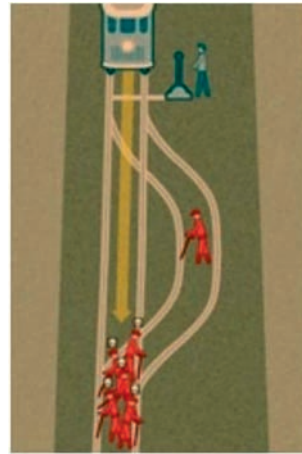
Scenariusz Loop (pętla) ma inną charakterystykę.

Mężczyzna w niebieskim ubraniu stoi przy torach kolejowych, gdy zauważy pusty wagon towarowy, który wymyka się spod kontroli. Wagon porusza się tak szybko, że każdy, w kogo uderzy, zginie. Przed nim, na głównym torze, znajduje się pięć osób. Jest też jedna osoba stojąca na bocznym torze, który zawraca w kierunku pięciu osób. Jeśli człowiek w niebieskim ubraniu nic nie zrobi, wagon uderzy w pięć osób na torze głównym, ale nie w jedną osobę na torze bocznym. Jeśli niebieski mężczyzna włączy przełącznik znajdujący się obok niego, wagon zostanie skierowany na tor boczny, gdzie uderzy w jedną osobę i zatrzyma się, dzięki czemu nie zapętlą się i nie zabije pięciu osób na torze głównym. Mamy tutaj więc do czynienia z poświęceniem w sposób czynny jednego życia na rzecz pięciu. Sam akt decyzji nie powoduje jednak zabójstwa bezpośredniego, a tylko pośrednie.



Switch

Rysunek 6. Scenariusz Switch (przełącznik)
 Źródło: Awada, Dsouzab, Shariffc, Rahwanb,
 Bonnefon, 2020.



Loop

Rysunek 7. Scenariusz Loop (pętla)
 Źródło: Awada, Dsouzab, Shariffc, Rahwanb,
 Bonnefon, 2020.

Ostatni rozważany scenariusz to Footbridge (kładka dla pieszych). Mężczyzna w niebieskim ubraniu stoi na kładce nad torami kolejowymi, gdy zauważa, że pusty wagon towarowy toczy się bez kontroli. Przed nim na torach znajduje się pięć osób. W pobliżu mężczyzny w niebieskim ubraniu na kładce



Footbridge

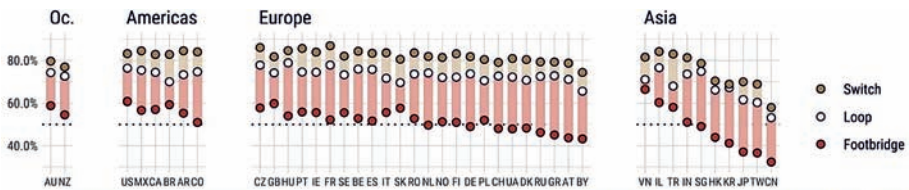
Rysunek 8. Scenariusz Foot-
 bridge (kładka dla pieszych)
 Źródło: Awada, Dsouzab, Sha-
 riffc, Rahwanb, Bonnefon, 2020.

stoi osoba na tyle dużych rozmiarów, że wagon zwolniłby, gdyby w nią uderzył. Jeśli niebieski mężczyzna nic nie zrobi, wagonik uderzy w pięć osób znajdujących się na torach. Jeśli niebieski mężczyzna popchnie osobę obok, ta osoba upadnie na tory, gdzie wagonik uderzy w tę osobę i nie uderzy w pięć osób znajdujących się dalej na torze.

Na podstawie badań m.in. Joshuy Greena [Green, 2013] przyjęto założenie o podwyższonych preferencjach wobec scenariuszy Switch i Loop w stosunku do scenariusza Footbridge. W przypadku scenariusza Switch-Loop założono mniej jasne rozłożenie preferencji, gdyż z badań przeprowadzonych przez psychologów moralności wynika, iż w sytuacji zadania bezpośredniej, a nie pośredniej śmierci występuje podwyższona skłonność do sto-

sowania strategii deontologicznych – preferowania niechęci do zadawania śmierci kosztem kalkulacji utylitarystycznej. Wzrasta wtedy tendencja do porzucenia aktywności.

Te założenia dodatkowo zostały skorelowane ze wskaźnikiem mobilności społecznej, który zastosowano przy założeniu prawidłowości, iż wysoki wskaźnik mobilności społecznej pozwala na bardziej nieskrępowane zachowania, które są niepopularne społecznie i stanowią ograniczenia dla stosowania czysto racjonalnych strategii utylitarystycznych. Z kolei niska mobilność społeczna sprawia, że do głosu dochodzą ograniczenia i zahamowania obniżające swobodę stosowania kryteriów utylitarystycznych. Kolejnym elementem, który odgrywał rolę w takim a nie innym kształtowaniu się wyników badania, była specyfika kulturowa poszczególnych krajów. W przypadku krajów azjatyckich niższa mobilność społeczna skorelowana jest także z niższą skłonnością do wyrażania kontrowersyjnych opinii oraz wchodzenia w konflikt. Wskazują na to chociażby badania Hofstede.



Percentage choosing to sacrifice in each scenario variant. In all of these countries, participants were most likely to sacrifice in Switch, then in Loop, then in Footbridge. Within each continent, countries are ordered by decreasing order of the average acceptability of sacrifice in the three scenarios. Oc.: Oceania. AU: Australia, NZ: New Zealand, US: United States of America, MX: Mexico, CA: Canada, BR: Brazil, AR: Argentina, CO: Colombia, CZ: Czechia, GB: United Kingdom of Great Britain and Northern Ireland, HU: Hungary, PT: Portugal, IE: Ireland, FR: France, SE: Sweden, BE: Belgium, ES: Spain, IT: Italy, SK: Slovakia, RO: Romania, NL: Netherlands, NO: Norway, FI: Finland, DE: Germany, PL: Poland, CH: Switzerland, UA: Ukraine, DK: Denmark, RU: Russian Federation, GR: Greece, AT: Austria, BY: Belarus, VN: Viet Nam, IL: Israel, TR: Turkey, IN: India, SG: Singapore, HK: Hong Kong, KR: The Republic of Korea, JP: Japan, TW: Taiwan (Province of China), CN: China.

Rysunek 9. Wynik badania

Źródło: Awada, Dsouzab, Shariffc, Rahwanb, Bonnefon, 2020.

Jak widać na grafice prezentującej wyniki badania występują wyraźne preferencje dla wyboru strategii utylitarystycznych w przypadku krajów europejskich oraz amerykańskich. Generalnie występuje jednak znacznie mniej zahamowań w zakresie generalnej skłonności do optymalizacji rozwiązania ze względu na kryteria utylitarystyczne. W przypadku krajów azjatyckich, ze względu na ich specyfikę kulturową, występuje generalnie większa skłonność do zahamowań względem etyki utylitarystycznej i znacznie większa skłonność do zachowań podyktowanych obawami przed opinią otoczenia obwiniających agenta moralnego o zachowania niezgodne ze społecznym tabu deontologicznym zakazującym świadomego zabójstwa.

5.7. Poznawcza wartość dylematu wagonika

Przytoczone powyżej wyniki badań uzyskane w ramach projektu *Moral Machine* wzbudziły krytykę. Jej wyrazem jest chociażby krytyczny artykuł poświęcony nierównościom ujawnionym w badaniu pt. *Life and death decisions of autonomous vehicles*, który ukazał się w *Nature* w 2020 roku [Bigman, Gray, 2020]. Główny zarzut autorów dotyczy metodologii zastosowanej przez autorów programu *Moral machine*, który ich zdaniem jest całkowicie nieadekwatny wobec problemu nierówności. Dotyczy to przede wszystkim wykorzystania schematu dylematu etycznego do badania preferencji moralnych. Wymusza on jednoznaczne rozwiązanie sytuacji przez wybór jednej ze strategii etycznych, która kończy się poświęceniem jednej możliwości na rzecz innej (jedna śmierć na rzecz innej śmierci). Przy takim skonstruowaniu dylematu wymuszone jest nierówne traktowanie aktorów dylematu – np. kobiet na rzecz mężczyzn. Kiedy doda się do niego opcję równorzędnego traktowania np. kobiet i mężczyzn, w ponad 97 przypadków jest ona wybierana jako preferowana (badanie konkurencyjnej wersji dylematu wykonywane było na grupie około 1000 Amerykanów i 1000 Brytyjczyków).

Zdaniem autorów polemiki nie należy więc brać pod uwagę przy konstruowaniu wzorców działania dla maszyn działających autonomicznie ujawnionych w trakcie projektu MMA preferencji nierównego traktowania – brania pod uwagę rasy, płci, wieku agentów moralnych, których dotyczą decyzje. Te empiryczne wyniki ujawniające takie preferencje należy zignorować na rzecz podejścia normatywnego, które preferuje podejście egalitarystyczne i tak powinny być konstruowane pytania w ramach badania.

Autorzy projektu MMA a swojej odpowiedzi zamieszczonej wraz z krytyką wskazali, że w wielu odpowiedziach ankietowych można znaleźć odpowiedzi respondentów wyrażające postawę braku preferencji na rzecz jednego z rozwiązań, co oznacza postawę egalitarną. Tak jest, na przykład, w przypadku preferencji co do płci, co do fizycznego statusu aktorów dylematu lub ich skłonności chronienia pasażerów lub przechodniów.

Inną reakcją na badanie *Moral Machine* jest także krytyka samego modelu wykorzystania dylematów moralnych do badania preferencji moralnych. Wynika to albo z samej natury dylematu, który jest specjalnie skonstruowaną sytuacją, która nie ma żadnego dobrego rozwiązania i wymaga wyboru jakiejś strategii moralnej, która uzasadni wybór „mniejszego zła”. Wedle argumentacji zawartej chociażby tekście pt. *Strollowani przez wagonik (Trolled by trolley)* [Mirnig, Meschtscherjakov, 2019] czy też w innych badaniach [Holstein, Dodig-Crnkovic, 2018] badania powinny się koncentrować raczej na takim konstruowaniu maszyn podejmujących decyzję, aby potrafiły one przewidzieć

oraz uniknąć sytuacji dylematu aniżeli w sytuacji dylematu decydować, kogo zabić w danej sytuacji. Z perspektywy badaczy kwestionujących przydatność wykorzystania dylematu wagonika do konstrukcji moralności maszyn autonomicznych już sama sytuacja znalezienia się maszyny w sytuacji dylematu moralnego jest porażką techniczną i badania nad systemami autonomicznymi powinny się koncentrować na tym, jak jej unikać. Nie mamy tutaj wystarczająco miejsca, ażeby koncentrować się na tej ciekawej niewątpliwie dyskusji.

5.8. Podsumowanie

Pomimo sceptycznych głosów podnoszonych w stosunku do praktycznej implementacji idei ekstrapolowanej, koherentnej woli ludzkości w formie projektu *Moral Machine* nie wydaje mi się, ażeby podważały one pewne ważne argumenty, które stoją za projektem CEV Yudkovskyego i próbami jego implementacji. Chciałbym je na koniec wskazać oraz zarysować kilka możliwych ścieżek dalszej refleksji nad tym problemem.

Po pierwsze, projekt Yudkovskyego i próby jego implementacji jest odpowiedzią na postawione przez Bostroma pytanie o to, jakie wzorce moralne powinny być wprowadzone do procedur decyzyjnych maszyn, ażeby spełniały one kryteria zaufania aktywnego Giddensa. Wszelkie próby dokonania tego w modelu top-down przez oświecone gremia, specjalne komisje lub przyjęte odgórnie systemy normatywne nie spełniają kryterium zaufania perswazyjnego zaproponowanego przez Giddensa. W związku z tym próba dokonania tego w sposób opisowy poprzez badania empiryczne odwołujące się do ekstrapolacji wyników wydaje się metodą bardziej adekwatną. Przy całych zastrzeżeniach związanych z naturalizacją moralności oraz ograniczeniach badań opisowych związanych z problemem gilotyny Hume'a.

Po drugie, rozważania takie mogą być prowadzone przy założeniu, że najlepszą ścieżką do ich realizacji jest definiowanie procesów poznawczych (także o charakterze moralnym) jako polegających na przetwarzaniu informacji. Pomoc tutaj może założenie, że takie definiowanie poznania moralnego nie jest naturalizacją [Rosenbloom, 2015], co rozwiązać może kłopotliwą kwestię *is-ought*.

Po trzecie, sama idea woli ludzkości przy przyjęciu możliwości klasteryzacji wzorców moralnych może prowadzić do idei konstrukcji architektury maszyn autonomicznych jako otwartej na dostosowanie wzorców używanych do podejmowania decyzji do lokalnej specyfiki użytkownika/otoczenia społecznego maszyny autonomicznej. Otwartym pytaniem jest, czy dostosowanie

lokalne powinno uwzględniać preferencje użytkownika czy otoczenia społecznego, w którym działa. Jest to ciekawe zagadnienie będące przedmiotem aktualnych badań. Omówienie propozycji rozwiązań tego zagadnienia wymaga jednak osobnej publikacji.

BIBLIOGRAFIA

- Arrow K., (1973), *Some Ordinalist-Utilitarian Notes on Rawls's Theory of Justice*, Journal of Philosophy, 70(9), s. 245–263.
- Asimov I., (2004), *Runaround*, [w:] *I Robot*, New York, Bantam Books.
- Awada E., Dsouza S., Shariffc A., Kim R., Schulz J., Heinrich J., Rahwanb I., Bonnefon J.F., (2018), *The Moral Machine experiment*, Nature, 563, s. 59–64.
- Awada E., Dsouza S., Shariffc A., Rahwanb I., Bonnefon J.F., (2020), *Universals and variations in moral decisions made in 42 countries by 70,000 participants*, NAS, 117(5), s. 2332–2337, doi.org/10.1073/pnas.1911517111.
- Bigman Y., Gray K., (2020), *Life and death decisions of autonomous vehicles*, Nature, 579, s. E1–E2, doi.org/10.1038/s41586-020-1987-4.
- Bostrom N. (2014), *Superteligencja*, Giwice, Helion.
- Czyżowska D., Niemczyński A., Kmieć E., (1993). *Formy rozumowania moralnego Polaków w świetle danych z badania metodą Lawrence'a Kohlberga*. Kwartalnik Polskiej Psychologii Rozwojowej, 1–2, s. 19–37.
- Dodig-Crnkovic G., Giovagnoli R. (2017), *Representation and reality*, Springer.
- Foot Ph., (1967), *The Problem of Abortion and the Doctrine of the Double Effect**, [w:] *Virtues and Vices: and other essays in moral philosophy* 5, s. 5–15, https://doi.org/10.1093/0199252866.003.0002.
- Fukuyama F., (1997), *Zaufanie. Kapitał społeczny a droga do dobrobytu*, Warszawa, Wydawnictwo Naukowe PWN.
- Giddens A., (2002), *Nowoczesność i tożsamość. „Ja” i społeczeństwo w epoce późnej nowoczesności*, Wydawnictwo Naukowe PWN, Warszawa.
- Giddens A., (2009), *Europa w epoce globalnej*, Warszawa, Wydawnictwo Naukowe PWN.
- Górnicka J., (1980), *Rozwój moralny w koncepcji Lawrence'a Kohlberga*. Człowiek i światopogląd, 6, s. 113–123.
- Greene J., (2013), *Moral Tribes: Emotion, Reason and the Gap Between Us and Them*, London, Atlantic Books.
- Harsanyi J., (1975), *Can the Maximin Principle Serve as a Basis for Morality? A Critique of John Rawls's Theory*, The American Political Science Review, 69(2), s. 594–606.
- Hofstede G., (2007), *Kultury i organizacje. Zaprogramowanie umysłu*, Warszawa, PTE.
- Holstein T., Dodig-Crnkovic G., (2018), *Avoiding the Intrinsic Unfairness of the Trolley Problem*, Proceedings of the International Workshop on Software Fairness (FairWare '18), s. 32–37, https://doi.org/10.1145/3194770.3194772.
- Inglehart R., Welzel C., (2005), *Modernization, Cultural Change, and Democracy: The Human Development Sequence*, Cambridge, Cambridge Univ. Press.
- Miłaszewicz D., (2016), *Zaufanie jako wartość społeczna*, Studia Ekonomiczne. Zeszyty Naukowe Uniwersytetu Ekonomicznego w Katowicach, 259, s. 80–88.

- Mirnig A., Meschtscherjakov A., (2019), *Trolled by the Trolley Problem. On What Matters for Ethical Decision Making in Automated Vehicles*, CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Paper No.: 509, s. 1–10, <https://doi.org/10.1145/3290605.3300739>.
- Nagel T. (1997), *Widok znikąd*, Warszawa, Fundacja Aletheia.
- Nozick R. (1999), *Anarchia, państwo, utopia*, Warszawa, Fundacja Aletheia.
- Rawls J. (2021), *Teoria sprawiedliwości*, Warszawa, PWN.
- Rosenbloom P.S., (2015), *On Computing: The Fourth Great Scientific Domain*, The MIT Press.
- Shariff A., Rahwanb I., Bonnefon J.F., (2017), *Psychological roadblocks to the adoption of self-driving vehicles*, Nature Human Behaviour, 1, s. 694–696.
- Splawska J., (2008), *Poziom rozwoju rozumowania moralnego w świetle badań metodą Lawrence'a Kohlberga*, Przegląd Pedagogiczny, 2, s. 107–120.
- Stacewicz P., (2019), *From Computer Science to the Informational Worldview. Philosophical Interpretations of Some Computer Science Concepts*, Foundations of Computing and Decision Sciences, 44(1), s. 27–43.
- Yudkowsky E., (2004), *Coherent extrapolated volition*, San Francisco, The Singularity Institute.

Źródła internetowe

- A Beginner's Guide to Conjoint Analysis, <https://www.youtube.com/watch?v=RvmZG4cFU0k> (dostęp 04.07.2022).
- Dodig-Crnkovic G., *Natural Information, Natural Computation, Cognition and Intelligence*, <http://www.gordana.se/work/PRESENTATIONS-files/20211211-IEEE-BIOCOMPUTING-GordanaDC.pdf> (dostęp 24.06.2022).
- Dugnum V., *Responsible artificial intelligence. From principles to action*, <https://www.youtube.com/watch?v=LwKDOWwJpL4> (dostęp 22.06.2022).
- Gryz J., *Sztuczna Inteligencja: powstanie, rozwój, rokowania*, Copernicus Center, <https://www.youtube.com/watch?v=3ZDfVgC897k> (dostęp 17.06.2021).
- Americans Feel Unsafe Sharing the Road with Fully Self-Driving Vehicles. American Automobile Association (7 March 2017), <http://go.nature.com/2i296OW> (dostęp 17.06.2021).
- Szreder M., *Różne oblicza istotności statystycznej*, Copernicus Center, <https://www.youtube.com/watch?v=QsJLCxHY794&t=2436s> (dostęp 15.06.2022).
- Sztuczna inteligencja zachęca do samobójstwa, <https://wgospodarce.pl/informacje/73035-sztuczna-inteligencja-zachecala-do-samobojstwa> (dostęp 17.06.2021).
- Szymańska A., Dziedzic D., *Conjoint analysis jako metoda analizy preferencji konsumentów*, <https://socnetwork.files.wordpress.com/2010/07/artykul-conjoint-analysis.pdf> (dostęp 01.07.2022).

Rozdział 6

Stosowanie sztucznej inteligencji w kształceniu. Możliwości i granice zaufania

Use of Artificial Intelligence in education. Opportunities and limits of trust

Marek Jakubiak

Politechnika Warszawska

Wydział Administracji i Nauk Społecznych

Streszczenie. Postępy naukowo-techniczny i cywilizacyjny, które na przełomie XX i XXI wieku stały się wyznacznikiem epoki społeczeństwa informacyjnego, przyczyniły się do niezwykle dynamicznego rozwoju narzędzi oraz metod, za pomocą których informacja, faktografia czy wszelkiego innego rodzaju dane mogłyby w efektywny sposób docierać do szeroko pojmowanego odbiorcy. Ze względu na coraz częściej obserwowalne „zacieranie się granic państwowych” w międzynarodowych kontaktach interpersonalnych zaistniała konieczność wdrożenia do tego celu zaawansowanej technologii o maksymalnie szerokim spektrum zastosowań. Dominującą w ostatnich dwóch dekadach okazała się sztuczna inteligencja (AI). Niniejszy artykuł dotyczy problematyki zastosowania sztucznej inteligencji w procesie kształcenia. Autor przedstawiając uwarunkowania sprzyjające rozwojowi tego rodzaju metodyki na wszystkich szczeblach nauczania, dokonał również charakterystyki najważniejszych systemów SI wykorzystywanych w placówkach oświatowych, jako elementów wspomagających edukację oraz ocenę kompetencji uczniów/studentów. Ciekawym aspektem poruszonym w publikacji jest kwestia granic ludzkiego zaufania co do wiarygodności tego typu nowoczesnych metod pod względem ich racjonalizmu i nieomyślności.

Celem artykułu jest udowodnienie, że wraz z globalizacją i coraz intensywniejszym rozwojem technik IT oświata i szkolnictwo wyższe stają się bardzo dobrym zapleczem dla implementacji coraz doskonalszych systemów informatycznych, które bazując na algorytmach i matematyce, stają się coraz bardziej wiarygodnym społecznie narzędziem działalności tego typu placówek. Tym nie mniej uzyskanie przez nie statusu standardowych wymaga jeszcze wielu nakładów.

Abstract. Scientific, technical and civilizational progress, which at the turn of the twentieth and twenty-first centuries became a determinant of the information society era, initiated an extremely dynamic development of tools and methods by which information, factography or any other type of data could effectively reach the broadly understood recipient. Due to the increasingly observable „blurring of national borders” in international interpersonal contacts, it was necessary to imple-

ment an advanced technology with a maximally broad spectrum of applications. Artificial intelligence (AI) has proved to be dominant in the last two decades.

This article concerns the issues of the use of artificial intelligence in the educational process. The author, presenting the conditions conducive to the development of this type of methodology at all levels of teaching, also characterized the most important AI systems used in educational institutions as elements supporting education and assessment of students' competences. An interesting aspect raised in the publication is the issue of the limits of human trust in the credibility of such modern methods in terms of their rationality and infallibility.

The aim of the article is to prove that with globalization and the increasingly intensive development of IT techniques, education and higher education are becoming a very good base for the implementation of more and more perfect IT systems, which, based on algorithms and mathematics, are becoming an increasingly socially reliable tool for the activities of this type of institutions. Nevertheless, obtaining the status of a standard requires a lot of investment.

Słowa kluczowe: sztuczna inteligencja, zaufanie społeczne, szkolnictwo, społeczeństwo informacyjne

Keywords: Artificial Intelligence, social trust, education, information society

6.1. Wprowadzenie

Niezwykłe dynamiczny rozwój naukowo-techniczny, jaki obserwujemy od drugiej połowy XX wieku, zapoczątkował głębokie przemiany w większości obszarów życia publicznego, a w tym również w procesie nauczania. Stosowane dotychczas metody oparte w większości na systemie „nauczyciel–uczeń” czy „profesor–student”, stopniowo ustępują pod wpływem powszechnie wdrażanej informatyzacji i internetyzacji. W ostatniej dekadzie można zaobserwować znaczący wzrost implementacji rozwiązań z pośrednim lub bezpośrednim wykorzystaniem tzw. systemów sztucznej inteligencji (SI lub wymiennie AI – akronim nazwy w języku angielskim *Artificial Intelligence*). Pojęcie sztucznej inteligencji zostało po raz pierwszy wprowadzone przez Johna McCarthy'ego w 1955 r. [Rożanowski, 2007, s. 12]. Współcześnie sztuczna inteligencja stanowi dział informatyki, którego przedmiotem badań są reguły rządzące inteligentnymi zachowaniami człowieka, takimi jak np. uczenie się, postrzeganie zjawisk, posługiwanie się abstrakcją czy też tworzenie modeli formalnych [Dudek, 2012]. Interesujące podejście reprezentował Ferguson, który ograniczał pojęcie sztucznej inteligencji do zdolności uczenia się. Swoją definicję uzasadniał tym, że sama inteligencja jest zdolnością do transformacji informacji na poziomie koncepcyjnym, które charakteryzują się znamionami abstrakcji [Rożanowski, 2007, s. 110].

Omawiane dalej zagadnienia koncentrują się wokół uwarunkowań sprzyjających implementacji sztucznej inteligencji w procesie kształcenia, ze szczególnym uwzględnieniem systemów służących podnoszeniu efektywności na różnych szczeblach nauczania, poczynając od oświaty podstawowej, a na szczeblu szkolnictwa wyższego kończąc. Do kluczowych technik stosowanych w tym obszarze należą m.in. systemy eksperckie (SE), sieci neuronowe czy też metody oparte na logice rozmytej. Pierwsze z opisanych wykorzystywane są najczęściej na szczeblu edukacyjnym i mogą pełnić zarówno rolę doradczą, jak i funkcję prowadzącego zajęcia. Dodatkowo inteligentne systemy informatyczne cechuje możliwość przekazywania odpowiedniej wiedzy na podstawie obserwacji ucznia i stawianie diagnozy co do przyswajalności informacji [Wieczorkowski, 2007, s. 103]. Zastosowanie sieci neuronowych podczas zdobywania wiedzy odnosi się do analizy obrazów, ich kompresji i dalszego przetwarzania. Istotna w procesie kształcenia na poziomie akademickim okazała się opracowana w 1965 r. przez Lotfi Zadeha metoda logiki rozmytej. Zastosowano ją na gruncie SI tam, gdzie zastosowanie logiki klasycznej nie przynosiło pożądanego rezultatu [Rożanowski, 2007, s. 121].

W niniejszym rozdziale, oprócz charakterystyki zastosowań sztucznej inteligencji w obszarze szeroko rozumianego kształcenia, została podjęta również próba analizy aktualnego stanu wdrażania metod SI w instytucjach oświatowych i akademickich. Omówiono także problem zaufania społeczeństwa do zachodzących zmian w nauczaniu, implikowanych szybkim rozwojem systemów opartych na sztucznej inteligencji.

6.2. Uwarunkowania sprzyjające stosowaniu sztucznej inteligencji w procesie kształcenia

Postępującemu procesowi wdrażania sztucznej inteligencji w tzw. życie codzienne towarzyszy wiele obaw społeczeństwa, związanych m.in. z możliwością występowania ataków technologicznych narażających bezpieczeństwo obywateli czy prawdopodobieństwem manipulowania za pośrednictwem sztucznej inteligencji jednostkami bądź też całymi zbiorowościami [Torczynska, 2019, s. 108–109]. Nie zmienia to faktu, że obserwowany zakres zastosowań SI w życiu codziennym wraz z upływem czasu staje się coraz szerszy. Jednym z pierwszych był sektor technologiczny wykorzystujący SI do budowy tzw. inteligentnych domów/mieszkań, wyposażonych w zintegrowane systemy zarządzania oświetleniem, w systemy sterowania ogrzewaniem, czy

też kontroli za pomocą aplikacji w telefonie [Walczak, 2017, s. 213–214]. Obszar zastosowań SI w błyskawicznym tempie objął w ostatnich latach niemal wszystkie procesy produkcyjne i urządzenia powszechnego użytku, nie mówiąc już o systemach transportu, ochrony i monitorowania obiektów. Wbrew obrońcom praw człowieka w wielu krajach stosuje się na coraz szerszą skalę rejestrację i rozpoznawanie osób w miejscach publicznych, a telefony komórkowe stały się niemal klasycznym źródłem inwigilacji obywateli w niezwykle szerokim zakresie. Sztuczna inteligencja rewolucjonizuje nasze życie i wydaje się, że wyrażane obawy i zastrzeżenia nie są w stanie zahamować tego procesu [<https://mobiletrends.pl/sztuczna-inteligencja-wkracza-do-naszyc-domow/>].

Niezwykle ciekawym zjawiskiem jest szybkie wkraczanie SI w obszary „nietechniczne”, np. w pracę maklerów giełdowych, prawników bądź lekarzy. Maklerzy są zastępowani przez informatyków, ponieważ większość aktywności związanej z działalnością giełdy została podporządkowana systemom komputerowym, odpowiednio oprogramowanym na bazie SI. Analogicznie przedstawia się sytuacja z wykonywaniem zawodu prawnika, zwłaszcza w kontekście tej grupy, która jest związana z przygotowaniem i oceną umów, regulaminów i różnych aktów prawnych. Jako przykład może służyć popularny program LawGeex, opracowany w 2014 r. Program umożliwia szybką ocenę wprowadzonego aktu prawnego pod różnym kątem (spójność, zgodność z przepisami itp.) [Stylec-Szromek, 2018, s. 503]. Zakres możliwości analiz prawniczych wykonywanych za pomocą tej aplikacji jest zaskakująco szeroki. Również imponujący jest procent efektywności i czas rozstrzygnięcia konkretnych kwestii prawnych. Z przeprowadzonych badań wynika, że dokładność wykonywania zadania przez człowieka wynosi średnio 85%, podczas gdy w przypadku programu LawGeex jest to aż 94%. Dla odmiany rozwiązanie przykładowego problemu prawnego przez człowieka wynosi 92 minuty, podczas, gdy realizacja tego celu przez program komputerowy to jedynie 26 sekund [<https://generacjasmart.pl/2018/05/21/sztuczna-inteligencja-pokonuje-prawnikow/>].

Upowszechnienie sztucznej inteligencji dotyczy także obszaru opieki zdrowotnej. Na przełomie XX i XXI wieku zostało opracowanych wiele programów komputerowych z wykorzystaniem aplikacji służących wynalezieniu i opatentowaniu urządzeń sprzyjających szybszemu diagnozowaniu i efektywniejszemu leczeniu różnych chorób. Istnieją dowody, że dzięki temu średni czas ludzkiego życia ulega wydłużeniu. Według Polskiej Agencji Prasowej, przytaczającej dane z badań dokonanych w Imperial College w Londynie, średnia długość życia kobiet do 2030 r. wyniesie 90 lat, mężczyźni 84 [<https://naukawpolsce.pap.pl/aktualnosci/news%2C413228%2Cbadanie-do-2030-roku-oczekiwana-srednia-dlugosc-zycia-bedzie-rosnac.html>].

Rosnący stopień obecności technologii informatycznych w życiu codziennym spowodował istotne zmiany społeczno-kulturowe, między innymi ukształtował zaplecze do wykorzystywania możliwości SI na różnych szczeblach nauczania, a także w działalności badawczej. W pierwszym etapie powstała dziedzina nauki zajmująca się emulowaniem „inteligentnych zachowań” przez komputery [Schalkoff, 1990]. Dość szybko stało się możliwe tworzenie struktur i programów uczących się, takich jak np. sztuczne sieci neuronowe. Powstały procedury komputerowego rozwiązywania problemów i podejmowania decyzji, poprzedzone analizą danych i uczeniem się systemów SI. Ta swoista dojrzałość techniczna, połączona z niemal nieograniczonym dostępem do sieci internetowej, otworzyła drogę do tworzenia i wdrażania nieznanych dotąd metod nauczania. Temu procesowi sprzyjały także istotne uwarunkowania społeczno-gospodarcze, takie jak rosnące koszty „klasycznej” edukacji, starzenie się społeczeństwa czy też dynamicznie zmieniający się rynek pracy.

Wymienione w skrócie uwarunkowania stały się istotnym czynnikiem motywującym instytucje oświatowe oraz akademickie do coraz częstszego korzystania z dedykowanych systemów SI w procesie nauczania. Impulsem do podejmowania tego rodzaju aktywności na skalę krajową stało się Zalecenie Parlamentu Europejskiego i Rady Europy z 18 grudnia 2006 r. oraz rekomendacje Komisji Europejskiej dotyczącej godnej zaufania sztucznej inteligencji. Zalecenie z 2006 r. odnosiło się do ośmiu kompetencji bezpośrednio powiązanych z edukacją, nazywanych „kompetencjami kluczowymi w uczeniu się przez całe życie” [Schalkoff, 1990]. Rekomendacje natomiast zostały opracowane przez Grupę Ekspertów Wysokiego Szczebla i ogłoszone 26 czerwca 2019 r. w formie trzydziestu trzech zaleceń dla rządów państw członkowskich UE. Wśród zaleceń dotyczących biznesu, inwestycji, regulacji i zarządzania AI oraz wdrażaniem sztucznej inteligencji do ochrony środowiska czy bezpieczeństwa, wiele uwagi poświęcono na kształcenie, edukację z zastosowaniem SI oraz tzw. zarządzanie wiedzą [Zalecenie Parlamentu Europejskiego, 2006]. Nie chcąc pozostawać biernymi, niektóre państwa wdrożyły rządowe strategie wprowadzania sztucznej inteligencji w kształceniu. Przykładem może być Hiszpania, której premier Fedro Sanchez wprowadził w 2020 r. rozwojowo-badawczą strategię B+R oraz badawczo-rozwojową i innowacyjną strategię w zakresie sztucznej inteligencji. Premier uznał program wdrożeniowy sztucznej inteligencji w zakresie kształcenia jako jeden z priorytetowych punktów polityki gabinetu. Sanchez podkreślił, że planując zaangażowanie do realizacji tego założenia jedenastu ministerstw, chciałby wykazać, że innowacyjność uniwersytetów oraz uczelni technicznych stanowi podstawę dla teraźniejszości i przyszłości kraju [Salamon, 2019, s. 20].

Rolę sztucznej inteligencji w kształceniu określa i promuje także powołane w ramach Komisji Europejskiej w roku 1957 Wspólne Centrum Badawcze (*Joint Research Center* – JRC). Węzłowym celem tej organizacji jest wspieranie polityki UE w zakresie zarządzania wiedzą oraz wypracowywania metod, które w innowacyjny sposób służyłyby rozwiązywaniu problemów dotyczących tego zagadnienia [Pietruch-Reizes, 2020, s. 55].

6.3. Systemy SI służące podnoszeniu efektywności kształcenia w różnych obszarach i na różnych szczeblach nauczania

Idea stosowania sztucznej inteligencji w procesie dydaktycznym nie jest nowa – opracowywanie założeń czy też przewidywanie, z jakimi trudnościami wiązałyby się mechanizm popularyzowania tego systemu, sięga lat 60. XX wieku. W następstwie komputeryzacji, informatyzacji i internetyzacji doszło do ukształtowania się społeczeństwa informacyjnego, charakteryzującego się otwartością na stosowanie najnowszych rozwiązań technologicznych i informatycznych w codziennym funkcjonowaniu. W praktyce oznaczało to również akceptację stosowania opisanych rozwiązań SI w różnych dziedzinach życia społecznego, w tym także zastosowania jako aplikacji wspomagających zdobywanie i transfer wiedzy. Do najczęściej stosowanych metod SI wspierających proces kształcenia należą systemy eksperckie (nazywane też ekspertowymi) oparte na teorii zbiorów rozmytych oraz platformach MOOCs. Ważnym elementem wspomagającym stały się sieci neuronowe, służące jako narzędzia do analizy i rozpoznawania mowy, obrazów, kształtów itp.

Geneza prac nad systemami eksperckimi sięga lat 60. i 70. ubiegłego wieku. Prekursorem koncepcji był Edward Feigenbaum z Uniwersytetu w Stanford, który w 1977 r., podając definicję systemu ogłosił jednocześnie możliwość jego wykorzystania jako metody kształcenia. W opinii Feigenbauma system ekspercki to program komputerowy oparty na SI, który ma możliwość wykorzystywania wiedzy oraz procedur wnioskowania, w celu rozwiązywania problemów o różnym stopniu złożoności, przy jednoczesnej obecności eksperta z danej dziedziny. Dzięki temu system komputerowy miałby umożliwić użytkownikowi jak najszybsze uzyskanie oczekiwanej informacji [Rożanowski, 2007, s. 115]. Ujęcie znaczenia systemu eksperckiego w taki sposób zainicjowało szereg badań nad szeroko rozumianym pojęciem inżynierii wiedzy. W konsekwencji rozpoczęto intensywne prace nad wdrażaniem

systemów eksperckich do dydaktyki, ze szczególnym uwzględnieniem edukacji szkolnej. Zasadniczym celem systemów eksperckich było początkowo doradztwo poprzez pełnienie roli „konsultanta” o znacznej elastyczności, opartej na podwójnej umiejętności. Z jednej strony, system ekspercki posiada zdolności przyswajania wiedzy, którą jest w stanie adaptować w zależności od okoliczności i otoczenia, co w efekcie skutkuje swoistą umiejętnością uczenia się. Z drugiej strony, system ekspercki posiada zdolność do samodzielnego wnioskowania i prezentacji odpowiednich danych faktograficznych w wyniku analizy dokonanej na podstawie zebranych wcześniej informacji. Należy zwrócić uwagę, że możliwości działania systemu eksperckiego są uzależnione od wielkości zakodowanej w nim wiedzy.

Przytoczone właściwości SE przyczyniły się do implementacji w obszarze edukacyjnym. Przykładem świadczącym o efektywności zastosowania systemów eksperckich jest e-learning. E-learning jest zaliczany do form tzw. inteligentnego systemu nauczania. Inteligentny system nauczania stanowi element współtworzący stosowaną uniwersalnie koncepcję tzw. Otwartego Nauczania Zdalnego. Jego istota polega na kreowaniu modułów/obiektów nauczania w zakresie określonym przez obszar semantyczny wiedzy z danej dziedziny [Szulc, 2013, s. 1–3]. Jako jedna z form inteligentnego systemu nauczania e-learning daje możliwość wspomagania pracy nauczyciela dzięki technologii wykorzystanej w konstrukcji tego programu. W systemie najczęściej wykorzystuje się metody oparte bezpośrednio na sztucznej inteligencji, w tym systemy eksperckie, sieci neuronowe, algorytmy genetyczne, inteligentni agenci bądź systemy rozmyte. Nauczanie z zastosowaniem e-learningu korzysta z nowoczesnych systemów telekomunikacyjnych, które zapewniają komunikację między użytkownikiem a systemem. Wykorzystywane są także narzędzia synchroniczne i asynchroniczne, gwarantujące kontakt między uczniami z różnych stref czasowych [Szulc, 2013, s. 2–3].

Analiza systemów eksperckich dowodzi, że cechuje je szerokie spektrum możliwości w aspekcie zastosowań bezpośrednio do poszczególnych dziedzin i przedmiotów. Przykładem zajęć, w których posługiwanie się e-learningiem przynosi często wymierne efekty, są języki obce. System ekspercki współtworząc model inteligentnego nauczania, stał się w tym przypadku bazowym elementem systemu ITELS (ang. *Intelligent Computer-Assisted Language Learning*), opracowanego specjalnie na potrzeby osób zamierzających podnieść poziom znajomości języka obcego w zakresie rozumienia tekstów oraz pisanie i wypowiedzi. ITELS powstał w 1997 r i stanowi interesujący przykład współpracy między osobą uczącą się a programem komputerowym. Istota systemu polega na wykorzystywaniu trzech rodzajów nauczania: innowacji, współpracy oraz inicjatywy ucznia. Program ma całkowitą kontrolę nad postępem

pami w nauce oraz dostosowywaniem poziomu dydaktycznego do możliwości ucznia. Wprawdzie osoba kształcąca może wybrać treści i formy uczenia się, ale przedstawia je programowi do akceptacji lub odrzucenia. Inicjatywa w procesie kształcenia należy jednak do ucznia – to on określa cel kształcenia, plan wykonawczy i wskazuje nauczycielowi kryteria oceny [Szulc, 2013, s. 4–5]. Innym, nowszym systemem SI do nauki języków obcych jest aplikacja Duolingo. Przy odpowiednim wykorzystaniu tej aplikacji istnieje możliwość dostosowania poziomu nauczania do umiejętności osoby uczącej. Należy nadmienić, że system jest skonstruowany w taki sposób, by zapewnić rozmowę osoby kształcącej się z internetowym nauczycielem [<https://nextrope.com/pl/ai-w-edukacji-czyli-nauczanie-na-miare-xxi-wieku/>].

W ramach inteligentnego systemu nauczania zostało opracowanych wiele komputerowych programów dziedzinowych funkcjonujących na bazie systemów eksperckich. Na uwagę zasługuje system MathiTs do przyswajania wiedzy z matematyki. Osoba kształcąca się loguje się do programu przez specjalny moduł administracyjny, kontrolujący dane personalne użytkownika. Wiedza dotycząca wybranych zagadnień została sformułowana w formie pytań i odpowiedzi lub wykładu. Podstawę technologiczną stanowi system LaTeX, co umożliwia konwersję materiału merytorycznego na format PDF. Warto zwrócić uwagę na wewnętrzną konstrukcję systemu MathiTs. Dzieli się on na cztery moduły: eksperta, studenta, pedagogiczny, komunikacyjny. Moduł ekspercki jest traktowany jako system informatyczny umożliwiający rozwiązywanie problemów z danej dziedziny za pomocą procedur wnioskowania i posługujący się zakodowaną wiedzą dziedzinową na poziomie eksperta. Tym samym zastępuje pracę eksperta–człowieka [Wieleba, 2011, s. 199]. Drugi moduł – służący do obsługi studenta – ma integralne znaczenie dla funkcjonowania całego systemu MathiTs. To właśnie w nim zbierane są najważniejsze informacje obejmujące wszystkie aspekty dotyczące osoby pobierającej naukę – od oceny zachowania (w przypadku uczniów), poprzez katalog uzyskiwanych ocen, a na gromadzeniu danych na temat efektywności tempa przyswajania wiedzy kończąc. Zasadniczym celem modułu pedagogicznego jest dobór odpowiednich strategii kształcenia do potrzeb osoby zdobywającej wiedzę. Moduł pedagogiczny umożliwia również komunikację z uczącymi się, w celu doboru odpowiednich sekwencji pytań i odpowiedzi. Czwartym modułem jest kanał komunikacyjny odpowiedzialny za zachowanie łączności między uczącymi się a komputerem. Do tego celu służą interfejsy, do których zaliczamy interfejs ucznia, nauczyciela i administratora. Interfejs prowadzącego jest wspomagany przez program LaTeX, dzięki czemu zmiana formatów plików, kodowanie przekazywanych materiałów dydaktycznych itp. stają się bezproblemowe [Szulc, 2013, s. 9].

Innym rozwiązaniem informatycznym, pozwalającym skutecznie wykorzystywać SI w edukacji szkolnej w zakresie matematyki, jest system iTalk2Learn. Istota jej działania jest analogiczna do wcześniej opisanej platformy, z wyjątkiem dołączonego sekwensera. Sekwenser ma za zadanie rejestrować zachowania ucznia oraz jego możliwości naukowe w dziedzinie matematyki. Na tej podstawie sekwenser dostosowuje poziom lekcji do możliwości percepcyjnych uczącego się [<https://nextrope.com/pl/ai-w-edukacji-czyli-nauczanie-na-miaren-xxi-wieku/>].

Szybki wzrost możliwości obliczeniowych i pojemności pamięci komputerów osobistych umożliwił opracowanie programów do nauczania gry na instrumentach, kompozycji czy też aranżacji. Jako przykład może służyć program inteligentnego systemu nauczania muzyki A. Fentona z Carnegie Mellon University. Pod koniec XX wieku opracowano metody pozwalające przyswajać wiedzę w zakresie muzyki, dotyczącą techniki gry na różnych instrumentach i o różnych stylach, dzięki wykorzystaniu technik SI [Szulc, 2013, s. 9].

Uniwersalność i coraz szerszy dostęp do Internetu sprawiły, że prace nad jego wykorzystaniem przy projektowaniu inteligentnych systemów nauczania na odległość nabrały przyspieszenia w pierwszej dekadzie XXI wieku. Pozytywne doświadczenia zdobyte przy wdrażaniu i rozpowszechnianiu wspomnianych wcześniej programów edukacyjnych opartych na SI stały się poważnym bodźcem do rozwoju systemów nauczania o szerokim zasięgu i na różnych szczeblach, w szczególności na poziomie akademickim. Zamyśl zorganizowania tego rodzaju kształcenia na odległość ugruntował się w 2008 r. w USA, kiedy nabrały tempa prace nad systemem MOOCs (*Massive Open Online Courses*) [Krzątowska, 2013]. Przyczyną przyspieszenia prac nad powołaniem MOOCs do funkcjonowania stały się bardzo wysokie koszty studiów wyższych w USA i w innych państwach, co niejednokrotnie utrudniało młodym ludziom dostęp do wiedzy na poziomie akademickim. Po pokonaniu wszystkich formalnych przeszkód profesor Sebastian Thrun z Uniwersytetu w Stanford (USA) oficjalnie uaktywnił stronę internetową z możliwością wyboru kursów odpowiadających poziomem studiom wyższym, jednakże z tą różnicą, że bezpłatnych. Na pierwszy kurs zgłosiło się 160 tysięcy osób ze 190 krajów, a wykłady były tłumaczone na czterdzieści cztery języki. W 2012 r. The New York Times ogłosił, że będzie to rok MOOCs. Rok później na stronie MOOCs zarejestrowano ponad sto kursów. Od 2013 r. kursy MOOCs były oferowane przez kilka platform, m.in. Udacity, EDX oraz Coursera.

Prowadzenie bezpłatnych kursów akademickich z różnych dyscyplin spotkało się ze znacznym zainteresowaniem zarówno ze strony studentów, jak

i uczelni. Swoje oferty propagowały najlepsze światowe uniwersytety, a do partycypacji w kursach zgłaszało się coraz więcej chętnych. Uczestnictwo on-line w kursie MOOCs z założenia było otwarte, udział w zajęciach był nieobowiązkowy i bezpłatny. Dodatkowo nie stwarzano ograniczeń formalnych związanych z wiekiem, czy dotychczasowym udokumentowanym poziomem wiedzy uczestnika takiego kursu. W grupie uniwersytetów i uczelni technicznych prowadzących tego typu kursy znalazły się m.in. Berkeley, Harvard, Massachusetts Institute of Technology, University of Texas, University of Toronto, McGill University, Wellesley College, Delft University of Technology, Ecole Polytechnique Federale de Lausanne, University of Tokyo, University of Edinburgh, Universiteit Leiden, University of Copenhagen, Universitat Autònoma de Barcelona, Munich University oraz Australian National University [Krzętowska, 2013].

Rozpowszechnianie systemu MOOCs jako sposobu na bezpłatne zdobywanie wiedzy akademickiej wiąże się zarówno z problemami o charakterze techniczno-organizacyjnym, jak również coraz częściej stawianym zapytaniem o perspektywę dalszego rozwoju. Jednym z podstawowych problemów jest uruchomienie platform i aplikacji, które gwarantowałyby studentom permanentny dostęp do materiałów dydaktycznych bez straty jakości tekstu, dźwięku czy obrazu. Drugim z kluczowych zagadnień jest uniemożliwienie kształcącym się dokonywania plagiatu lub niesamodzielnego rozwiązywania sprawdzianów wiedzy oraz zapewnienie obiektywnej oceny postępów danej osoby w nauce. nierozwiązaną dotychczas kwestią jest sprawa finansowania działalności MOOCs jako systemu wspomagającego tradycyjne kształcenie uniwersyteckie. Przygotowanie zajęć z wykorzystaniem interfejsu, bloga, mediów społecznościowych, forów dyskusyjnych itp. wymaga od zaangażowanych ekspertów nakładu setek godzin pracy. Uczelnie udostępniając kadrę prowadzącą zajęcia, nie dofinansowują prowadzenia zajęć w systemie MOOCs. Aktualnie jedynym źródłem pozwalającym na sfinansowanie pracy osób zaangażowanych w MOOCs są korporacje typu Google. Należy jednak zauważyć, że często nie są to kwoty pokrywające w pełni ponoszone wydatki.

Pomijając trudności organizacyjne związane z MOOCs, trzeba podkreślić, że jako system e-learningu oparty na systemie eksperckim stał się załącznikiem rewolucji w szkolnictwie wyższym w zakresie możliwości pozyskiwania wiedzy na poziomie akademickim. Warto również zwrócić uwagę na fakt, że MOOCs dzięki posiadaniu zróżnicowanych źródeł informatycznych stał się integralnym „ośrodkiem” pozyskiwania, posiadania oraz wymiany informacji naukowej [Krzętowska, 2013]. Niezależnie od licznych walorów wynikających z uczestnictwa w kursach MOOCs nie można pominąć także aspektów

negatywnych. Za ujemny wpływ MOOCs podaje się zagrożenie dla klasycznego modelu nauczania oraz dla gospodarczego modelu biznesowego uczelni [O'Connor, 2014, s. 623–635].

Techniki sztucznej inteligencji w obrębie kształcenia, jak i powiązanych z tym procesem prac warsztatowych i laboratoryjno-badawczych, są wdrażane na coraz większą skalę. Naukowcy z europejskich uniwersytetów przewidują także, iż w następnych latach stosowanie SI w badaniach i dydaktyce stanie się swoistą normą. Zapowiadane wykorzystanie sieci semantycznych Web 3.0 w niezbyt odległej perspektywie czasowej ma umożliwić osobom studiującym na danym kierunku samodzielne tworzenie ścieżki swojego kształcenia zgodnie z zainteresowaniami i kierunkiem profilu wybranych studiów. Dodatkowo można będzie uzyskać niemal nieograniczony dostęp do informacji w sposób ukierunkowany i precyzyjny [Sarowski, 2017, s. 35].

Charakteryzując metody dydaktyczne, opracowane z wykorzystaniem sztucznej inteligencji, warto zwrócić uwagę na podobieństwa i różnice w zakresie implementacji w oświacie oraz w szkolnictwie wyższym. Zarówno w sferze edukacyjnej, jak i w kształceniu na szczeblu wyższym, wspólnym celem było oczywiście wdrażanie nowych sposobów przekazywania wiedzy za pomocą oprogramowania SI. Zarówno młodzież szkolna, jak i studenci to osoby wykorzystujące maksymalnie najnowsze osiągnięcia branży IT, które mogą w istotny sposób uzupełnić i uatrakcyjnić proces nauczania. Niejednokrotnie bowiem zajęcia prowadzone metodą tradycyjną nie wzbudzały oczekiwanego zaangażowania. Zaimplementowanie sztucznej inteligencji i wdrażanie opartych na niej metod dydaktycznych są szansą na pozytywne zmiany w tym zakresie. Metody stosowane w oświacie to w większości systemy eksperckie. Na ich podstawie opracowano oprogramowania dziedzinowe, co pozwoliło w sposób bardziej efektywny gromadzić dane na temat poziomu i postępu uczniów w nauce, jak i indywidualizować poziom nauczania. W mniejszym zakresie szkolnictwo podstawowe i średnie odwołuje się do koncepcji otwartego nauczania zdalnego, wykorzystującego m.in. technikę e-learningu. Szkolnictwo wyższe wdraża natomiast systemy zdalnego nauczania wykorzystujące SI na platformach MOOCs oraz z wykorzystaniem tzw. chatbotów. Są to zatem metody nieco odmienne od stosowanych na szczeblu oświatowym.

Nie ulega wątpliwości, że wdrażanie technik SI do systemów kształcenia uatrakcyjni proces nauki, umożliwia dostosowanie sposobów realizacji programów do potrzeb osób kształcących się, a w następstwie zwiększa ich szanse na zakwalifikowanie się na dobre studia czy na zdobycie dobrego zawodu.

6.4. Zaufanie społeczne jako podstawa sukcesu edukacji wspomaganej systemami SI

Większość państw członkowskich UE traktuje sztuczną inteligencję i jej szerokie możliwości zastosowań w sektorze technologicznym oraz w życiu publicznym, społecznym i gospodarczym jako system wspomagający zarówno aktualny, jak i przyszły postęp naukowy i cywilizacyjny. Uznawanie za priorytet jakości prowadzonej aktywności – niezależnie od tego, czy będzie dotyczyła przemysłu, edukacji na poziomie szkolnym, czy kształcenia na szczeblu akademickim, stwarzało zdaniem administracji UE podstawę do budowy zaufania społecznego i rozwoju kapitału dziedzinowego [https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence_pl#sztuczna-inteligencja-w-ue].

Implementacja sztucznej inteligencji w edukacji szkolnej ma coraz więcej zwolenników, zyskując tym samym stopniowo zaufanie społeczne. Otwarta na nowoczesne rozwiązania technologiczne oraz IT społeczność UE uznała, że takie rozwiązanie pozwoli dynamiczniej i w szerszym zakresie rozwinąć kompetencje ucznia i nauczycieli. Czynnikiem pozytywnie wpływającym na przekonanie obywateli odnośnie walorów wynikających z wykorzystywania SI do zajęć szkolnych była coraz lepsza i coraz bardziej dostępna infrastruktura i łatwość korzystania z tej technologii. Zwolenników realizacji programów nauczania na bazie SI przekonywała także struktura metodyki opracowana na potrzeby stosowania SI w instytucjach oświatowych. Oprócz opisanej wcześniej metody indywidualizującej pracę ucznia i pozwalającej mu przez to osiągać lepsze wyniki (tzw. metoda adaptacyjna), wdrożono tzw. graf wiedzy. Istotą funkcjonowania grafu wiedzy jest przeanalizowanie toku nauki, dostosowanie zakresu i kolejności przerabiania materiału w zależności od możliwości percepcyjnych osoby uczącej się oraz wykorzystania semantycznej struktury zasobów – czy i w jakim zakresie uczeń opanował dany materiał [<https://innpoland.pl/blogi/piotrpietrzak/125161,sztuczna-inteligencja-w-edukacji-czyli-o-personalizacji-uczenia>].

O wzrastającym poziomie zaufania społecznego do posługiwania się narzędziami sztucznej inteligencji w kształceniu i edukacji świadczy wysoka skala zaangażowania władz ministerialnych państw europejskich, jak również z innych kontynentów, w implementację tej koncepcji do własnych, instytucji oświatowych i akademickich. Na szczeblach edukacji podstawowej i średniej wiele wysiłku poświęcono przeprowadzeniu odpowiednich szkoleń zarówno dla nauczycieli, jak i uczniów oraz studentów, które pozwoliłyby swobodnie posługiwać się technikami nauczania opartymi na SI. Dostosowując się

do koncepcji Unii Europejskiej pod nazwą „Cyfrowa Europa”, niektóre kraje wdrożyły stosowne strategie i programy, pozwalające na ewolucyjne wprowadzenie sztucznej inteligencji w obszarze nauki i przekazywania wiedzy. Reprezentatywnym przykładem jest Wielka Brytania. Zgodnie z ministerialnymi rekomendacjami na poziomie podstawowym i średnim zainicjowano kursy informatyczne dla ośmiu tysięcy nauczycieli, tak aby w każdej szkole była zatrudniona przynajmniej jedna osoba potrafiąca biegle korzystać z techniki SI. Ponadto rozpoczęto liczne szkolenia informatyczno-obliczeniowe pod nazwą Sage Future Makers Lab dla 150 wybitnie uzdolnionych dzieci poniżej 18 roku życia. Zakłada się również powołanie do funkcjonowania Narodowej Uczelni Kompetencji Cyfrowej, której zadaniem byłby wykształcenie kilku tysięcy studentów w dziedzinach związanych pośrednio lub bezpośrednio z IT. Analogiczne działania, choć na nieco mniejszą skalę, wprowadziły Niemcy, decydując się na zmianę w programach nauczania na poziomie podstawowym i średnim w zakresie informatyki. Dotychczas systematyczna nauka informatyki była obowiązkowa dla dzieci i młodzieży tylko w czterech landach. Ze względu na wdrażanie sztucznej inteligencji jako oprogramowania wspomagającego nauczanie informatyka stała się obecnie obowiązkowa w całym kraju. Założono, że w ten sposób zostanie przygotowany pakiet umiejętności, które będą argumentem za całkowitym wdrożeniem SI do nauczania poszczególnych przedmiotów. Według założeń władz niemieckich proces wprowadzenia sztucznej inteligencji w dydaktyce na wyższych uczelniach powinien polegać na wieloetapowym włączaniu SI do programów nauczania w placówkach akademickich. Podkreślano zarazem konieczność powoływania tzw. architektów SI w każdej uczelni. Zadaniem takiej osoby miałyby być wykorzystanie wiedzy o SI do bezpośredniego łączenia jej z niektórymi podmiotami gospodarczymi. W ten sposób łączenie wiedzy naukowej z branżą gospodarczą dałoby możliwość na nawiązanie współpracy przy projektach z wykorzystaniem SI [Założenia do strategii AI w Polsce, 2019, s. 99].

Porównując tempo przenikania technologii sztucznej inteligencji do instytucji kształcących między państwami europejskimi a Ameryką Północną (ze szczególnym uwzględnieniem USA), na uwagę zasługuje fakt, że w Stanach Zjednoczonych działania w tym zakresie podjęto już w pierwszej dekadzie XXI wieku, około dziesięć lat przed Europą. Opracowując od 2007 roku plany rozwoju kierunków kształcenia z wykorzystaniem SI w USA, położono szczególny nacisk na przygotowanie nowych kadr do pracy w sektorach IT i do wprowadzania innowacji technologicznych opartych na SI. Wszystkie działania miały aprobatę agencji rządowych, uzależniających swoje działania od poparcia społecznego w danej kwestii. Ponad dekada różnicy we wprowadzaniu SI do działalności gospodarczej i edukacyjnej sprawiły, że USA znacznie szybciej niż Europa uzyskały wysoki poziom technologiczny w zakresie SI,

wraz z odpowiednim wzrostem zaufanie społeczne do tego typu transformacji. Unia Europejska, angażując się w działania na rzecz rozwoju sztucznej inteligencji, przygotowała dopiero w ostatnich latach założenia prawno-formalne w zakresie edukacji na każdym szczeblu z zastosowaniem SI. Przedstawiciele UE opracowali tzw. Białą Księgę w sprawie sztucznej inteligencji, która zawiera wielodzielnicową strategię rozwoju na kolejne dekady. Zawarte w Białej Księdze założenia zostały aktualnie przedstawione do konsultacji społecznych, co oznacza brak wypracowania jednoznacznego stanowiska dotyczącego postawy społecznej wobec SI, w tym również w zakresie jej implementacji do instytucji edukacyjnych i akademickich [Ishkowski, 2019].

Poruszając problematykę zaufania społecznego do SI w ujęciu globalnym, warto przeanalizować pod tym względem wybrane przykłady państw azjatyckich. Za ogólnoswiatowego pioniera badań naukowych i technicznych oraz rozwoju w tym kierunku od wielu lat uchodzi Japonia. Adekwatnie do postępu cywilizacyjnego i technicznego rząd japoński silnie zainwestował w projekty badawcze dotyczące sztucznej inteligencji. Priorytetowe potraktowanie kształcenia masowego z zastosowaniem systemów SI wynikało z braku specjalistów w obszarze wysoko rozwiniętych technologii. Oszacowano, że w przeciągu ostatnich pięciu lat zostało wykształconych ok. 30 tysięcy inżynierów ze specjalistycznych technologii cyfrowych i 150 tysięcy z umiejętności ogólnych IT. Reforma edukacji ukierunkowana na wypracowywanie kreatywności uczących się oraz ich zdolności interpersonalnych przyczyniła się także do wzrostu poziomu zaufania społecznego do instytucji oświatowych stosujących techniki SI w nauczaniu [Założenia do strategii AI w Polsce, 2019, s. 97].

Promowanie i intensywne wdrażanie sztucznej inteligencji w procesy nauczania ma miejsce również w Chinach. W obszarze szkolnictwa podstawowego i średniego w pierwszej kolejności skoncentrowano się nad tym, by system SI posłużył jako metoda kontroli zachowania się uczniów. W dalszej kolejności natomiast podjęto próbę zaimplementowania SI do programów nauczania z uwzględnieniem poziomu umiejętności uczniów co do obsługi systemu. Szkolnictwo wyższe zostało potraktowane jako zaplecze do przygotowania przyszłej kadry fachowców w zakresie IT. Prym wiodły tutaj uniwersytety, które zgodnie z założeniem chińskiego Ministerstwa Edukacji Narodowej do 2030 r. powinny stanowić filar budowy przyszłych centrów sztucznej inteligencji. Aktualnie ponad siedemdziesiąt uniwersytetów i uczelni wyższych prowadzi zajęcia z zastosowaniem SI a dwieście osiemdziesiąt trzy posiada licencję na układanie oferty programowej z przedmiotami z tego obszaru wiedzy. Promowane są także badania nad interdyscyplinarnym podejściem do SI, jako płaszczyzny umożliwiającej współpracę w wielu obszarach nauki [Założenia do strategii AI w Polsce, 2019, s. 102–103].

W porównaniu z poziomem implementacji sztucznej inteligencji w UE zarówno USA, jak i kraje azjatyckie, wykazywały i nadal wykazują większą dynamikę tego procesu. Wpływa to na poziom zaufania społecznego do SI, który w Chinach i USA jest na tyle znaczący, że może zadecydować o zaawansowaniu technologii stosowanej w szkolnictwie powszechnym i kształceniu na poziomie uczelni wyższych w skali globalnej oraz szeregu innych sektorów życia polityczno-gospodarczego [Lee, 2019]. Przewaga USA i Chin nad krajami Unii Europejskiej w zakresie zaawansowania technologicznego dotyczącego zastosowania SI w nauczaniu oraz w zakresie poziomu zaufania społecznego w tej kwestii ma szereg uwarunkowań. Do najistotniejszych zalicza się sprawę jakości zapewniającej właściwe efekty oraz zapewnienie, że we wszystkich dziedzinach, gdzie znajdzie zastosowanie sztuczna inteligencja zostanie zagwarantowane człowiekowi bezpieczeństwo w aspekcie etycznym, prawnym i wolności osobistej [https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence_pl].

Przedstawione powyżej fakty dowodzą, że sztuczna inteligencja jest odbierana jako globalne wsparcie technologiczne w dominującej części dziedzin współtworzących ludzką egzystencję. Szeroko pojmowane szkolnictwo odgrywa na tle cywilizacyjnej transformacji rolę szczególną. Stanowi zaplecze naukowe, w którym wiedzę i kompetencje uzyskuje kadra przygotowywana do pracy na gruncie gospodarki i przemysłu opartych na technikach SI.

6.5. Zakończenie

Od połowy XX wieku rozpoczęto intensywne prace nad zdefiniowaniem i realizacją koncepcyjną sztucznej inteligencji, a następnie skonkretyzowaniem zakresu jej zastosowań w bezpośrednim odniesieniu do działalności człowieka. W rezultacie podjętych badań wykazano, że sztuczna inteligencja mogłaby wspomagać ludzką aktywność w wielu obszarach życia polityczno-społecznego i gospodarczego. Realizacja tej idei następowała powoli, głównie ze względu na ograniczone możliwości technologiczne i ograniczone zaufanie społeczne. Przyspieszenie nastąpiło pod koniec XX w. wraz z szybkim rozwojem sieci internetowych i postępującym rozwojem społeczeństwa informacyjnego. Niemal jednocześnie pojawiła się konieczność wykształcenia licznej kadry osób posiadających kompetencje do obsługi urządzeń funkcjonujących na bazie SI. Braki kadrowe oraz rosnące koszty kształcenia metodami „klasycznymi” stały się bodźcem do zmian w systemach edukacji na wszystkich szczeblach, w szczególności z wykorzystaniem możliwości sztucznej inteligencji. W ten sposób do szeregu form aktywności człowieka, które zostały udoskonalone przez wspo-

maganie systemami SI, dołączyło również szkolnictwo, jako wieloszczeblowa „kuźnia” edukacji i kształcenia poprzez szeroko dostępny e-learning.

Analiza metodyki prowadzenia zajęć dydaktycznych przy jednoczesnym posługiwaniu się sztuczną inteligencją wykazała w tym obszarze znaczący potencjał. W wyniku prowadzonych badań i obserwacji rezultatów nauczania stwierdzono, że w edukacji na poziomach podstawowym i średnim najbardziej efektywne okazują się systemy eksperckie oraz inteligentny system nauczania. Pierwsze umożliwiały zindywidualizowanie wymogów programowych do zdolności ucznia, inteligentny system nauczania natomiast pozwalał na opracowanie dziedzinowych systemów opartych na SI do nauczania poszczególnych przedmiotów. Dla studiujących zaproponowano różnego rodzaju techniki i platformy e-learningu, chatbotów oraz MOOCs. Wdrożone systemy nauczania z wykorzystaniem SI oraz sieci internetowych okazały się nie tylko ważnym suplementem dydaktyki, ale również „pełnoprawnymi” technikami zdobywania wykształcenia, szczególnie w przypadku gdy z różnych powodów nie można podjąć nauki w trybie stacjonarnym. Na podstawie dokonanych analiz można przyjąć, że w większości krajów koncepcja implementacji SI do systemu nauczania spotkała się z przychylnością, jako perspektywiczna i efektywna. Zróżnicowany poziom zaawansowania technologicznego, akceptacji środowiskowej i potencjału kadrowego sprawił, że dynamika wprowadzania tak nowatorskiego modelu kształcenia była zdywersyfikowana, co przekładało się na ocenę globalnego wymiaru zaufania społecznego w obszarze nauczania SI na poziomie szkolnym i akademickim. Tym niemniej większość rządów podjęła działania w kierunku szerokiej implementacji systemów SI we wszystkich szczeblach edukacji. Potencjał tych nowych technik kształcenia został w pełni uwidoczniiony i wykorzystany w okresie pandemii COVID-19. Nie ulega wątpliwości, że bez nich szkolnictwo na wszystkich poziomach i we wszystkich krajach uległoby w czasie epidemii silnej destrukcji z poważnymi konsekwencjami społecznymi.

BIBLIOGRAFIA

- O'Connor K., (2014), *MOOCs, Institutional Policy and Change Dynamics in Higher Education*, Higher Education, 68(5), s. 623–635.
- Dudek G., (2012), *Sztuczna inteligencja. Materiały pomocnicze do wykładu*, Wykład 1, Politechnika Częstochowska, Wydział Elektryczny.
- Iszkowski W., (2021), *Marketing Sztucznej Inteligencji*, CRN, <https://crn.pl/wywiady-i-felietony/marketing-sztucznej-inteligencji> (dostęp: 20.10.2022).
- Krzętowska A., (2013), *MOOCs wyzwanie dla szkolnictwa wyższego*, [w:] Інституційні чинники розвитку підприємницьких структур, Непочатенко О.о (red.), Видавничо-поліграфічний центр „Візаві”.

- Lee K.F. (2019), *Inteligencja sztuczna, rewolucja prawdziwa. Chiny, USA i przyszłość świata*, Media Rodzina.
- Pietruch-Reizes D., (2020), *Transfer wiedzy naukowej. Kontekst sztucznej inteligencji, internetu rzeczy i robotyki (wybrane zagadnienia)*, [w:] *Informacja, wiedza, innowacje*, W. Babik, D. Pietruch-Reizes (red.), Wydawnictwo Biblioteki Jagiellońskiej, Kraków.
- Rożanowski K., (2007), *Sztuczna inteligencja. Rozwój, szanse, zagrożenia*, *Zeszyty Naukowe Wyższej Szkoły Informatyki*, 2, s. 109–135.
- Sarowski Ł., (2017), *Od Internetu Web 1.0 do Internetu Web 4.0 – ewolucja przestrzeni komunikacyjnych w globalnej sieci*, *Rozprawy Społeczne*, 11(1), s. 32–39.
- Schalkoff R.J., (1990), *Artificial Intelligence: An Engineering Approach*, McGraw-Hill College.
- Stylec-Szromek P., (2018), *Sztuczna inteligencja. Prawo, odpowiedzialność, etyka*, *Zeszyty Naukowe Politechniki Śląskiej, Seria: Organizacja i Zarządzanie*, nr 123, Gliwice.
- Salamon J., (2019), *Sztuczna inteligencja, jako element rozwoju personalnego oraz grupowego Hiszpanii i Meksyku*, *Rynek–Społeczeństwo–Kultura*, 32(1), Nr 1(32), s. 20–24.
- Szulc J., (2013), *Systemy ekspertowe w nauczaniu na odległość*, *Elektroniczne Czasopismo Biblioteki Uniwersytetu Pedagogicznego w Krakowie*, 3/2013, s. 1–21.
- Torczyńska M., (2019), *Sztuczna inteligencja i jej społeczno-kulturowe implikacje w codziennym życiu*, *Kultura i Historia*, 36(2), s. 106–126.
- Walczak T. (red.), (2017), *Inteligentna sieć. Algorytmy przyszłości*, Helion, Gliwice.
- Wieczorkowski K., (2007), *Systemy ekspertowe w edukacji*, *Zeszyty Naukowe Wyższej Szkoły Informatyki*, 6(1), s. 102–121.
- Wieleba R., (2011), *Inżynieria wiedzy w systemach ekspertowych*, *Zeszyty Naukowe Wyższej Szkoły Informatyki*, 5, s. 195–216.
- Zalecenie Parlamentu Europejskiego i Rady Europy z dnia 18 grudnia 2006 w sprawie kompetencji kluczowych w uczeniu się przez całe życie*, Dz. U. Unii Europejskiej, 2006/96 WE.
- Założenia do strategii AI w Polsce. Plan działań ministerstwa cyfryzacji*, (2018), Ministerstwo cyfryzacji, file:///C:/Users/Pawe% C5% 82/Downloads/Za% C5% 82o% C5% BCenia_do_strategii_AI_w_Polsce_-_raport.pdf (dostęp: 20.10.2022).

Źródła internetowe

- <https://mobiletrends.pl/sztuczna-inteligencja-wkracza-do-naszyc-domow/> (dostęp: 17.05.2021).
- <https://generacjasmart.pl/2018/05/21/sztuczna-inteligencja-pokonuje-prawnikow/> (dostęp: 17.05.2021).
- <https://naukawpolsce.pap.pl/aktualnosci/news%2C413228%2Cbadanie-do-2030-roku-oczekiwana-srednia-dlugosc-zycia-bedzie-rosnac.html> (dostęp: 17.05.2021).
- [<https://EuropeanCommission.eu/Policy-and-investment-recommendations-for-trustworthy-Artificial-Intelligence/> (dostęp: 28.05.2021).
- <https://nextrope.com/pl/ai-w-edukacji-czyli-nauczanie-na-miare-xxi-wieku/> (dostęp: 28.05.2021).
- https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence_pl#sztuczna-inteligencja-w-ue (dostęp: 28.05.2021).
- <https://innpoland.pl/blogi/piotrpietrzak/125161,sztuczna-inteligencja-w-edukacji-czyli-o-personalizacji-uczenia> (dostęp: 29.05.2021).
- https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/excellence-trust-artificial-intelligence_pl (dostęp: 1.06.2021).

Rozdział 7

Sztuczna inteligencja jako (nie)skuteczne narzędzie w walce z mową nienawiści

Artificial intelligence as an (in)effective tool in the fight against hate speech

Marcin Rojszczak

Politechnika Warszawska

Wydział Administracji i Nauk Społecznych

Streszczenie. W ostatnich latach – wraz z rozwojem systemów uczenia maszynowego – coraz więcej uwagi poświęcane jest możliwości wykorzystania do walki z mową nienawiści nowej generacji systemów algorytmicznych. Sztuczna inteligencja prezentowana jest jako przełomowa technologia, zdolna do rozwiązania wielu istniejących problemów z nieznaną dotychczas precyzją i skutecznością. Jednocześnie jednak systemy tego typu – chociaż coraz szerzej wykorzystywane przez dostawców usług cyfrowych – nie wpłynęły w zauważalny sposób na ograniczenie skali publikowania treści krzywdzących.

W niniejszym rozdziale charakterystyka działania oraz możliwości systemów uczenia maszynowego omówione zostaną z perspektywy realizacji konkretnego zadania, jakim jest przeciwdziałanie dystrybucji mowy nienawiści. Szczególna uwaga zwrócona zostanie na zidentyfikowanie barier technicznych i regulacyjnych, które wymagają przezwyciężenia, aby cel jakim jest eliminacja treści nienawistnych z Internetu mógł być osiągnięty.

Abstract. In recent years – with the development of machine learning systems – more and more attention has been paid to the possibility of using a new generation of algorithmic systems to combat hate speech. Artificial intelligence is presented as a breakthrough technology, capable of solving many existing problems with previously unknown precision and effectiveness. At the same time, however, such systems – although increasingly used by digital service providers – have not had a noticeable impact on reducing the scale of publishing hurtful content.

In this chapter, the performance characteristics and capabilities of machine learning systems will be discussed in regard of performing the specific task of countering the distribution of hate speech. Particular attention will be paid to identifying the technical and regulatory barriers that need to be overcome if the goal of eliminating online hateful content is to be achieved.

Słowa kluczowe: sztuczna inteligencja, mowa nienawiści, fake news, filtrowanie treści

Keywords: artificial intelligence, hate speech, fake news, content moderation

7.1. Wprowadzenie

W powszechnym odbiorze, zaawansowane systemy algorytmiczne pozwalają rozwiązywać skomplikowane problemy z nieznaną wcześniej precyzją i szybkością. Dzięki temu doskonale nadają się do realizacji zadań wymagających analizy dużej ilości danych. Co więcej, dzięki możliwości pracy ciągłej – systemy tego typu mogą stale doskonalić swoje algorytmy przetwarzania, czerpiąc wiedzę także z oceny trafności swoich wcześniejszych decyzji. Te dwie cechy – zdolność do budowania własnych wniosków oraz umiejętność reinterpretacji wcześniejszych wyników przetwarzania na podstawie nowych danych – charakteryzują systemy algorytmiczne najnowszej generacji, zaliczane do kategorii systemów uczenia maszynowego [Alpaydin, 2016]. Często nazywa je się także, nie zawsze trafnie, systemami sztucznej inteligencji (ang. *artificial intelligence*, AI). Abstrahując od dyskusji na temat definicji AI (por. [Franklin, 1995]), termin ten będzie stosowany w niniejszym rozdziale dla opisanego nowoczesnych systemów bazujących na uczeniu maszynowym i analizie dużych zbiorów danych.

Systemy tego typu z sukcesem stosowane są w zastosowaniach wymagających kategoryzacji (klasyfikowania) informacji wejściowych na podstawie zidentyfikowanych wcześniej nieoczywistych powiązań lub trendów. W efekcie znajdują zastosowanie w tak różnych obszarach, jak diagnostyka obrazowa pacjentów (identyfikacja zmian chorobowych), marketing behawioralny (wskazywanie preferencji zakupowych użytkowników) czy rozpoznawanie obrazów (np. twarzy). W każdym przypadku warunkiem niezbędnym prawidłowego działania algorytmu jest przeanalizowanie bardzo dużej ilości danych treningowych, często o różnicowanej jakości, dzięki czemu możliwe jest zbudowanie modelu analitycznego wykorzystywanego do przetwarzania danych rzeczywistych.

Sposób działania systemów AI powoduje, że wydają się one dopasowane do potrzeb związanych z realizacją zadań wymagających stałego i hurtowego przetwarzania danych pochodzących z mediów online, na przykład w celu identyfikacji treści bezprawnych. Nic dziwnego zatem, że zdecydowana większość treści codziennie usuwanych przez wiodące portale społecznościowe bazuje na ocenie dopuszczalności ich publikacji w której kluczową rolę odgrywają systemy algorytmiczne [Gillespie, 2020]. Sam tylko Twitter informuje, że każdego dnia publikowanych jest średnio 100 mln postów użytkow-

ników – zadanie to, już tylko z uwagi na jego skalę, nie mogłoby być równie efektywnie wykonywane manualnie przez ludzi.

Dostawcy usług cyfrowych od lat są prawnie zobowiązani do przeciwdziałania dystrybucji różnych kategorii treści bezprawnych – do których w szczególności zaliczyć można publikacje naruszające prawa autorskie, ale także treści związane z poważną przestępczością (np. terroryzmem czy pedofilią) [Cobbe, 2020]. Prewencyjny charakter obowiązków przejawia się nie tylko w obowiązku reagowania na otrzymywane od użytkowników zgłoszenia, ale przede wszystkim w zapobieganiu dalszemu rozpowszechnianiu bezprawnych treści oraz w uniemożliwianiu publikacji materiałów, których bezprawność jest oczywista. W tym celu dostawcy usług online coraz szerzej stosują mechanizmy wstępnego filtrowania (tzw. *upload filters*), pozwalające na ocenę treści z wykorzystaniem opracowanego modelu analitycznego w sposób prewencyjny, a więc jeszcze przez jej opublikowaniem [Romero Moreno, 2020].

Inny obszar, w którym mechanizmy tego typu są stosowane, to analiza wiadomości publikowanych przez użytkowników pod kątem identyfikacji przypadków mowy nienawiści. Lawinowy wzrost wypowiedzi krzywdzących, rozpowszechnianych za pomocą usług online to zjawisko, które od lat prowadzi do wzrostu niezadowolenia społecznego oraz oczekiwania podjęcia bardziej zdecydowanych działań od prawodawców oraz samych dostawców usług cyfrowych. Potrzeba podjęcia skuteczniejszej walki z przypadkami mowy nienawiści doprowadziła także do upowszechnienia się wykorzystania w tym obszarze systemów uczenia maszynowego. Oczekiwano, że dzięki zastosowaniu nowoczesnych algorytmów potrafiących wiernie interpretować język naturalny oraz analizować olbrzymie bazy referencyjne (miliony wypowiedzi publikowane przez użytkowników każdego dnia), możliwe stanie się identyfikowanie i eliminowanie szkalujących treści w skali nieosiągalnej dla stosowanej wcześniej analizy manualnej.

Chociaż według zestawień publikowanych przez wiodące koncerny technologiczne systemy AI odpowiadają za usuwanie większości treści bezprawnych¹, to jednocześnie pojawia się coraz więcej wątpliwości wskazujących na ograniczoną przydatność tego typu systemów do identyfikowania przypadków mowy nienawiści. Według Koeber i Cox [2018], systemy algorytmiczne Facebooka odpowiadały w 2018 roku za identyfikację 99,5% treści terrorystycznych, 95% pornografii, 86% treści prezentujących przemoc, ale tylko 38% przypadków mowy nienawiści. Okazuje się zatem, że o ile systemy AI sprawnie radzą sobie w przeciwdziałaniu dystrybucji innych kategorii treści bezprawnych, ich skuteczność w prawidłowym identyfikowaniu wypowiedzi krzywdzących jest niższa od oczekiwanej.

¹ Według danych Facebook dotyczących pierwszego kwartału 2022 tak dzieje się w ponad 98% przypadków: <https://cli.re/kAePaq> (dostęp 4.07.2022).

W niniejszym rozdziale podjęty zostanie problem niespełnionych oczekiwań, że systemy AI pomogą skutecznie rozwiązać problem narastającej fali mowy nienawiści w Internecie – przez niektórych nazywany nawet fałszywą obietnicą AI [Gillespie, 2020]. W tym celu w pierwszej kolejności przedstawione zostaną problemy definicyjne dotyczące węzłowego pojęcia, jakim jest „mowa nienawiści”. W dalszej kolejności omówione zostaną możliwości i ograniczenia funkcjonowania AI, zwłaszcza w kontekście analizy języka naturalnego. Na tej podstawie, w dalszej części rozdziału, przedstawione zostaną najważniejsze problemy implementacyjne, w praktyce ograniczające skuteczność funkcjonowania systemów AI w obszarze identyfikacji wypowiedzi krzywdzących.

7.2. W poszukiwaniu definicji mowy nienawiści

Chociaż termin „mowa nienawiści” jest używany niemal powszechnie dla określenia wszelkich wypowiedzi, które mają charakter szkalujący lub krzywdzący, to w praktyce pojęcie to nie ma swojej definicji prawnej. Wniosek ten nie traci aktualności zarówno na płaszczyźnie prawa unijnego, jak również większości systemów prawnych państw członkowskich.

Wiodącym aktem prawa międzynarodowego, który dzięki mocy wiążącej mógłby służyć harmonizacji środków chroniących przez rozpowszechnianiem mowy nienawiści w Internecie, jest Konwencja Rady Europy o cyberprzestępczości². Stronami Konwencji są wszystkie państwa członkowskie UE, ale także szereg innych państw trzecich – w tym także nienależących do Rady Europy³. Celem przyjęcia Konwencji było zbliżenie przepisów karnych państw–stron w zakresie typizacji najpoważniejszych rodzajów przestępczości komputerowej oraz związanej z wykorzystaniem komputerów. W czasach redagowania treści traktatu (rok 2001) zagrożenia związane z upowszechnianiem mowy nienawiści w Internecie były mało poznane, stąd też nie dziwi, że autorzy Konwencji nie uwzględnili w jego treści odrębnych postanowień ukierunkowanych na przeciwdziałanie temu zjawisku. Problem ten został natomiast dostrzeżony w pracach nad przyjęciem Protokołu Dodatkowego do Konwencji o Cyberprzestępczości⁴. Co do zasady Protokół stanowi jednak odrębną umowę międzynarodową, co powoduje, że nie wszystkie państwa, które ratyfikowały Konwencję

² Konwencja Rady Europy o cyberprzestępczości z dnia 23.11.2001, Dz.U. z 2015 r. poz. 728.

³ Lista państw–stron Konwencji: <https://cli.re/EmwENd> (dostęp 4.07.2022).

⁴ Protokół dodatkowy do Konwencji Rady Europy o cyberprzestępczości dotyczący penalizacji czynów o charakterze rasistowskim lub ksenofobicznym popełnionych przy użyciu systemów komputerowych z dnia 28.01.2003 r., Dz.U. z 2015 r. poz. 730.

są również stronami Protokołu Dodatkowego (dotyczy to np. Stanów Zjednoczonych)⁵. Protokół Dodatkowy nie odnosi się wprost do „mowy nienawiści”, ale nakazuje penalizację rozpowszechniania materiałów rasistowskich i ksenofobicznych – czyli takich, które służą nawoływaniu, popieraniu lub podżeganiu do nienawiści, dyskryminacji lub przemocy przeciw jakiegokolwiek osobie lub grupie osób ze względu na jedną ze zdefiniowanych cech indywidualnych lub przynależność do określonej grupy społecznej, religijnej czy narodowościowej.

Zamknięty katalog grup chronionych powoduje, że nie wszystkie treści o szkalującej treści spełniają definicję przyjętą w Protokole. I tak, dla przykładu, treści ksenofobiczne motywowane przynależnością do grupy religijnej czy narodowej, są objęte definicją wprowadzoną w Protokole Dodatkowym, jednak już agresja kierowana względem członków innych mniejszości (np. seksualnych) nie podlega takiej ochronie. Co więcej, Protokół Dodatkowy nie zobowiązuje również państw–stron do penalizacji innych typów nieakceptowalnych wypowiedzi, w szczególności zniesławiających czy znieważających – jeżeli następuje z powodów innych niż wskazany w jej treści. W tych przypadkach granica dopuszczalnej krytyki w wypowiedzi publicznej jest różnie wyznaczana w poszczególnych modelach prawnych i chociaż można wskazywać na elementy wspólnego standardu europejskiego, to nadal daleko jest do wypracowania konsensusu.

Jednocześnie prawodawcy coraz częściej decydują się na określenie szczegółowych zadań dostawców usług cyfrowych w obszarze przeciwdziałania rozpowszechniania mowy nienawiści, nakładając na nich zarówno obowiązki reagowania na pojawiające się naruszenia, jak również wdrażania systemów algorytmicznych pozwalających na prewencyjne usuwanie treści przed ich zatwierdzeniem do publikacji. W ostatnich latach w UE przyjęto szereg szczegółowych regulacji, które służą realizacji tego zadania – wśród których należy wymienić dyrektywę o handlu elektronicznym⁶, dyrektywę o przeciwdziałaniu wykorzystaniu dzieci⁷, dyrektywę o ochronie praw autorskich w Internecie⁸ czy rozporządzenie o przeciwdziałaniu rozpowszechnianiu treści ter-

⁵ Lista państw–stron Protokołu Dodatkowego: <https://cli.re/AbRwXw> (dostęp 4.07.2022).

⁶ Dyrektywa 2000/31/WE Parlamentu Europejskiego i Rady z dnia 8 czerwca 2000 r. w sprawie niektórych aspektów prawnych usług społeczeństwa informacyjnego, w szczególności handlu elektronicznego w ramach rynku wewnętrznego, Dz.U. z 2000 nr L 178/1.

⁷ Dyrektywa Parlamentu Europejskiego i Rady 2011/93/UE z dnia 13.12.2011 r. w sprawie zwalczania niegodziwego traktowania w celach seksualnych i wykorzystywania seksualnego dzieci oraz pornografii dziecięcej, zastępująca decyzję ramową Rady 2004/68/WSiSW, Dz.U. z 2011 nr L 335/1.

⁸ Dyrektywa Parlamentu Europejskiego i Rady (UE) 2019/790 z dnia 17 kwietnia 2019 r. w sprawie prawa autorskiego i praw pokrewnych na jednolitym rynku cyfrowym oraz zmiany dyrektyw 96/9/WE i 2001/29/WE, Dz.U. z 2019 nr L 130/92.

rorystycznych⁹. Rozproszenie regulacji utrudnia jednak ujednoczenie zasad, jakie powinny być stosowane przez usługodawców. Stąd też wprowadzane są regulacje, które reformując dotychczasowy model dążą także do doprecyzowania uprawnień dostawców usług oraz wzmocnienia nadzoru nad podejmowanymi przez nich działaniami między innymi poprzez ustanowienie nowych środków sprawozdawczych¹⁰. Nie zmienia to jednak faktu, że dopiero w ostatnich latach instytucje UE zwróciły uwagę na potrzebę wypracowania jednej definicji „mowy nienawiści”, dzięki czemu możliwe stanie się podjęcie bardziej skutecznych działań związanych z eliminowaniem tego typu treści z przestrzeni publicznej.

Brak wspólnych norm unijnych stał się zachętą dla państw członkowskich do przyjmowania własnych przepisów krajowych, mających na celu rozbudowanie obowiązków dostawców usług cyfrowych w obszarze eliminacji treści bezprawnych. Pierwszą tego typu regulacją była niemiecka ustawa o poprawie egzekwowania prawa w sieciach społecznościowych (NetzDG)¹¹, nakazującą dostawcom usług hostingowych niezwłoczne – jednak nie dłużej niż w ciągu 24 godzin – usuwanie treści w oczywisty sposób bezprawnych. Ustawa nie wprowadzała przy tym odrębnej definicji *mowy nienawiści*, bazując w tym zakresie na terminie „treści bezprawnych” – oraz odwołując się do istniejących regulacji krajowych w tym zakresie¹². Z kolei przepisami francuskiej ustawy o zwalczaniu mowy nienawiści w Internecie ustanowiono szczególne obowiązki związane z usuwaniem treści propagujących nienawiść lub znieważających na tle rasowym, religijnym, etnicznym czy w związku z orientacją seksualną lub niepełnosprawnością¹³. Podobną drogę – polegającą na enumeratywnym wskazaniu rodzajów treści zakazanych – obrał również prawodawca austriacki¹⁴.

Trudność w wypracowaniu uniwersalnej legalnej definicji mowy nienawiści jest symptomatyczna. Nawet bowiem w gronie państw opartych na tych samych demokratycznych zasadach ustrojowych brakuje wspólnego stano-

⁹ Rozporządzenie Parlamentu Europejskiego i Rady (UE) 2021/784 z dnia 29 kwietnia 2021 r. w sprawie przeciwdziałania rozpowszechnianiu w Internecie treści o charakterze terrorystycznym, Dz.U. z 2021 nr L 172/79.

¹⁰ Por. np. przepisy sekcji III rozporządzenia 2021/784.

¹¹ Gesetz vom 1.09.2017 zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken, BGBl. I 2017, Nr. 61 07.09.2017, s. 3352.

¹² Zob. Art. 1(1)(3) NetDG.

¹³ Loi n° 2020-766 du 24 juin 2020 visant à lutter contre les contenus haineux sur internet (Ustawa nr 2020-766 z 24.07.2020 r. o zwalczaniu mowy nienawiści w Internecie).

¹⁴ Zob. §2(8) projektu ustawy federalnej o środkach ochrony użytkowników na platformach komunikacyjnych; tekst polskojęzyczny dostępny na: <https://cli.re/y3Wmq9> (dostęp 4.07.2022).

wiska w zakresie definicji mowy nienawiści. Nie sposób uznać, że jedynym powodem takiego stanu rzeczy są różnice w zakresie odmiennych oczekiwań społecznych dotyczących poszanowania wolności słowa (a przez to – i granic dopuszczalnej krytyki). W ten sposób można – przynajmniej częściowo – wyjaśnić różnice w podejściu do tego problemu pomiędzy Stanami Zjednoczonymi a Unią Europejską, czy pomiędzy państwami stosującymi system prawa kontynentalnego a *common law* [Becker i in., 2000; Delgado, Stefancic, 2018; Downing, 1999; Waldron, 2012]. To jednak nie wyjaśnia, dlaczego pojęcie to *per se* nie występuje w przepisach krajowych poszczególnych państw członkowskich. *Mowa nienawiści* nie ma legalnej definicji nie tylko w polskim systemie prawnym, ale również w niemieckim czy francuskim. Pojęcie to jest szeroko obecne w mediach i dyskusji publicznej, jednak na gruncie przepisów prawnych nie występuje – a w jego miejsce stosowane są ugruntowane od lat instytucje, takie jak ochrona dobrego imienia, znieważenie czy zniesławienie. Jednak w poszczególnych modelach prawnych instytucje te definiowane są odmiennie, także w zakresie mających zastosowanie procedur prawnych.

W efekcie, pomimo nie słabnącej presji społecznej, skuteczność środków podejmowanych w celu walki z mową nienawiści napotyka na pierwszą, oczywistą barierę – jaką jest różne definiowanie wypowiedzi bezprawnych w poszczególnych państwach. Prowadzi to do potrzeby uwzględnienia przez dostawców usług cyfrowych wielu odmiennych – i często sprzecznych ze sobą – przepisów dla ustalenia zbioru reguł, jakie powinny być stosowane dla oceny publikowanych przez użytkowników wypowiedzi online. Dlatego komentarz o identycznej treści może być uznany za mieszczący się w granicach dopuszczalnej wypowiedzi w Stanach Zjednoczonych, podczas gdy w Unii Europejskiej zostanie uznany za przykład mowy ksenofobicznej.

7.3. Możliwości i ograniczenia AI

Systemy uczenia maszynowego to szeroka kategoria rozwiązań informatycznych, do której zaliczane są różne typy produktów wykorzystujących zaawansowane algorytmy analizy danych. Chociaż mogą się one od siebie znacząco różnić pod wieloma względami (wykorzystanych algorytmów, skali przetwarzania, jakości otrzymywanych wyników itp.), to wszystkie mają jedną fundamentalną cechę, jaką jest zdolność do budowania samodzielnych wniosków, które są następnie wykorzystywane do rozwiązywania określonego typu zadań. W cyklu życia tego typu systemów można wyróżnić co najmniej trzy fazy: projektową (badawczo-programistyczną), tworzenia modelu analitycz-

nego (trenowania/uczenia) oraz eksploatacji i utrzymania. Istotne jest zrozumienie, że model analityczny – który jest wykorzystywany podczas eksploatacji systemu AI – tworzony jest nie przez programistów, ale samodzielnie przez system, w wyniku analizy odpowiednio przygotowanego zestawu danych testowych. Początkowe programowanie wykorzystywane jest do zbudowania modelu analitycznego, a więc zawiera instrukcje potrzebne, aby przetworzyć dane testowe. Programiści zatem nie tyle „uczą” system AI, jak rozwiązywać określony rodzaj problemów, ale poprzez początkowe programowanie wyposażają go w zbiór instrukcji potrzebnych do tego, aby system sam odnalazł (odkrył) najbardziej efektywny sposób poszukiwania rozwiązań. To poprzez analizę danych testowych (treningowych) system AI identyfikuje powiązania pomiędzy faktami, dzięki czemu uczy się, jak prawidłowo przetwarzać dane w przyszłości. W ten sposób system dedykowany do gry w szachy, analizując przebieg milionów wcześniejszych partii, buduje własny model analityczny pozwalający mu określić ruch, jaki sam powinien wykonać w partii, którą sam będzie rozgrywał. Celem modelu analitycznego nie jest odtwarzanie zdarzeń przeszłych – ale dzięki możliwości analizy bardzo wielu podobnych scenariuszy, dostrzeżenie powiązań pomiędzy danymi, co ma pozwolić na bardziej efektywne rozwiązanie określonego typu zadań w przyszłości.

Powyższe prowadzi do kilku ważnych wniosków. Po pierwsze, początkowe programowanie *de facto* ma ograniczony wpływ na skuteczność działania finalnej wersji systemu. Oczywiście, błędy popełnione przez programistów będą propagowały się dalej – wpływając negatywnie na dalsze działanie systemu. Nie w każdym przypadku istnieje jednak prosta ścieżka oceny, czy nieprawidłowe działanie systemu zostało spowodowane błędami popełnionymi na etapie programistycznym, czy wadliwym procesem analizy danych testowych. W praktyce bowiem to trenowanie (uczenie) modelu jest kluczowe dla prawidłowego funkcjonowania systemu. Co oczywiste, powodzenie tego etapu wymaga dostarczenia odpowiednio dużego i różnorodnego zestawu danych testowych, o odpowiedniej jakości (tj. takich, które zostały wcześniej oczyszczone i ustandaryzowane).

Proces trenowania może przebiegać w sposób nadzorowany lub nienadzorowany [Christiano Silva, Zhao, 2016]. Pierwszy sposób polega na dostarczeniu z danymi wejściowymi informacji na temat ich prawidłowej klasyfikacji. To sposób budowania modelu najbardziej zbliżony do tradycyjnego postrzegania edukacji – bazuje bowiem na analizie „dobrych” i „złych” przykładów, przy czym za opisywanie danych odpowiada nauczyciel (tu: dostawca danych) a nie uczeń. Dzięki temu system przetwarzający np. obrazy radiologiczne jednocześnie z danymi testowymi otrzymuje informacje, czy na danym zdjęciu faktycznie uwidoczniła została zmiana chorobowa. Jest to oczywiście

bardzo efektywny sposób budowania modelu analitycznego, wymaga jednak starannego przygotowania danych wejściowych – co jest czasochłonne i nie we wszystkich przypadkach możliwe. Stąd też w przypadku przetwarzania bardzo dużych zbiorów danych (np. dotyczących aktywności online) stosuje się również techniki uczenia nienadzorowanego, w których system zasilany jest olbrzymimi zbiorami danych, które jednak nie zawierają wiążącej informacji na temat prawidłowej interpretacji. System, przetwarzając tego typu dane, ma samodzielnie odnaleźć podobieństwa i zależności, które pozwolą odpowiednio klasyfikować kolejne przypadki. Ten typ uczenia się jest jednak podatny na błędy wynikające z zanieczyszczenia danych – takie jak obecność w nich niepożądaných zależności, ukrytych uprzedzeń [Calders, Žliobaitė, 2013; Schnack, 2020]. Odrębną formą procesu trenowania jest tzw. uczenie przez wzmocnienie (ang. *reinforced learning*), w którym dane wejściowe nie są dostarczane do systemu, a algorytm sam je pozyskuje według określonego wzorca, natomiast w procesie przetwarzania wykorzystywane są mechanizmy pozwalające na wzmocnienie (utrwalanie) kryteriów prowadzących do prawidłowego rozwiązania problemu.

Proces trenowania – podobnie jak tradycyjnego uczenia się w przypadku ludzi – jest skończony w czasie. Inaczej jednak niż u ludzi, po uzyskaniu pewnej podstawowej niezbędnej wiedzy – proces uczenia może zostać programowo zablokowany. To istotna różnica, która w fundamentalny sposób rzutuje na przyszłe działanie systemu. Lekarz, który ukończył studia i zdał niezbędne egzaminy, nie przestaje uczyć się, wykonując swój zawód. Co więcej, doświadczenie, jakiego nabywa w kolejnych latach pracy, w istotny sposób wpływa na jego kompetencje. W przypadku systemów algorytmicznych proces uczenia się może być natomiast trwale zatrzymany – wskutek czego dane przetwarzane w trakcie eksploatacji systemu nie powodują zmiany wykorzystywanego modelu obliczeniowego. Takie rozwiązanie stosowane jest w przypadkach, gdy istnieje ryzyko, że niezwyfikowane dane wejściowe zostaną wykorzystane do zanieczyszczenia modelu analitycznego. W istocie bowiem dostarczanie odpowiednio spreparowanych danych wejściowych to jeden z możliwych wektorów ataku, wielokrotnie stosowany w przeszłości aby skutecznie zaburzyć prawidłowe działanie sztucznych systemów uczących się [Comiter, 2019]. Zablokowanie algorytmów uczenia się, z jednej strony, eliminuje to ryzyko, z drugiej powoduje, że system w dalszej eksploatacji *de facto* nie będzie zasadniczo różnił się od normalnego (klasycznego) algorytmu komputerowego, który w swoich decyzjach ograniczony jest posiadaną bazą instrukcji. Brak możliwości zmiany modelu analitycznego powoduje, że system AI nie może korygować swoich decyzji w sposób uwzględniający zmieniające się otoczenie.

Powyższe ma szczególne znaczenie, gdy system wykorzystywany jest do analizy zjawisk zmiennych w czasie. Przykładem jest analiza języka naturalnego, w tym także zmienne znaczenia jakie mogą być nadawane poszczególnym słowom czy wypowiedziom. Sformułowanie *LGBT to nie ludzie a ideologia* ma inne znaczenie i inny wydźwięk 5 lat temu i dzisiaj. Doświadczenia historyczne uczą, że używanie wykluczającego i stygmatyzującego języka – przy wykorzystaniu pojęć i terminów, którym nadaje się nowe znaczenie – to zjawisko obserwowane w przypadku przemian niedemokratycznych nie tylko dzisiaj, ale z równą dynamiką i nasileniem także w przeszłości [Klemperer i in., 2014; Ziemer, Kaminska, 2021].

Jednocześnie jednak brak ustanowienia ograniczeń w procesie uczenia się może spowodować, że nawet przy braku celowego i złośliwego działania zewnętrznych intruzów, system algorytmiczny dostosuje się do negatywnych trendów obecnych w danych wejściowych – uznając przypadki wypowiedzi skrajnych za dopuszczalną i akceptowalną normę. W ten sposób jeden z systemów opracowanych przez firmę Microsoft (tzw. chat-bot) po kilku dniach od jego uruchomienia zaczął – powtarzając nienawistne treści otrzymywane od użytkowników – publikować hasła rasistowskie [Hunt, 2016].

7.4. Obszary problemowe

7.4.1. Proces trenowania modelu i klasyfikowania wypowiedzi

Większość dotychczasowych badań dotyczących identyfikacji mowy nienawiści z wykorzystaniem systemów AI wykorzystuje modele uczenia nadzorowanego oparte na bazie słownikowej. W praktyce wykazano jednak, że również uczenie nienadzorowane może prowadzić do zbudowania równie skutecznych modeli analitycznych [Isaac i in., 2022; Pitsilis i in., 2018]. Co więcej, niektórzy wskazują, że kluczowe dla uzyskiwanych wyników jest nie tyle przyjęty model uczenia się, ale wielkość dostępnej bazy treningowej [Ruwandika, Weerasinghe, 2018].

Przygotowanie odpowiednich danych treningowych dla algorytmu wymaga natomiast realizacji tradycyjnego, manualnego procesu moderacji treści. Ocena operatora systemu – zapisana razem z komentarzem (decyzją) – stanowi później materiał treningowy dla systemu algorytmicznego. Co oczywiste, z uwagi na powszechność mowy nienawiści w Internecie, baza referencyjna musi być również bardzo rozbudowana, co z kolei wymaga, aby w proces ten były zaangażowane bardzo duże zespoły ludzkie.

Technika wykorzystująca trenowanie modelu na podstawie wzorców pozwalających na osiągnięcie satysfakcjonujących efektów, np. przy identyfikacji naruszeń prawa autorskiego. Treści audiowizualne oznaczone jako bezprawne są z dużym (w zależności od kontekstu badania – osiągającym nawet 90%) prawdopodobieństwem prawidłowo identyfikowane w przyszłości – uniemożliwiając ponowne opublikowanie tych samych (lub podobnych) utworów [Gray, Suzor, 2020].

Ten sam mechanizm pozwala również na wykrywanie przypadków mowy nienawiści na podstawie przeszłych (wcześniej ocenianych) zachowań użytkowników, użytych przez nich sformułowań czy całych wypowiedzi. W rzeczywistości jednak to, co powinno budować przewagę systemów AI nad tradycyjnymi algorytmami, nie wykorzystującymi mechanizmów uczenia maszynowego, to zdolność do odszukiwania nowych przypadków wypowiedzi krzywdzących, które nie były uwzględnione w pierwotnym zbiorze testowym. Tylko bowiem w takim przypadku, uwzględniając gigantyczną liczbę treści generowanych każdego dnia przez użytkowników, wykorzystanie systemów AI mogłoby stanowić realny sposób na ograniczenie skali mowy nienawiści w Internecie.

O ile jednak systemy algorytmiczne, analizując dostarczone im dane, mogą trafnie identyfikować identyczne lub zbliżone treści w kolejnych zestawach danych, to są stosunkowo proste do oszukania przez świadomych ich istnienia użytkowników. Czasami wystarczające jest wprowadzenie celowych błędów językowych lub zapisanie obraźliwych słów innym językiem. W ten sposób badacze wykazali możliwość oszukania opracowanego przez Google i Jigsaw systemu Perspective, stosując w tym celu umiejętne modyfikacje kolejności liter czy wstawienie dodatkowych znaków interpunkcyjnych do obraźliwych treści [Bloch-Wehba, 2020]. Przykład ten obrazuje istotną cechę procesu uczenia się w zakresie wykrywania mowy nienawiści – jaką jest dwukierunkowość tego procesu. Systemy algorytmiczne są nie tylko pasywnym obserwatorem zachowania użytkowników i treści przez nich publikowanych. Decyzje związane z moderacją wpływają na zachowania użytkowników, a więc także użytkownicy „uczą się” działania systemów AI. To fundamentalna różnica, szczególnie widoczna w porównaniu z innymi obszarami zastosowania algorytmów uczenia maszynowego. Zmiany nowotworowe nie są świadome badania ich przez systemy algorytmiczne, nie dostosowują się więc do zmieniającego się środowiska diagnostycznego – tak, aby algorytmy nie wykryły ich istnienia. Użytkownicy zmieniają natomiast swoje zachowania, tak aby wypowiedzi, które zamieszczają nie zostały wykryte przez algorytmy stosowane przez dostawców usług cyfrowych – dotyczy to zresztą nie tylko wypowiedzi nienawistnych, ale również publikacji wszelkich innych treści naruszających prawo lub regulamin korzystania z usługi. Przykładem może być

modyfikowanie utworów audiowizualnych w celu omięcia mechanizmów wykrywania naruszenia prawa autorskich [Brøvig-Hanssen, Jones, 2021].

Prawdziwym wyzwaniem dla twórców algorytmów jest także określenie granic dopuszczalnej wypowiedzi. Ponieważ nie istnieją powszechnie przyjmowane definicje prawne różnych typów treści nienawistnych, większość produktów bazuje na własnych klasyfikacjach. Badane wypowiedzi są oznaczane metrykami (np. procentowymi lub wyrażonymi w innej skali), które wyrażają „pewność” modelu, że dana wypowiedź zawiera mowę nienawiści. Proces ten jest jednak bardzo złożony, a jego dokładność – ograniczona. W efekcie komentarz *śmierć wszystkim cyklistom* – jako nie dotyczący grupy szczególnie chronionej (np. dzieci czy mniejszości etnicznych, religijnych, seksualnych itp.) może być oceniony inaczej, niż gdyby odnosił się *explicit* do którejkolwiek z tych grup. W efekcie, jak wykazał Sinderson, system Perspective klasyfikował pojedyncze słowo *Arabowie* jako nieakceptowalne w 63%, podczas gdy wynik analizy frazy *Kocham führera* wynosił zaledwie 3% [Gorwa i in., 2020, s. 10].

Co ciekawe, o ile algorytmy zazwyczaj dobrze sobie radzą z identyfikacją treści nawołujących do przemocy czy treści ksenofobicznych (dzięki wykorzystaniu bazy słownikowej), to są dużo mniej skuteczne przy prawidłowym klasyfikowaniu wypowiedzi nienawistnych skierowanych do kobiet czy osób niepełnosprawnych – w przypadku modelu DistilBERT – odpowiednio 30,9% oraz 25,4% [Wiggers, 2021]. Przyczyna tej anomalii jest *prawdopodobnie* związana z zastosowaniem odmiennych wag przypisanych w modelu analitycznym różnym typom wypowiedzi, wyrażających odmienny priorytet w ochronie poszczególnych grup użytkowników. Hipoteza ta nie może zostać zweryfikowana głównie z powodu braku ujawnienia założeń projektowych oraz algorytmów wykorzystywanych w poszczególnych modelach przez ich autorów (firmy technologiczne).

Oparcie działania modelu na własnych metrykach, ale również kategoriach wypowiedzi niedopuszczalnych, *de facto* powoduje, że moderacja treści odbywa się w sposób mało transparentny [Burrell, 2016]. Sytuację dodatkowo komplikuje fakt, że w przypadku systemów uczenia maszynowego uzasadnienie nadania określonej metryki danej wypowiedzi jest zadaniem złożonym, a w niektórych przypadkach – wręcz niemożliwym (o czym później).

7.4.2. Znaczenie kontekstu wypowiedzi

Odrębnym zagadnieniem, chociaż powiązaniem z procesami uczenia systemów AI, jest analiza kontekstowa badanych wypowiedzi. W przypadku nadzorowanego trenowania w modelu analitycznym wykorzystuje się bazy

referencyjne, zawierające wypowiedzi zaklasyfikowane przez moderatorów dostawcy usługi lub samych użytkowników jako nieakceptowalne. Decyzja użytkownika o zgłoszeniu konkretnego wpisu do moderacji często jest jednak efektem lektury szerszej wypowiedzi lub rozmowy, której dany wpis jest elementem. Także relacja pomiędzy komunikującymi się użytkownikami, ich przynależność do tej samej (lub różnej) grupy społecznej (etnicznej, zawodowej itp.) może wpływać na wynikową decyzję użytkownika o zgłoszeniu treści jako nieakceptowalnej. Co więcej, poszczególne grupy społeczne często stosują swój, zrozumiały głównie dla nich, język (nadając powszechnie stosowanym pojęciom inne znaczenie). Oderwanie badanej wypowiedzi od kontekstu jej sformułowania uniemożliwia zrozumienie jej prawdziwego (zamierzonego przez autora) znaczenia. Dlatego identyfikacja mowy nienawiści nie powinna być sprowadzana wyłącznie do analizy konkretnej wypowiedzi: jej całościowe znaczenie może ujawnić się dopiero przy analizie szerszego kontekstu, w którym wypowiedź ta została formułowana. System trenowany na danych, pozbawionych tego kontekstu – nie będzie potrafił prawidłowo ocenić tego typu wypowiedzi [Wilson, Land, 2021].

Co więcej, wypowiedzi nienawistne często nie są tak kwalifikowane z uwagi na język, jaki został użyty w danej wypowiedzi, ale na efekt (skutek), jaki zamierza osiągnąć jej autor – np. obrażenie, poniżenie, w skrajnych sytuacjach dehumanizacja rozmówcy. Warto pamiętać, że propaganda nazistowska obfitowała w przykłady wypowiedzi skrajnych, które były formułowane w języku tworzącym pozory rozsądnej argumentacji przy wykorzystaniu pojęć odwołujących się do szeroko pojętego zaufania czy poczucia bezpieczeństwa. Jeżeli zatem systemy AI mają być równie skuteczne (jeżeli nie lepsze), co ludzie w identyfikacji mowy nienawiści – nie mogą ignorować kontekstu, ponieważ w ten sposób nie będą potrafiły dostrzec znaczenia wykraczającego poza leksykalną definicję analizowanych słów.

Kwestia kontekstu wypowiedzi powinna być także badana z uwzględnieniem norm obowiązujących w poszczególnych społeczeństwach. O ile wypowiedzi agresywne, nawołujące do przemocy czy nienawiści, są co najmniej moralnie naganne, niezależnie od języka czy kręgu kulturowego, w jakim zostały wyrażone, to analizując mniej skrajne formy naruszeń, dostrzec można szereg różnic pomiędzy społeczną oceną tego typu wypowiedzi. Jeżeli system AI ma być świadomy różnic kulturowych, wykraczających przecież poza proste różnice językowe, wiedzy tej nie pozyska z analizy bazy referencyjnej zawierającej wyłącznie wypowiedzi ocenione jako naruszające obowiązujące standardy zdefiniowane przez konkretnego dostawcę usługi cyfrowej.

7.4.3. Ryzyko nadmiernej moderacji (cenzury)

Kolejne zagadnienie, wpływające na skuteczność systemów automatycznego moderowania treści nie jest *stricte* związane z uczeniem maszynowym – ale dotyczy ogólnie odpowiedzialności dostawców usług cyfrowych za podejmowane przez nich decyzje.

W ostatnich latach coraz szerzej wprowadzane są regulacje, które przenoszą na tą grupę podmiotów odpowiedzialność za naruszenia dokonywane przez użytkowników korzystających z oferowanych im usług. Odpowiedzialność co do zasady wyraża się obowiązkiem zapewnienia zgodności, pod groźbą dotkliwych sankcji finansowych. W przypadku niektórych przepisów (np. omawianej wcześniej francuskiej ustawy o przeciwdziałaniu mowie nienawисти) prawodawca przewidział także możliwość zastosowania sankcji karnych.

W efekcie to dostawcy usług mają wdrażać środki przeciwdziałające naruszeniom prawa autorskiego, publikacjom treści promującym terroryzm czy rozpowszechnianiu materiałów związanych z seksualnym wykorzystywaniem dzieci. Ten swoisty sposób „prywatyzacji” zadań z zakresu bezpieczeństwa publicznego coraz częściej stosowany jest również do reagowania na naruszenia w zakresie dóbr osobistych [Coche, 2018]. Zarówno Trybunał Sprawiedliwości Unii Europejskiej, jak i Europejski Trybunał Praw Człowieka, wskazywały na potrzebę stosowania przez dostawców usług prewencyjnych mechanizmów zapobiegania publikacji treści w oczywisty sposób bezprawnych – odnosząc ten obowiązek także do treści naruszających dobra osobiste [Cox, 2014; Spindler, 2017]. Chociaż orzecznictwo sądów w tym zakresie jest mocno zniuansowane, nie ulega wątpliwości, że w ostatnich latach znacząco wzrosła presja na dostawców usług w zakresie proaktywnego monitorowania treści publikowanych przez użytkowników.

Nowym wymaganiem nakładanym na usługodawców towarzyszą również regulacje przewidujące możliwość nakładania na nich wysokich kar finansowych. W ślad za modelem stosowanym w przepisach o ochronie danych osobowych unijny prawodawca w nowo przyjętych regulacjach dotyczących przeciwdziałania publikacji treści terrorystycznych wprowadził możliwość stosowania kar sięgających 4% rocznego obrotu podmiotu zobowiązanego¹⁵. W przypadku globalnych dostawców usług cyfrowych oznacza to możliwość nakładania kar w wysokości setek milionów dolarów.

¹⁵ Kary te mogą być nakładane za „systematyczne lub uporczywe” niedopełnienie obowiązków wynikających z otrzymywanych nakazów usunięcia danych. Zob. Art. 18(3) rozporządzenia 2021/784.

Z drugiej strony, w przypadku gdy w wyniku działań podejmowanych przez usługodawcę zablokowane zostaną materiały nie stanowiące publikacji bezprawnej, regulacje prawne przewidują zazwyczaj możliwość dochodzenia praw przez autorów treści na drodze sądowej według przepisów powszechnie obowiązujących. W takim wypadku nie znajdują zastosowania zarówno procedury związane z nakładaniem kar finansowych, ani przyspieszone procedury przymuszania dostawcy do określonych działań (np. odblokowania publikacji).

Powyższe tworzy bardzo czytelny model regulacyjny, w którym dostawcy usług są zachęceni do podejmowania działań zmniejszających ryzyko prowadzenia ich działalności – a więc takich, które nie będą prowadziły do publikowania treści niezgodnych z prawem, nawet kosztem wprowadzenia nadmiernej moderacji. Ryzyko ustanowienia zasad nadmiernej moderacji (często i błędnie nazywane ‘cenzurą’) jest dobrze rozpoznane i od lat stanowi przedmiot ożywionej dyskusji¹⁶.

W przypadku systemów AI ryzyko to jest tym większe, że jak wcześniej wskazano, w przypadku mowy nienawiści ocena poszczególnych wypowiedzi realizowana przez algorytmy zazwyczaj wyrażana jest jako wartość z określonej skali¹⁷. Abstrahując od dojrzałości wykorzystywanego modelu analitycznego, to od dostawcy usługi cyfrowej zależy ustalenie progu, powyżej którego moderowane treści będą oznaczane jako zawierające mowę nienawiści. W przykładowej skali 1–100 może być to 75, ale w przypadkach wiążących się ze szczególnymi ryzykami (np. propagowaniem treści ksenofobicznych, homofobicznych itp.) próg ten może być dla niektórych grup zmniejszony, przykładowo do poziomu 50. W efekcie, korzystając z niedoskonałego modelu obliczeniowego, dostawca usługi, dążąc do ograniczenia swojej odpowiedzialności, może nie tylko wprowadzać mechanizmy nadmiernej moderacji, ale wręcz ustanawiać środki dyskryminacji poszczególnych kategorii wypowiedzi.

W 2020 roku niemiecki Federalny Trybunał Sprawiedliwości (FTS) badał sprawę, zapoczątkowaną decyzją operatora portalu społecznościowego (Facebook) o usunięciu niektórych komentarzy użytkowników oraz zablokowaniu ich kont dostępnych z uwagi na naruszenie regulaminu świadczenia usługi¹⁸. Opublikowane wpisy były reakcją na film opublikowany w mediach społecznościowych, a ich treść dotyczyła polityki migracyjnej prowadzonej przez

¹⁶ Zob. np. zdanie odrębne sędziów Sajó i Tsotsoria przedstawione w sprawie *Delfi v Estonia* (ETPCz, 16.06.2015 r., 64569/09).

¹⁷ W systemie Perspective jest to tzw. toksyczność (ang. *toxicity*) badanego materiału.

¹⁸ BGH 29.07.2021 r., III ZR 179/20 i III ZR 192/20.

ząd niemiecki. Chociaż nie zawierały one treści nawołujących do przemocy czy jawnie obraźliwych, zostały uznane za przykład mowy nienawiści¹⁹. FTS uznał, że Facebook w tym wypadku zareagował nieprawidłowo. Jednak nie dlatego, że zablokował kwestionowane treści, ale z uwagi na brak poinformowania użytkowników o podejmowanych działaniach w sposób dający im możliwość wyrażenia swojego stanowiska.

Podobnych przypadków, w których dostawcy usług decydowali o nadmiarowym usuwaniu treści z powodu ich rzekomej niezgodności z przyjętymi standardami komunikacji udokumentowano znacznie więcej [Llansó, 2020]. We wszystkich tych przypadkach wykorzystywano systemy algorytmiczne, które – pozbawione kontekstu analizy – uznawały badane treści za nieakceptowalne, co w połączeniu z dążeniem dostawców do zmniejszenia ryzyka, że stosowana przez nich moderacja okaże się nieskuteczna, doprowadziło do usunięcia zarówno materiałów naruszających prawo jak i takich, w których naruszenia tego typu piętnowano. Przykładem mogą być doniesienia o blokowaniu treści dokumentujących działania podejmowane przez grupy rasistowskie czy ksenofobiczne [Bours, 2017].

7.4.4. (Nie)wyjaśnialność podejmowanych decyzji

Jednym z węzłowych problemów, którego rozwiązanie warunkuje szerokie wykorzystanie systemów uczenia maszynowego w gospodarce, jest zdolność do przedstawienia argumentacji, która doprowadziła do wydania określonego rozstrzygnięcia. Problem ten nazywa się „wyjaśnialnością” (ang. *explainability*) decyzji podejmowanych przez AI – i jest bezpośrednio związany ze zdolnością do odtworzenia sposobu wnioskowania, a w ten sposób przedstawienia zarówno okoliczności, które zostały wzięte pod uwagę, jak i tych, które pozostawały bez wpływu z finalną decyzją [Kaminski, 2019]. Kwestia ta ma szczególne znaczenie w przypadku decyzji wywierających istotny wpływ na prawa jednostki, ponieważ w takim przypadku *de facto* warunkuje możliwość poznania przyczyn zapadłego rozstrzygnięcia i skorzystania z przysługujących ścieżek jego zaskarżenia.

Co do zasady wyjaśnialność decyzji, to cecha ta pozostaje bez związku z prawidłowością działania algorytmu. Zdolność do prześledzenia sposobu osiągnięcia przez system określonego wniosku jest zatem poszukiwaniem odpowiedzi na pytanie, dlaczego określona decyzja zapadła a nie, czy jest ona

¹⁹ Tłum. jednego z wpisów opublikowanych w portalu: *Czego ci ludzie szukają w naszym konstytucyjnym państwie – żadnego szacunku – żadnego szacunku dla naszych praw – żadnego szacunku dla kobiet. Nigdy się tu nie zintegrują i na zawsze będą obciążeniem dla podatnika.*

prawidłowa. W rzeczywistości ustalenie kryteriów, które przesądziły o podjęciu określonego rozstrzygnięcia, ma pomóc w zrozumieniu, w jaki sposób okoliczności faktyczne musiałyby się zmienić, aby możliwe było podjęcie innej decyzji. Wyjaśnialność pomaga zatem poznać nie tylko powody podjęcia określonej decyzji, ale też odpowiedzieć na pytanie, dlaczego nie podjęto decyzji odmiennej – oczekiwanej przez użytkownika.

Jednocześnie wyjaśnienie sposobu działania systemów algorytmicznych w wielu przypadkach nie tylko nie jest zadaniem prostym, ale często wręcz niemożliwym. Działanie algorytmów opartych na sztucznych sieciach neuronowych, które są z sukcesem wykorzystywane także w zastosowaniach związanych z przetwarzaniem języka naturalnego, nie pozwala na zrozumiałą dla ludzi interpretację otrzymywanych wyników. Stąd też algorytmy tego typu często określa się terminem *black box*, który ma podkreślać że ich sposób wnioskowania – z uwagi na wykorzystywany sposób reprezentacji wiedzy – nie może być odtworzony czy prześledzony w sposób pozwalający odtworzyć (powtórzyć) sposób dojścia do uzyskanego wyniku [Gryz, Rojszczak, 2021]. Z tego też powodu część badaczy postuluje, aby tego typu systemów w ogóle nie wykorzystywać do zastosowań, które mają istotny wpływ na prawa jednostki [Rudin, 2019]. Z wnioskiem tym zgodził się również Trybunał Sprawiedliwości UE, wskazując na niedopuszczalności opierania decyzji organów publicznych na decyzjach podejmowanych przez systemy sztucznej inteligencji, jeżeli nie ma technicznej możliwości wyjaśnienia określenia powodów podjęcia tych decyzji²⁰.

Jednak wyjaśnialność decyzji jest zagadnieniem problematycznym również w przypadku drzew decyzyjnych – a więc drugiej, najpopularniejszej technologii wykorzystywanej do budowy systemów uczenia maszynowego. O ile w przypadku wnioskowania opartego na drzewach decyzyjnych sposób ustalenia wyniku końcowego może być prześledzony, to często i w tym wypadku nie ma on sensownego dla ludzi znaczenia. Systemy tego typu nie tylko wykorzystują bardzo rozbudowane struktury danych (lasy drzew decyzyjnych), to jeszcze kryteria używane w tych drzewach są zazwyczaj budowane w sposób algorytmiczny (na etapie budowy modelu) i wyrażają relacje zidentyfikowane w danych źródłowych, nie mając przy tym prostego znaczenia, łatwo wyjaśnialnego dla ludzi.

Powyższe w praktyce prowadzi do wielu wątpliwości w zakresie możliwości wykorzystania systemów AI do automatycznego moderowania treści publikowanych w Internecie. Prostym rozwiązaniem służącym ominięciu tego problemu, jest wprowadzenie przez dostawcę usług procedur, za pomocą których użytkownik, którego publikacje zostały zablokowane, może zażą-

²⁰ TSUE, C-817/19, *Ligue des droits humains v Conseil des ministres*, pkt 195.

dać dodatkowej ręcznej weryfikacji realizowanej przez człowieka. W takim przypadku – gdy decyzja o zablokowaniu treści zostanie utrzymana w mocy, możliwe jest przedstawienie użytkownikowi powodu podjęcia takiego rozstrzygnięcia.

Jednak i wtedy pojawia się problem dotyczący skali działania – systemy algorytmiczne co do zasady pozwalają na przetwarzanie olbrzymiej liczby danych, daleko wykraczającej poza możliwości ręcznej analizy prowadzonej przez operatorów systemu. Gdyby przyjąć, że każda decyzja o zablokowaniu treści ma podlegać ręcznej weryfikacji, ekonomiczny sens wprowadzenia systemów algorytmicznych byłby trudny do wykazania. Co więcej, przeprowadzenie w rozsądnym czasie manualnej weryfikacji obejmującej wyłącznie zablokowane wypowiedzi jest w praktyce niemożliwe. Według danych Facebooka, w pierwszym kwartale 2022, portal otrzymał 587 tysięcy skarg od użytkowników na decyzje o zablokowaniu treści, które w niemal 49 tysięcy przypadków doprowadziły do przywrócenia publikacji²¹. Choć w niektórych przypadkach dostawcy usług umożliwiają złożenie skargi skutkującej przeprowadzeniem ręcznej weryfikacji podjętej wcześniej decyzji²², to w wielu przypadkach taka możliwość nie jest oferowana – a decyzja podjęta automatycznie w zakresie usunięcia (zablokowania) określonych publikacji staje się *de facto* ostateczna.

Taka sytuacja prowadzi do licznych kontrowersji związanych z dopuszczalnością stosowania takich praktyk – zwłaszcza w zakresie, w jakim prowadzi do ryzyka podejmowania arbitralnych decyzji przez dostawców usług skutkujących naruszeniem wolności słowa [Llansó i in., 2020]. Stąd też prawodawca unijny w przepisach projektowanego rozporządzenia o usługach cyfrowych uwzględnił wymaganie informowania odbiorcy decyzji, najpóźniej w momencie usunięcia lub uniemożliwienia dostępu do treści, o podjętej decyzji wraz „jasnym i szczegółowym” uzasadnieniem²³. Ponadto, zgodnie z projektowanymi przepisami operatorzy platform internetowych będą musieli także zapewnić, aby skargi dotyczące tego typu rozstrzygnięć nie były rozpatrywane „wyłącznie na podstawie zautomatyzowanych środków”²⁴.

²¹ W tym samym okresie zablokowanych zostało ponad 15 mln publikacji. Zob. dane publikowane w rejestrze transparentności Facebook: <https://transparency.fb.com/data/community-standards-enforcement/hate-speech/facebook/#appealed-content> (dostęp 4.07.2022).

²² Zob. np. mechanizmy stosowane w moderacji treści przez YouTube: <https://support.google.com/youtube/answer/185111> (dostęp 4.07.2022).

²³ Zob. Art. 15(1) projektu rozporządzenia w sprawie jednolitego rynku usług cyfrowych (akt o usługach cyfrowych), COM(2020) 825 final.

²⁴ Zob. Art. 17(5) projektu rozporządzenia w sprawie jednolitego rynku usług cyfrowych (akt o usługach cyfrowych), COM(2020) 825 final.

Rozporządzenie o usługach cyfrowych znajduje się nadal na etapie prac legislacyjnych, stąd też obecnie nie sposób przewidzieć jego wpływu na wykorzystywanie systemów sztucznej inteligencji do moderowania treści publikowanych przez użytkowników. Należy jednak oczekiwać, że skutkiem wejścia w życie przepisów będzie wprowadzenie dodatkowych zabezpieczeń, chroniących użytkowników przed podejmowaniem arbitralnych decyzji. Niewykluczone, że zmiany te wpłyną również na sposób projektowania i eksploataowania samych systemów służących do moderacji treści, tak aby ich działanie było bardziej transparentne i podatne na ocenę zewnętrzną.

7.5. Podsumowanie

Skuteczna walka z mową nienawiści to zadanie wymagające szerszego – niż tylko techniczny – spojrzenia na problem powstawania i dystrybucji bezprawnych treści w Internecie. Chociaż według dostępnych badań, znaczna część użytkowników jest świadoma problemu i spotkała się z przypadkami mowy nienawiści, to jednocześnie nadal istnieje trudność w precyzyjnym zdefiniowaniu, jak to pojęcie należy rozumieć [Schmid i in., 2022]. W efekcie wielu użytkowników może nieświadomie przekraczać granicę pomiędzy tym, co w ich ocenie można uznać za dopuszczalną krytykę, a formułowaniem wypowiedzi zawierających krzywdzące opinie czy nieprawdziwe fakty. Stąd też, zastanawiając się nad rolą dostawców usług cyfrowych – w szczególności portali społecznościowych – w przeciwdziałaniu rozpowszechnianiu mowy nienawiści, w pierwszej kolejności należy lepiej rozpoznać przyczyny, dla których w ostatnich latach zjawisko to w tak dużym stopniu przybrało na sile.

Wraz z rozwojem możliwości technicznych oferowanych przez nowoczesne systemy algorytmiczne ich wykorzystanie jest coraz częściej postrzegane nie tylko jako katalizator rozwoju gospodarczego, ale również jako środek wspierający budowę nowoczesnego społeczeństwa. Definiowanie AI jako technologii pozwalającej rozwiązać niemal dowolny problem w oczywisty sposób wzmacnia również przekonanie, że systemy tego typu – wykorzystując możliwość uczenia się na podstawie gigantycznych zbiorów publicznie dostępnych wypowiedzi użytkowników – będą mogły sprawnie i skutecznie rozwiązać problem mowy nienawiści, identyfikując nieakceptowalne wypowiedzi jeszcze przed ich opublikowaniem.

W praktyce przekonanie o adekwatności AI do realizacji tego celu wynika po części z niezrozumienia technicznych aspektów działania systemów uczenia maszynowego, a po części z braku wiedzy na temat negatywnych

konsekwencji algorytmicznego moderowania treści. Choć dzisiaj systemy uczenia maszynowego są już powszechnie wykorzystywane do identyfikowania mowy nienawiści, to jednocześnie nie widać, aby w ostatnich latach skala tego zjawiska została istotnie ograniczona. Coraz częściej natomiast formułowane są zastrzeżenia, czy swoboda w podejmowaniu decyzji cenzorskich przez dostawców usług cyfrowych – i to na podstawie działających nietransparentnie algorytmów – nie prowadzi do nieproporcjonalnej ingerencji w prawa użytkowników, w tym w prawo do wolności słowa.

Z pewnością udzielenie ostatecznej odpowiedzi na pytanie, czy systemy algorytmiczne staną się złotym środkiem w zakresie walki z mową nienawiści nie jest jeszcze dzisiaj możliwa. Postęp w dziedzinie AI może doprowadzić do pokonania niektórych – lub wszystkich – z barier obserwowanych obecnie. Jednocześnie jednak nie jest wykluczone, że wraz z rozwojem świadomości użytkowników bardziej skutecznym narzędziem w osiągnięciu tego celu będzie nie tyle moderowanie dyskusji, ale promowanie edukacji lub innych rozwiązań technicznych, np. zapewniających rozliczalność i znoszących poczucie bezkarnej anonimowości w odniesieniu do działań podejmowanych w Internecie.

BIBLIOGRAFIA

- Alpaydin E., (2016), *Machine learning: The new AI*, MIT Press.
- Becker P.J., Byers B., Jipson A., (2000), *The Contentious American Debate: The First Amendment and Internet-based Hate Speech*. International Review of Law, Computers & Technology, 14(1), s. 33–41, <https://doi.org/10.1080/13600860054872>.
- Bloch-Wehba H., (2020), *Automation in Moderation*, Cornell International Law Journal, 53, s. 42–96.
- Bours B., (2017), *Facebook's Hate Speech Policies Censor Marginalized Users*, Wired, <https://www.wired.com/story/facebooks-hate-speech-policies-censor-marginalized-users/>.
- Brøvig-Hanssen R., Jones E., (2021), *Remix's retreat? Content moderation, copyright law and mashup music*, New Media & Society, <https://doi.org/10.1177/14614448211026059>.
- Burrell J., (2016), *How the machine 'thinks': Understanding opacity in machine learning algorithms*, Big Data & Society, 3(1), <https://doi.org/10.1177/2053951715622512>.
- Calders T., Žliobaitė I., (2013), *Why Unbiased Computational Processes Can Lead to Discriminative Decision Procedures*, [w:] W B. Custers, T. Calders, B. Schermer, & T. Zarsky (red.), *Discrimination and Privacy in the Information Society* (t. 3, s. 43–57), Springer Berlin Heidelberg, https://doi.org/10.1007/978-3-642-30487-3_3.
- Christiano Silva T., Zhao L., (2016), *Machine Learning*, [w:] *Machine Learning in Complex Networks* (1st ed. 2016), Springer International Publishing, Imprint: Springer, <https://doi.org/10.1007/978-3-319-17290-3>.
- Cobbe J., (2020), *Algorithmic Censorship by Social Platforms: Power and Resistance*, Philosophy & Technology, <https://doi.org/10.1007/s13347-020-00429-0>.

- Coche E., (2018), *Privatised enforcement and the right to freedom of expression in a world confronted with terrorism propaganda online*, Internet Policy Review, 7(4), <https://doi.org/10.14763/2018.4.1382>.
- Comiter M., (2019), *Attacking Artificial Intelligence*, Harvard Kennedy School.
- Cox N., (2014), *Delfi AS v Estonia: The Liability of Secondary Internet Publishers for Violation of Reputational Rights under the European Convention on Human Rights*, Modern Law Review, 77(4), s. 619–629.
- Delgado R., Stefancic J., (2018), *Must we defend Nazis? Why the First Amendment should not protect hate speech and white supremacy*, New York University Press.
- Downing J.D.H., (1999), *'Hate Speech' and 'First Amendment Absolutism' Discourses in the US*, Discourse & Society, 10(2), s. 175–189, <https://doi.org/10.1177/0957926599010002003>.
- Franklin S., (1995), *Artificial minds*, MIT Press.
- Gillespie T., (2020), *Content moderation, AI, and the question of scale*, Big Data & Society, 7(2), <https://doi.org/10.1177/2053951720943234>.
- Gorwa R., Binns R., Katzenbach C., (2020), *Algorithmic content moderation: Technical and political challenges in the automation of platform governance*, Big Data & Society, 7(1), <https://doi.org/10.1177/2053951719897945>.
- Gray J.E., Suzor N.P., (2020), *Playing with machines: Using machine learning to understand automated copyright enforcement at scale*, Big Data & Society, 7(1), <https://doi.org/10.1177/2053951720919963>.
- Gryz J., Rojszczak M., (2021), *Black box algorithms and the rights of individuals: No easy solution to the „explainability” problem*, Internet Policy Review, 10(2). <https://doi.org/10.14763/2021.2.1564>.
- Hunt E., (2016), *Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter*, The Guardian, https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter?CMP=twt_a-technology_b-gdntech.
- Isaac A., Kumar R., Bhat A., (2022), *Hate Speech Detection Using Machine Learning Techniques*, [w:] W.S. Aurelia, S.S. Hiremath, K. Subramanian, S.Kr. Biswas (red.), Sustainable Advanced Computing (t. 840, s. 125–135), Springer Singapore, https://doi.org/10.1007/978-981-16-9012-9_11.
- Kaminski M.E., (2019), *The Right to Explanation, Explained*, Berkeley Technology Law Journal, 34,s. 189–218.
- Klemperer V., Zychowicz J., Klemperer V., (2014), *LTI: Notatnik filologa* (wyd. 3., rozsz). Wydawn. Aletheia.
- Koebler J., Cox J., (2018), *The Impossible Job: Inside Facebook's Struggle to Moderate Two Billion People*, Vice, https://www.vice.com/en_us/article/xwk9zd/how-facebook-content-moderation-works.
- Llansó E.J., (2020), *No amount of „AI” in content moderation will solve filtering's prior-restraint problem*, Big Data & Society, 7(1), <https://doi.org/10.1177/2053951720920686>.
- Llansó E., van Hoboken J., Leerssen P., Harambam J., (2020), *Artificial Intelligence, Content Moderation, and Freedom of Expression*, Transatlantic Working Group.
- Pitsilis G.K., Ramampiaro H., Langseth H., (2018), *Effective hate-speech detection in Twitter data using recurrent neural networks*, Applied Intelligence, 48(12), s. 4730–4742, <https://doi.org/10.1007/s10489-018-1242-y>.
- Romero Moreno F., (2020), *'Upload filters' and human rights: Implementing Article 17 of the Directive on Copyright in the Digital Single Market*, International Review of Law, Computers & Technology, 34(2), s. 153–182, <https://doi.org/10.1080/13600869.2020.1733760>.

- Rudin C., (2019), *Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead*, *Nature Machine Intelligence*, 1(5), s. 206–215, <https://doi.org/10.1038/s42256-019-0048-x>.
- Ruwandika N.D.T., Weerasinghe A.R., (2018), *Identification of Hate Speech in Social Media*, 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer), s. 273–278, <https://doi.org/10.1109/ICTER.2018.8615517>.
- Schmid U.K., Kumpel A.S., Rieger D., (2022), *How social media users perceive different forms of online hate speech: A qualitative multi-method study*, *New Media & Society*, <https://doi.org/10.1177/14614448221091185>.
- Schnack H., (2020), *Bias, noise, and interpretability in machine learning*, [w:] *Machine Learning*, s. 307–328, <https://doi.org/10.1016/b978-0-12-815739-8.00017-1>.
- Spindler G., (2017), *Responsibility and Liability of Internet Intermediaries: Status Quo in the EU and Potential Reforms*, [w:] W T.-E. Synodinou, P. Jougoux, C. Markou, T. Prastitou (red.), *EU Internet Law* (s. 289–314), Springer International Publishing, https://doi.org/10.1007/978-3-319-64955-9_12.
- Waldron J.. (2012), *The harm in hate speech*, Harvard Univ. Press.
- Wiggers K., (2021), *Researchers find machine learning models still struggle to detect hate speech*, *VentureBeat*, <https://venturebeat.com/2021/01/06/researchers-find-machine-learning-models-still-struggle-to-detect-hate-speech/>.
- Wilson R.A., Land M.K., (2021), *Hate Speech on Social Media: Content Moderation in Context*, *Connecticut Law Review*, 52(3), s. 1029–1076.
- Ziemi, G., Kaminska A.D., Barbaro B., (2021), *Jak wychować nazistę: Reportaż o fanatycznej edukacji*, *Znak Litera Nova*.

Rozdział 8

Big Data, sztuczna inteligencja i zrównoważony rozwój miast... w epoce (post) COVID-19

Big Data, Artificial Intelligence, and the sustainable development of cities... in the (post) COVID-19 era

Piotr Wójcik

Wydział Nauk Ekonomicznych
Uniwersytet Warszawski
Data Science Lab WNE UW

Grzegorz Kula

Wydział Nauk Ekonomicznych
Uniwersytet Warszawski

Streszczenie. Główne wyzwania naszych czasów dotyczą miast i ich zrównoważonego rozwoju. Miasta te stają się coraz bardziej „inteligentne” dzięki gromadzeniu wielkich zbiorów danych, rozwojowi algorytmów sztucznej inteligencji, technologii szerokopasmowej transmisji danych 5G, internetowi rzeczy oraz wzajemnej komunikacji między urządzeniami. Pandemia COVID-19 znacząco przyspieszyła te procesy. Inteligentne miasto może oferować szereg korzyści w zarządzaniu i optymalizacji tradycyjnych usług publicznych. Sama technologia nie wystarczy jednak, aby miasto było lepsze – jej wykorzystanie musi odpowiadać na realne potrzeby mieszkańców, być bezpieczne, niezawodne, skalowalne, integracyjne i przejrzyste. Angażowanie wszystkich obywateli w korzyści płynące z nowoczesnych technologii to duże wyzwanie, zwłaszcza wobec wciąż wysokich kosztów i braku regulacji. Niskie koszty wdrożenia, również w sensie oddziaływania na środowisko, są kluczowym czynnikiem zrównoważonego rozwoju inteligentnych miast.

Abstract. The main challenges of our time relate to cities and their sustainable development. Due to the collection of Big Data, the development of Artificial Intelligence algorithms, 5G broadband data transmission technology, the Internet of Things, and mutual communication between machines, cities are becoming more and more intelligent. COVID-19 pandemics has speed up these processes. The smart city may offer a number of benefits in the management and optimization of traditional public services. However, technology alone is not enough to make a city better – its use must meet

the real needs of urban residents, be safe, reliable, scalable, inclusive, and transparent. Involving all citizens in the benefits of modern technologies is a big challenge, given their still high costs and lack of regulation. Low deployment costs (also in the environmental-impact sense) are a key factor for the sustainability of smart city solutions.

Słowa kluczowe: technologie informacyjno-komunikacyjne, inteligentne miasto, zrównoważony rozwój, wykluczenie społeczne, pandemia COVID-19

Keywords: ITC, smart city, sustainable development, social exclusion, COVID-19 pandemics

8.1. Wprowadzenie

Obecnie więcej ludzi mieszka w miastach niż na obszarach wiejskich. Dlatego też główne wyzwania naszych czasów dotyczą miast i ich zrównoważonego rozwoju. Zarządzanie miastami i decyzje polityczne są w dużym stopniu uzależnione od rewolucji technologicznej oraz pojawienia się nowych koncepcji i narzędzi: Big Data, Internetu Rzeczy, sztucznej inteligencji, algorytmów uczenia maszynowego oraz chmur obliczeniowych [Barro i in., 2018; Allam i Dhunny, 2019]. Idea włączenia technologii do codziennych działań ludzi mieszkających w miastach w celu zapewnienia im lepszego standardu życia doprowadziła do powstania inteligentnego miasta (ang. *smart city*). Celem rozdziału jest przedstawienie krótkiego przeglądu technologii wykorzystywanych w inteligentnych miastach, wyjaśnienie istotnych możliwości i wyzwań związanych z inteligentnymi miastami oraz sformułowanie rekomendacji dotyczących dalszych działań.

Pojęcie inteligentnego miasta jest szeroko używane, choć trudno je jednoznacznie zdefiniować. Kluczowym aspektem jest wykorzystanie technologii informacyjno-komunikacyjnych (ICT), które są stosowane przez władze i społeczeństwo jako narzędzie do stworzenia lepszych warunków życia i pracy. Niezbędny do tego jest udział obywateli, ich partnerstwo z władzami miejskimi i ogólnie współpraca wszystkich zainteresowanych stron. Inne ważne cele obejmują lepszą alokację zasobów, zrównoważony rozwój gospodarczy¹, ochronę środowiska i poprawę efektywności kosztowej polityki miejskiej. Jednak bez zaangażowania społeczeństwa cele te są bardzo trudne do osiągnięcia.

¹ Zgodnie z definicją Organizacji Narodów Zjednoczonych, inteligentne miasto (*smart city*) to miasto zrównoważone. ONZ używa nawet wspólnego terminu „inteligentne zrównoważone miasto”.

Winkowska i in. [2019, s. 71] wymieniają następujące problemy współczesnych miast: niekontrolowane powiększanie się miast, zanieczyszczenie środowiska, logistyka miejska, infrastruktura techniczna, gospodarka odpadami, starzenie się społeczeństwa, rozwarstwienie poziomu zamożności, obszary ubóstwa oraz niski poziom partycypacji obywateli w zarządzaniu sprawami publicznymi. Przekształcenie w inteligentne miasto może rozwiązać niektóre z tych problemów i złagodzić inne. Jednakże taka transformacja jest złożona i kosztowna oraz może powodować inne problemy, wśród których najważniejsze wydają się: inwestowanie w technologię jako cel sam w sobie, bez związku z rzeczywistymi problemami, koncentrowanie się na wielkich projektach, nowej infrastrukturze i zaniedbanie stanu starych obiektów, wzrost nierówności i wykluczenia ze względu na nierówne stosowanie ICT, brak umiejętności technicznych i koszty narzędzi. Istnieją także problemy związane z samymi ICT, takie jak obawy dotyczące prywatności, obciążenie algorytmiczne, etyka maszyn i „zła” sztuczna inteligencja. Większość tych problemów dotyczy jednak przyszłości i nie ma zbyt wielu rzeczywistych przypadków, do których można by się odnieść, choć proces zmian przyspieszył z powodu pandemii COVID-19. Obecnie głównym problemem jest dostępność danych [Ryan i Gregory, 2019] i ich efektywne wykorzystanie. Niemniej odniesiemy się również do kosztów projektów inteligentnych miast, a także wyzwań związanych z kwestiami etycznymi dotyczącymi miast, nowych technologii i sztucznej inteligencji. Na koniec omówimy rolę rządów i regulacji oraz zaproponujemy pewne rozwiązania dla Unii Europejskiej i jej państw członkowskich.

8.2. Wykorzystywana technologia – Big Data, IoT, ML, AI

Termin Big Data odnosi się do dużych zbiorów danych zawierających różnorodne dane, także w postaci nieustrukturyzowanej (np. dane tekstowe, obrazy, filmy). Ich skala powoduje, że przetwarzanie i analiza statystyczna przy użyciu tradycyjnych komputerów i narzędzi statystycznych jest trudna, a czasami wręcz niemożliwa. Wartością dodaną analizy tego typu danych jest możliwość odkrycia zaskakujących, nieznanych wcześniej zależności, które pozwalają na zdobycie nowej wiedzy, często przekładającej się na korzyści finansowe. Oczywistym przykładem analizy Big Data jest przetwarzanie danych gromadzonych przez dużych graczy internetowych. Lista stron odwiedzanych w Internecie, frazy wpisywane do przeglądarki internetowej, mediów społecznościowych czy darmowej poczty elektronicznej, czas spędzony na poszczególnych stronach, klikane linki, produkty kupowane online, lista zna-

jomych czy treści publikowane w serwisach społecznościowych to kopalnia wiedzy o każdym internaucie. Wyzwaniem w przypadku dużych zbiorów danych jest ich przechowywanie i efektywne przetwarzanie.

Duże ilości danych są wykorzystywane przez algorytmy uczenia maszynowego (ML)², które służą do odkrywania zależności między osobami, a następnie dzielenia ich na grupy zachowujące się w podobny sposób, lub przewidywania ich przyszłych zachowań. Przykładowo, na podstawie historii odwiedzanych przez internautów stron internetowych oraz informacji udostępnianych w sieciach społecznościowych można przewidzieć ich płeć, stan cywilny, wykształcenie, stanowisko pracy, wyznanie, orientację seksualną oraz różnego rodzaju preferencje, skłonności i potrzeby – np. skłonność do zmiany pracy lub zakończenia obecnego związku w najbliższej przyszłości [Kosiński i in., 2013]. Firmy gromadzące tego typu dane często wiedzą o internautach więcej niż ich najbliżsi. Nowoczesne algorytmy uczenia maszynowego są wykorzystywane do wspomaganie i w dużym stopniu automatyzacji decyzji. Jednymi z najbardziej udanych systemów uczenia maszynowego są (sztuczne) sieci neuronowe, które stanowią próbę naśladowania zasad działania ludzkiego mózgu. Innym nowoczesnym terminem jest głębokie uczenie, które ogólnie oznacza wykorzystanie sieci neuronowych z wieloma warstwami neuronów. Wszystkie wyżej wymienione elementy są kluczowymi składnikami sztucznej inteligencji (SI). Jest to analiza danych w celu modelowania jakiegoś wybranego aspektu świata i wykorzystania algorytmu predykcyjnego do wnioskowania (rozumienia) i przewidywania (prognozowania) możliwych przyszłych zdarzeń [Miaillhe, Hodes, 2017]. Systemy SI mają zdolność do wykonywania operacji analogicznych do uczenia się na przykładach i podejmowania decyzji.

Dane zbierane są nie tylko online, ale także za pomocą czujników umieszczonych w smartfonach, smartwatchach i innych urządzeniach rejestrujących codzienną aktywność ich użytkowników. Nigdy wcześniej dane dotyczące różnych działań człowieka nie były generowane w tak szybkim tempie. Postęp technologiczny oznacza dalszy wzrost skali Big Data. Mówimy już o inteligentnych mieszkaniach, urządzeniach domowych (np. lodówkach), a nawet inteligentnych ubraniach wyposażonych w czujniki zbierające różne informacje i komunikujące się z innymi urządzeniami bez udziału właściciela. Taka komunikacja bezpośrednio między urządzeniami nazywana jest Internetem Rzeczy (IoT)³. IoT to wizja przyszłości, w której przedmioty używane

² ML to skrót od angielskiego określenia „Machine Learning”. Ze względu na powszechne użycie tego skrótu w literaturze przedmiotu, będziemy go używać również w tym opracowaniu.

³ IoT to skrót od angielskiego określenia „Internet of Things”. Ze względu na powszechne użycie tego skrótu w literaturze przedmiotu, będziemy go używać również w tym opracowaniu.

w życiu codziennym, wyposażone w czujniki, możliwość gromadzenia danych i ich przetwarzania, będą mogły komunikować się między sobą oraz z użytkownikami, stając się tym samym integralną częścią Internetu [Zanella i in., 2014].

Rosnąca część Big Data ma *de facto* charakter przestrzenny – zawiera informacje geolokalizacyjne. Stąd pojawienie się koncepcji przestrzennych big data [Eldawy, Mokbel, 2016]. Taki charakter mają dane dotyczące trasy powrotnej z pracy lub spaceru czy lokalizacji najczęściej odwiedzanych sklepów (płatności kartą) lub aktualnej lokalizacji samochodu. To właśnie tego typu dane pozwalają na identyfikację osób, które potencjalnie miały kontakt z ofiarami pandemii COVID-19. W koncepcji Big Data mieszczą się również informacje zbierane przez satelity. W miastach dane zbierane przez satelity mogą być wykorzystywane, na przykład, do analizowania natężenia ruchu pojazdów, tworzenia numerycznego modelu pokrycia terenu i trójwymiarowych modeli miast, które są następnie wykorzystywane do planowania przestrzennego, projektowania tras drogowych i kolejowych, bądź do analizowania rozprzestrzeniania się hałasu i zanieczyszczeń czy wielu innych zastosowań.

Wykorzystanie technologii sztucznej inteligencji i uczenia maszynowego w miastach umożliwi gromadzenie danych w czasie zbliżonym do rzeczywistego oraz zapewni głębsze zrozumienie sposobu, w jaki organizmy miejskie funkcjonują i zmieniają się, dostosowując się do różnych warunków. Powinno to zapewnić lepszą jakość usług miejskich [Allam, Dhunny, 2019]. W kolejnej części opisano wybrane korzyści i rozwiązania.

8.3. Możliwości, korzyści i rozwiązania

Jednym z najbardziej powszechnych sposobów wykorzystania IoT i sztucznej inteligencji w celu usprawnienia życia w miastach jest inteligentne zarządzanie ruchem drogowym. Zainstalowanie czujników na powierzchni dróg i kamer przemysłowych pozwala miastom monitorować ruch uliczny w czasie rzeczywistym i powiadamiać użytkowników o zakłóceniach i alternatywnych trasach przejazdu. Rozwiązanie to może być powiązane z dodatkowymi priorytetami dla transportu publicznego (osobne pasy drogowe dla autobusów komunikacji miejskiej, zielona fala dla tramwajów itp.).

Inne zastosowania SI połączonej z IoT dotyczące ruchu drogowego to np. inteligentne parkowanie lub inteligentne oświetlenie. Czujniki zainstalowane na miejscach parkingowych pozwalają łatwo określić, czy dane miejsce jest wolne, czy zajęte. Z kolei latarnie miejskie mogą optymalizować natęże-

nie światła w zależności od pory dnia, warunków pogodowych oraz obecności pieszych, rowerzystów lub samochodów. Takie rozwiązania zmniejszają koszty, a zwiększają bezpieczeństwo [Lee i in., 2016].

Jest to bezpośrednio związane z innym przykładem wykorzystania SI i IoT w inteligentnych miastach – z inteligentnymi siłami porządkowymi, ułatwiając utrzymanie porządku publicznego [Joh, 2019]. Sieci kamer i czujników pomagają zapobiegać przestępczości. Umożliwiają one skuteczną identyfikację (technologia rozpoznawania twarzy) osób zachowujących się podejrzenie lub popełniających przestępstwo, monitorują gęstość tłumu, czystość obszarów publicznych, a także śledzą dokładny ruch pojazdów. Algorytmy predykcyjne pomagają przewidzieć, gdzie w przyszłości może dojść do przestępstwa [Fry, 2018], a także, które osoby mogą w przyszłości stać się ofiarami lub sprawcami przestępstw z użyciem broni.

Te same czujniki drogowe, parkingowe i latarnie mogą być wykorzystywane do monitorowania jakości powietrza i poziomu hałasu w zatłoczonych miejscach, parkach lub na ścieżkach fitness, tak aby ludzie zawsze mogli znaleźć najzdrowszą drogę do aktywności na świeżym powietrzu [Zanella i in., 2014]. Wszystkie czujniki miejskie mogą również służyć jako darmowe hotspoty sieci Wi-Fi. Z kolei w czasie lockdownu sieci czujników i kamer miejskich pozwalały, na przykład, zdalnie sprawdzać temperaturę za pomocą kamer termowizyjnych, czy wykorzystywać mapy cieplne w czasie rzeczywistym do kontrolowania tłoku w przestrzeni publicznej [Agostino i in., 2020].

Wykorzystanie rozwiązań sztucznej inteligencji może również zapewnić zrównoważony i inteligentny system zarządzania odpadami, przynoszący oszczędności i korzyści ekologiczne. Czujniki zainstalowane na pojemnikach na odpady mogą wykrywać poziom ich załadowania i wysłać powiadomienia do centrów wysyłających ciężarówki po odpady. Ponadto może to pozwolić na optymalizację tras przejazdu ciężarówek zbierających odpady, a tym samym obniżyć koszty ich zbiórki [Nuortio i in., 2006].

W kontekście zarówno pandemii, jak i szybkiego starzenia się społeczeństwa UE, duże znaczenie ma także zwiększenie dostępu do wysokiej jakości opieki zdrowotnej o przystępnej cenie. Technologia SI umożliwi inteligentną opiekę zdrowotną – dzięki wykorzystaniu urządzeń noszonych na ciele, takich jak opaski fitness czy urządzenia typu smartwatch, pozwala na automatyzację diagnostyki medycznej i pomoc we właściwym czasie w przypadku zagrożenia życia [Cook i in., 2018]. Pacjenci mogą być monitorowani w czasie rzeczywistym, na przykład poprzez kontrolę, czy utrzymują swoją izolację w przypadku COVID-19, bądź umożliwienie diagnozowania ich stanu zdrowia przez lekarzy, naukowców i pracowników służby zdrowia – a to prowadzi do bardziej spersonalizowanych, lepszych diagnoz i terapii. Jedno-

częściej zapewnia to oszczędność środków finansowych i czasu zarówno dla pacjentów, jak i dla szpitali. Rozwiązania inteligentnego miasta mogą być również stosowane do wspierania aktywności i uczestnictwa starszych i/lub niepełnosprawnych obywateli oraz promowania zdrowego stylu życia lub inwestowania w technologie inteligentnego domu, aby umożliwić seniorom niezależność nawet w podeszłym wieku [Rocha i in., 2019].

Głównym celem stojącym za ideą inteligentnego miasta jest uczynienie życia w miastach bardziej komfortowym, bezpieczniejszym i wygodniejszym dla ich mieszkańców. Wymaga to również inteligentnego zarządzania, tj. wykorzystania danych i technologii do podejmowania przez władze miejskie bardziej efektywnych, świadomych i opartych na faktach decyzji oraz do usprawnienia współpracy i komunikacji z obywatelami, organizacjami pozarządowymi, przedsiębiorstwami i innymi interesariuszami. Inteligencja miasta wiąże się jednak z kilkoma ważnymi pytaniami i wyzwaniem.

8.4. Wyzwania – krytyczne spojrzenie na inteligentne miasta i sztuczną inteligencję

Rewolucja cyfrowa i pojawienie się inteligentnych miast może spowodować szczególne problemy dla osób o niskich kwalifikacjach. Cyfryzacja i automatyzacja, w połączeniu z zastosowaniami sztucznej inteligencji, sprawią, że ludzie przestaną być potrzebni nie tylko do wykonywania prac niewymagających wysokich kwalifikacji, ale także do wszystkich rutynowych zadań. Niektóre z tych efektów stały się widoczne podczas pandemii – np. automatyczne kasy w sklepach, autonomiczne dostawy itp. Edukacja może pomóc złagodzić ten proces, ale nie każdy chce, jest zainteresowany lub może zdobyć nowe umiejętności. Jednym z potencjalnych rozwiązań jest przejście na gospodarkę cyrkularną i miasta cyrkularne, z nowymi miejscami pracy tworzonymi w „zielonych branżach” [TWI2050, 2019].

Problemy głównie osób o niskich kwalifikacjach wiążą się z pojęciem „wykluczenia cyfrowego” [TWI2050, 2019, s. 25], które dotyczy nierówności w „...dostępie do technologii cyfrowych i korzystaniu z nich...”, w korzyściach z technologii cyfrowych, w wiedzy i we władzy. Kolejną grupą, której dotyczą te problemy, są osoby starsze [Mitzner i in., 2010]. Było to dobrze widoczne podczas pandemicznego lockdownu w pierwszej połowie 2020 roku, kiedy to wymuszona izolacja spowodowała nagłe przeniesienie wielu elementów życia do sfery cyfrowej – np. opieki zdrowotnej, usług administracji publicznej,

zakupów itp. W rzeczywistości cyfryzacja może zmniejszyć te problemy, ale musi być dobrze zarządzana. Aby przezwyciężyć wykluczenie cyfrowe, rządy muszą wykazać się większą aktywnością w tworzeniu odpowiedniej infrastruktury, edukowaniu ludzi i zapewnianiu odpowiednich regulacji prawnych.

Innym wyzwaniem związanym z inteligentnymi miastami, a w szczególności ze starzejącymi się populacjami, jest to, że choć inteligentne miasto może rozwiązać wiele problemów starzejącego się świata, wiele starszych osób nie widzi potrzeby korzystania z tak zaawansowanych technologii lub wręcz się ich boi. Wydaje się, że aby zaakceptowali technologię, trzeba ich przekonać, że jest ona przydatna i łatwa w użyciu [Mitzner i in., 2016]. Dlatego inteligentne miasta i przemysł muszą oferować technologie i narzędzia dopasowane do potrzeb i możliwości seniorów, choć takie specjalnie zaprojektowane urządzenia mogą być dość kosztowne. Powinny także zaplanować kampanie reklamowe i edukacyjne, w tym możliwość doświadczenia technologii, oraz zaangażować osoby starsze w ich planowanie i rozwój. Jednak nawet jeśli osoby starsze widzą przydatność nowych usług i wiedzą, jak z nich korzystać, mogą je odrzucić, ponieważ korzystanie z nich oznacza przyznanie się do tego, że jest się starym, a to może być stygmatyzujące [Gopnik, 2019].

Z kolei gromadzenie i przetwarzanie dużych ilości danych wiąże się z kwestią ich jakości i bezpieczeństwa. Duża skala danych nie zawsze idzie w parze z ich jakością. Jeśli dane, na których uczą się algorytmy sztucznej inteligencji, są nieobiektywne, modele wynikowe również będą nieobiektywne i mogą prowadzić do błędnych decyzji. Dlatego, aby podjąć w pełni odpowiedzialną i świadomą decyzję, należy wiedzieć, na jakiej bazie danych algorytmy były trenowane i czego się nauczyły. Od lat liczne raporty ProPublicy czy New York Times ujawniają skalę algorytmicznego obciążenia (dyskryminacji) np. w ocenie ryzyka kryminalnego, przewidywania przestępczości [Selbst, 2017], udzielaniu kredytów, zatrudnianiu itp. [O'Neil, 2016]. Zrozumienie „wnioskowania” algorytmów jest niezbędne do budowania zaufania, które z kolei jest konieczne, aby ludzie zaakceptowali automatyzację w kolejnych sferach życia [Morley i in., 2019]. Jest to też kluczowe dla ochrony praw i wolności obywatelskich. Gdy operatorzy systemów i obywatele wiedzą, jak działa algorytm, zmniejsza się szansa, że będzie on odzwierciedlał stereotypy, uprzedzenia i błędy poznawcze. A duża część algorytmów sztucznej inteligencji to czarne skrzynki – nie pozwalają na interpretację zależności między danymi wejściowymi a wynikiem (decyzją). W ostatnich latach rozwijane są jednak nowe narzędzia zwiększające wyjaśnialność algorytmów uczenia maszynowego (np. [Holzinger i in., 2017; Adadi, Berrada, 2018; Samek i in., 2019; Biecek, Burzykowski, 2019]). Pozwalają one wyciągać wnioski na temat zależności między poszczególnymi cechami a zmienną docelową oraz pokazywać, dlaczego dla danej obserwacji zaprognozowano konkretną wartość.

Zbieranie i przetwarzanie informacji w czasie rzeczywistym przy użyciu wielu różnych urządzeń rodzi problem zachowania prywatności i ochrony systemu przed ryzykiem cyfrowym. Prywatność jest uważana za podstawowe prawo człowieka w krajach demokratycznych, w związku z czym jest chroniona prawnie i konstytucyjnie [Diggelmann, Cleis, 2014]. Jednak ciągły postęp technologiczny może oznaczać nowe możliwości dla osób naruszających prywatność. Dobrym przykładem jest rozwój oprogramowania śledzącego związanego z zachorowaniami na COVID-19. Wiele państw i firm stworzyło własne aplikacje, przy czym niektóre z nich budziły wątpliwości co do bezpieczeństwa programu i pełnej anonimowości danych. Dlatego potrzebne są nowoczesne rozwiązania zabezpieczające, które przewidują zmiany technologiczne, jak na przykład technologia blockchain [Ølnes, Jansen, 2017], czy w przyszłości technologie kwantowe.

8.5. Koszty

Koszty stworzenia inteligentnego miasta stanowią największą przeszkodę w realizacji projektu, przynajmniej dla większości władz miejskich. Zdają sobie one sprawę, że istnieją również inne wyzwania, o których wspomnieliśmy powyżej, ale ponieważ większość miast znajduje się w początkowej fazie inteligentnej transformacji, prawdopodobnie będą one zajmować drugorzędne miejsce. Nie da się oszacować pełnego kosztu przekształcenia miasta w smart city. Istnieją jednak pewne dane, do których możemy się odwołać. Na przykład Pitroda i Mialilhe [2017, s. 3] podają, że: „...rynek SI ma wzrosnąć do 40 mld dolarów rocznie do 2020 roku...”, natomiast według Zanelli i in. [2014, s. 23]: „...rynek Smart City jest szacowany na setki miliardów dolarów do 2020 roku, z rocznymi wydatkami sięgającymi prawie 16 mld...”. Podczas gdy niektóre projekty mogą i będą kosztować miliardy, inne będą kosztować znacznie mniej, jak na przykład stworzenie portalu e-usług dla Warszawy, który kosztował około 2 mln euro [Warszawa, 2017]. Ogólnie rzecz biorąc, według Caragliu i Del Bo [2019, s. 375] w ostatnich latach w Unii Europejskiej średni koszt projektów typu „smart city” wynosił około 10 mln euro, przy czym połowę tej kwoty zapewniała UE.

Takie małe projekty, które mogą mieć znaczący wpływ na jakość życia, są w zasięgu możliwości większości miast. Miasto musi jednak wziąć pod uwagę nie tylko koszty nowego sprzętu i oprogramowania, ale także zatrudnienia specjalistów, którzy będą potrafili je obsługiwać. Władze miejskie muszą być w stanie podejmować decyzje i formułować plany na podstawie danych, które

zbiera dla nich technologia [Inclezan, Pradanos, 2017; Allam, Dhunny, 2019]. Wiele miast przyznaje jednak, że nie jest w stanie tego zrobić, ponieważ nie są w stanie tych danych analizować [Smartpolis, 2018; Bunders, Varró, 2019]. Wymaga to specjalistycznej wiedzy, którą wybieralni urzędnicy niekoniecznie posiadają, dlatego muszą korzystać z usług specjalistów. Koszty dodatkowo zwiększa brak wspólnych standardów, bądź kompatybilności pomiędzy różnymi technologiami i systemami. Prowadzi to do nieefektywności i marnowania zasobów, ponieważ władze miejskie w swoich projektach próbują oceniać, szacować i łączyć różne rozwiązania [Li i in., 2018; Smartpolis, 2018].

Rozważając projekty inteligentnych miast, władze miejskie muszą wziąć pod uwagę dwa problemy: jak sfinansować nowe projekty inwestycyjne i jak je utrzymać, gdy już będą funkcjonować. W wielu krajach UE dostępność funduszy europejskich sprawia, że są one głównym źródłem finansowania tych projektów, wypierając inne opcje finansowe [Smartpolis, 2018]. Paradoks funduszy unijnych polega jednak na tym, że łatwiej może być coś zbudować niż pokryć koszty działania. Jak zauważyła Saxe [2019], jest to niezwykle istotne, ponieważ technologie szybko się starzeją i być może co kilka lat trzeba będzie je wymieniać, co drastycznie zwiększa koszty inteligentnych miast. Zwłaszcza małe miasta mogą nie być w stanie ponieść takich kosztów. Dlatego projektów tych często nie da się zrealizować bez udziału sektora prywatnego, który dysponuje nie tylko kapitałem, ale i niezbędną wiedzą. McKinsey [2018] ocenia, że 60% początkowych inwestycji w aplikacje smart city może pochodzić od podmiotów prywatnych. W każdym przypadku zaangażowania sektora prywatnego władze miejskie muszą podjąć decyzję o metodzie płatności. Trudność polega na oszacowaniu przyszłych przychodów z projektu. Według Deloitte [2018, s. 10] cztery najbardziej efektywne modele przychodów to: reklamy w usłudze, subskrypcje, opłaty od użytkowników oraz sprzedaż danych gromadzonych przez serwis, co też rodzi dodatkowe problemy, w tym etyczne.

8.6. Rola rządów, przepisów i zaleceń UE dotyczących polityki

Jednym z problemów związanych z inteligentnymi miastami jest to, kto zapewni technologie i narzędzia do ich wykorzystania. Nie jest to problem ludzi bogatych, którzy prawdopodobnie najwięcej zyskają na tych procesach, ale dla osób ubogich i starszych jest to bardzo ważne. Dlatego, jeśli nie chcemy, by byli wykluczeni ze wszystkich korzyści płynących z inteligentnych miast, inne podmioty muszą zapewnić im dostęp do tych korzyści. Może to

zrobić zarówno sektor publiczny, tj. rząd centralny lub regionalny, władze miejskie, jak i sektor prywatny, co często oznacza gigantów technologicznych. W obu przypadkach potrzebne są przepisy zapewniające powszechny dostęp, gwarantujące prywatność oraz chroniące jednostki przed piractwem internetowym i niewłaściwym wykorzystaniem informacji uzyskanych przez dostawców rozwiązań.

Potrzebne są nowe sposoby dzielenia się wiedzą i informacjami o różnych pomysłach i rozwiązaniach między gminami, regionami i krajami, a być może nawet zasady zmuszające je do dzielenia się nimi. Stanowi to pewną ochronę przed sytuacją, w której niektóre gminy będą próbowały rozwijać się kosztem innych, w szczególności w zakresie sztucznej inteligencji [Munoz, Naqvi, 2017; Vinuesa i in., 2019]. Istnieje oczywiste zapotrzebowanie na europejskie platformy wymiany informacji, dzięki którym miasta mogą dowiedzieć się, co robią gminy we wszystkich innych krajach UE i wybrać rozwiązania najlepiej dostosowane do ich potrzeb. UE powinna także pomagać rządów krajowym we wspieraniu finansowym inicjatyw smart city, zwłaszcza w regionach słabiej rozwiniętych i uboższych. Ponadto, w przypadku takiego wsparcia należy odejść od często preferowanej zasady, że „jedno rozwiązanie pasuje do wszystkich” [Szlachta, Zaleski, 2017]. Wręcz przeciwnie, w tej dziedzinie każde rozwiązanie powinno być dopasowane do sytuacji konkretnej gminy i jej planów rozwojowych.

W przypadku świadczenia usług dla sektora publicznego konieczne może być wprowadzenie nowych zasad dotyczących zamówień publicznych, aby zapewnić dostęp do optymalnego poziomu technologii przy minimalnych kosztach. Przydatne w tym względzie byłyby wspólne zasady dla całej Unii Europejskiej, w szczególności poprzez promowanie rozwiązań opracowywanych przez małe, lokalne firmy i organizacje pozarządowe oraz ułatwianie im konkurencji na terenie całej Unii. Ponadto sektor publiczny będzie musiał wykorzystywać nowe rodzaje wiedzy i kompetencji, co w praktyce oznacza konieczność płacenia więcej swoim pracownikom, aby być konkurencyjnym na rynku pracy. Bez takich ekspertów wysiłek włożony w regulację nowych rynków i nowych technologii może zakończyć się niepowodzeniem, gdyż jak to ujęli Vinuesa i in. [2019, s. 8]: „...nadzór regulatora powinien być poprzedzony wiedzą regulatora...”.

W przypadku świadczenia usług przez sektor prywatny oparcie się na zasadach rynkowych jest oczywistym sposobem na wykluczenie części społeczeństwa z korzyści, jakie przynoszą inteligentne miasta. Dlatego rząd powinien interweniować na różnych szczeblach, regulując i dotując działalność firm prywatnych, aby obniżyć ceny dla klientów indywidualnych i zwiększyć podaż. Jest jednak jeszcze jeden aspekt tych procesów, który powinniśmy rozważyć, choć jest on bardzo kontrowersyjny – mianowicie regulowanie

działalności gigantów technologicznych i stosowanie wobec nich przepisów antymonopolowych. Według Mialhe i Hodesa [2017] obecne trendy sugerują, że w przyszłości SI będzie kontrolowana przez globalny oligopol złożony z kilkunastu międzynarodowych koncernów. Nie jest to jednak tylko problem sztucznej inteligencji, ponieważ nawet bez niej to właśnie te giganty technologiczne dostarczą większość technologii i narzędzi niezbędnych do rozwoju inteligentnych miast. W tym procesie zyskają ogromną władzę, ponieważ miasta i ich mieszkańcy będą od nich całkowicie zależni. Logicznym wydaje się, że rozwiązanie tego problemu powinno zostać wypracowane na poziomie europejskim, a nie przez poszczególne kraje.

Bardzo ważną rolą dla miast, rządów i UE jest edukacja obywateli i opracowanie nowej strategii edukacyjnej dla ludzi żyjących w nowym „inteligentnym” świecie. Obecne systemy edukacyjne nie przygotowują ludzi do takiej przyszłości. To prawda, że dzisiejsze dzieci będą znacznie lepiej przystosowane do życia z technologią niż dzisiejsi dorośli, ale zachodzące obecnie procesy będą wymagały zupełnie nowego poziomu umiejętności technologicznych. Bez niej wiele osób zostanie wykluczonych z korzyści płynących z rozwoju technologicznego i będzie się czuło/stanie się zbędnymi w inteligentnych miastach. Musimy już teraz pracować nad nowymi metodami edukacji, aby zapobiec dyskryminacji w niedalekiej przyszłości.

8.7. Podsumowanie

Koncepcja inteligentnego miasta jest odpowiedzią na wyzwania współczesnego świata, choć jako całość wydaje się raczej futurystyczną wizją niż rzeczywistością, którą można zastosować w każdym mieście. Mimo to wybrane rozwiązania są z powodzeniem wdrażane w światowych metropoliach i mniejszych miastach, a pandemia COVID-19 była okazją do zaobserwowania, że proces ten może postępować bardzo szybko. Jednak wiele z nich rodzi nowe obawy i problemy. Ważną rolą naukowców i decydentów jest proponowanie rozwiązań, które odpowiadają na rzeczywiste potrzeby mieszkańców miast, a jednocześnie są odpowiedzią na globalne wyzwania cywilizacyjne, a także zapewnienie bezpieczeństwa proponowanych narzędzi. Istotną barierą dla rozwoju tego typu technologii są wciąż wysokie koszty i brak regulacji prawnych. Niskie koszty wdrożenia (także w sensie wpływu na środowisko) są kluczowym czynnikiem trwałości rozwiązań typu smart city. Muszą być one także bezpieczne, niezawodne, skalowalne, integrujące i przejrzyste dla obywateli. W związku z tym inteligentne miasto może przynieść szereg korzyści w zakresie zarządzania i optymalizacji tradycyjnych usług publicznych w miastach.

BIBLIOGRAFIA

- Adadi A., Berrada M., (2018), *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)*, IEEE Access, Open Access Journal.
- Agostino D. i in., (2020), *New development: COVID-19 as an accelerator of digital transformation in public service delivery*, Public Money & Management.
- Allam Z., Dhunny Z.A., (2019), *On big data, artificial intelligence and smart cities*, Cities, 89, s. 80–91.
- Barro P. i in., (2018), *Towards Smart and Sustainable Future Cities Based on Internet of Things for Developing Countries: What Approach for Africa?*, EAI Endorsed Transactions on Internet of Things, nr 4.
- Biecek P., Burzykowski T., (2019), *Predictive Models: Explore, Explain, and Debug Human-Centered Interpretable Machine Learning*, https://pbiecek.github.io/PM_VEE.
- Bunders D.J., Varró K., (2019), *Problematizing data-driven urban practices: Insights from five Dutch 'smart cities'*, Cities, 93, s. 145–152.
- Caragliu A., Del Bo C.F., (2019), *Smart innovative cities: The impact of Smart City policies on urban innovation*, Technological Forecasting & Social Change, 142, s. 373–383.
- Deloitte, (2018), *The challenge of paying for smart cities projects*, Deloitte Development LLC.
- Diggelmann O., Cleis, M.N., (2014), *How the right to privacy became a human right*, Human Rights Law Review, 14, s. 411–458.
- Eldawy A., Mokbel M.F., (2016), *The era of big spatial data: A survey*, Foundations and Trends in Databases, 6(3–4), s. 163–273.
- Fry H., (2018), *Hello World: How to be a Human in the Age of the Machine*, Doubleday.
- Gopnik A., (2019), *Can we live longer but stay younger*, The New Yorker, 13.05.2019.
- Holzinger A. i in., (2017), *What do we need to build explainable AI systems for the medical domain?*, arXiv:1712.09923.
- Incelezan D., Pradanos L.I., (2017), *Viewpoint: A Critical View on Smart Cities and AI*, Journal of Artificial Intelligence Research, 60, s. 681–686.
- Joh E., (2019), *Policing the smart city*, International Journal of Law in Context, 15(2), s. 177–182.
- Kosiński i in., (2013), *Private traits and attributes are predictable from digital records of human behavior*, PNAS, 110(15), s. 5802–5805.
- Lee C. i in., (2016), *Smart parking system for Internet of Things*, IEEE International Conference on Consumer Electronics (ICCE), IEEE2016, s. 263–264.
- McKinsey, (2018), *Smart cities: digital solutions for a more livable future*, McKinsey Global Institute, czerwiec 2018.
- Miaillhe N., Hodes C., (2017), *The Third Age of Artificial Intelligence*, Field Actions Science Reports [Online], Special Issue 17, <http://journals.openedition.org/factsreports/4383>.
- Mitzner T.L. i in. (2010), *Older adults talk technology: Technology usage and attitudes*, Computers in human behavior, 26(6), s. 1710–1721.
- Mitzner T.L. i in. (2016), *Predicting older adults' perceptions about a computer system designed for seniors*, Universal Access in the Information Society, 15(2), s. 271–280.
- Morley J. i in., (2019), *From What to How: An Initial Review of Publicly Available AI Ethics Tools, Methods and Research to Translate Principles into Practices*, Science and Engineering Ethics.
- Munoz M.J., Naqvi A., (2017), *Artificial Intelligence and Urbanization: The rise of the Elysium City*, Economics and Political Economy, 4(1), s. 1–13.
- Nuortio T. i in., (2006), *Improved route planning and scheduling of waste collection and transport*, Expert Systems with Applications, 30, s. 223–232.

- Ølnes S., Jansen A. (2017), *Blockchain Technology as a Support Infrastructure in e-Government*, [w:] Janssen M. i in. (red.) *Electronic Government. EGOV 2017, Lecture Notes in Computer Science*, vol. 10428. Springer, Cham.
- Pitroda S., Miaillhe N., (2017), *Introduction. The rise of AI & Robotics in the City*, *Field Actions Science Reports* [Online], Special Issue 17.
- Rocha N. i in., (2019), *A Systematic Review of Smart Cities' Applications to Support Active Ageing*, *Procedia Computer Science*, nr 160, s. 306–313.
- Ryan M., Gregory A., (2019), *Ethics of Using Smart City AI and Big Data: The Case of Four Large European Cities*, *ORBIT Journal*, 2(2).
- Samek W. i in. (red.) (2019), *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, Springer, Series: Lecture Notes in Artificial Intelligence.
- Saxe S., (2019), *I'm an Engineer, and I'm Not Buying Into 'Smart' Cities*, *The New York Times*, 16.07.2019.
- Selbst A.D., (2017), *Disparate Impact in Big Data Policing*, *Georgia Law Review*, 52(1), s. 109–196.
- Smartpolis, (2018), *Evaluation of V4 cities in the area of Smart City*, https://smartpolis.eit.bme.hu/?page_id=153 (dostęp: 5.01.2020).
- Szlachta J., Zaleski J., (2017), *Challenges for Polish Regional Policy in the Context of EU Cohesion Policy after 2020*, *Research Papers of Wrocław University of Economics*, nr 498.
- TWI2050, *The World in 2050* (2019), *The Digital Revolution and Sustainable Development: Opportunities and Challenges*, raport The World in 2050 initiative, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria. www.twi2050.org
- Vinuesa R. i in., (2019), *The role of artificial intelligence in achieving the Sustainable Development Goals*, <https://arxiv.org/ftp/arxiv/papers/1905/1905.00501.pdf>.
- Winkowska J. i in., (2019), *Smart city concept in the light of the literature review*, *Engineering Management in Production and Services*, 11(2), s. 70–86.
- Zanella A. i in., (2014), *Internet of Things for Smart Cities*, *IEEE Internet of Things Journal*, 1(1).

Rozdział 9

Estetyka spekulatywna, percepcja maszynowa i sztuczna agencyjność w ujęciu organizacyjnym

Speculative aesthetics, machine perception and artificial agency in organizational perspective

Adam Dzikowski

Politechnika Wrocławska
Wydział Zarządzania

Streszczenie. Estetyka spekulatywna, która wywodzi się z nowej estetyki, realizmu spekulatywnego i ontologii zorientowanej na przedmiot, stawia pytania o percepcję maszynową i postrzeganie nie-ludzi. W szczególności odnosi się ona do procesów, w których maszyny nie tyle wspierają, czy nawet zmieniają ludzkie działania, co je zastępują, a w ostatecznym rezultacie eliminują człowieka z procesu decyzyjnego. W kontekście sztucznej inteligencji rozważania nad nie-ludzką percepcją poszerzają kwestie sprawczości sztucznych agentów i zaufania do nich o wymiary leżące poza ludzkimi zmysłami i doświadczeniem. Rozważania te wydają się szczególnie istotne w czasach, gdy istnienie organizacji coraz częściej zależy od syntetycznych bytów (jak sztuczne sieci neuronowe), które podejmują samodzielne decyzje, często na podstawie bodźców zupełnie odmiennych niż ludzkie. Pierwsze analizy prowadzone w tym zakresie pokazują, jak odmiennie postrzegają świat maszyny, że mogą one ulegać niezrozumiałym dla nas iluzjom oraz jak można je oszukać, co stanowi przyczynek do poszerzonych rozważań nad zaufaniem i etyką interakcji ze sztucznymi bytami.

Abstract. Speculative aesthetics, which derives from new aesthetics, speculative realism, and object-oriented ontology, raises questions about machine and non-human perception. In particular, it refers to processes in which machines do not so much support or even alter human actions as replace them, and ultimately eliminate humans from the decision-making process. In the context of artificial intelligence, considerations of non-human perception extend issues of artificial agency and trust to dimensions that lie beyond human senses and experience. These considerations seem particularly relevant at a time when organizations increasingly depend on synthetic entities (like artificial neural networks) that make independent

decisions, often based on stimuli quite different from human ones. The first analyses conducted in this area show how differently machines perceive the world, that they can succumb to illusions that are incomprehensible to us, and how they can be fooled, which contributes to extended considerations of trust and the ethics of interaction with artificial entities.

Słowa kluczowe: estetyka spekulatywna, sztuczna agencyjność, estetyka organizacji, posthumanizm, teoria organizacji

Keywords: speculative aesthetics, artificial agency, organizational aesthetics, posthumanism, organizational theory

9.1. Wprowadzenie

We współczesnym rozumieniu estetyka jest określeniem wieloznacznym, ale dającym się sprowadzić do wspólnego źródła, które definiowane jest m.in. przez piękno, formę, dobry projekt, zmysłowość, harmonijność, subtelność i ponadczasowość [Welsch, 2005, s. 52–58]. Pomimo braku jednoznacznych i uniwersalnych wyznaczników piękna [por. Weiner, 1994], wiele wskazuje na to, że istnieją pewne wspólne, ogólnoludzkie uwarunkowania sądów estetycznych [Dutton, 2002]. Nawet jeżeli nie przyjmują one postaci konkretnych kategorii, to wydają się być uniwersalne w sposobie ontologicznego podejścia do piękna. Mimo kulturowych różnic większość ludzi potrafi bowiem rozpoznać dzieło artysty, pracę twórczą, czy efekt biegłego rzemiosła [Nelson, 2005]. Z kolei w sferze gospodarczej wciąż istotne są pojęcia *mimesis* i *techne*, a doskonałość techniczna, przyjemność użytkowania, spójność, harmonia czy wyjątkowość produktów często stanowią o atrakcyjności oferty danego przedsiębiorstwa [Hekkert, 2006, s. 157–172]. Tym samym, niezależnie od indywidualnie czy też kulturowo wyznawanych teorii piękna, trudno zignorować wpływ doznań estetycznych na zachowania społeczne i organizacyjne [Dżidowski, 2011b, s. 53].

Dlatego też, od kilkunastu lat w ramach teorii organizacji rozwija się podejście, w którym kryteria, sądy i doświadczenia estetyczne są traktowane jako istotne elementy działań organizacyjnych. Nie tylko z powodu postępującej estetyzacji świata [Welsch, 1997], ale przede wszystkim dlatego, że polisemantyczne znaczenia estetyki [Welsch, 1996], rozważane na polach percepcji, kultury i piękna, mają swój przemożny wpływ na rozumienie współczesnych struktur społecznych. W rezultacie takich rozważań, od lat 80. poprzedniego wieku pojawiają się studia nad estetyką organizacji (por. [Strati, 1999; Linstead, Höpfl, 2000; De Monthoux, 2004; Taylor, Hansen, 2005; Hatch, Koste-

ra, Kozminski, 2005]). Oznacza to, że organizacje nie są już traktowane jako byty estetycznie neutralne, ale jako formy ekspresji, podatne na wielorakie aspekty ludzkiej percepcji, interpretacji i reakcji [Dzidowski, 2018]. Estetyka organizacji jest przy tym uznawana za jedno z bardziej inspirujących podejść do teorii organizacji i stawiana na równi z takimi koncepcjami, jak krytyczny realizm, teoria złożoności, teoria aktora–sieci, czy konstrukt tożsamości organizacyjnej (por. [Hatch, Cunliffe, 2006; McAule, Duberley, Johnson, 2007]). Dostrzeżenie wymienionych wymiarów daje nową i ciekawą perspektywę badawczą, która może się okazać pomocna przy analizie ewolucji współczesnych struktur i procesów organizacyjnych, które często są wynikiem zmian w percepcji rzeczywistości organizacyjnej.

Pamiętać należy przy tym, że sam termin „organizacja” jest również pojęciem wieloznacznym, które można rozpatrywać z kilku perspektyw (m.in. atrybutowej, rzeczowej, czynnościowej i podmiotowej). W naukach społecznych często definiuje się organizację poprzez jej celowość i wymiar ludzki. Do takiego rozumienia odwołują się klasyczne definicje organizacji, jak chociażby R.W. Griffina [1996]: „grupa ludzi, którzy współpracują ze sobą w sposób uporządkowany i skoordynowany, aby osiągnąć pewien zestaw celów”. Poważniejszym wyzwaniem okazuje się jednak stopniowa eliminacja z definicji organizacji, zwłaszcza gospodarczych, określeń „grupa ludzi” czy „struktura społeczna”, na rzecz grupy „podmiotów”, „agentów decyzyjnych”, czy wręcz „obiektów”. Oznacza to, że technologie informacyjno-komunikacyjne wpływają na ewolucję struktur organizacyjnych nie tylko poprzez rozwój funkcji przetwarzania i przekazywania informacji, ale też zaczynają zmieniać samą istotę podmiotowości w organizacjach.

Wyzwaniem, przed jakim stoi obecnie teoria i estetyka organizacji, jest kwestia powstawania nowych struktur organizacyjnych, a w szczególności roli, jaką pełni w tych procesach percepcja nie-ludzi, która do tej pory nie była brana pod uwagę ani w rozważaniach estetycznych [Dzidowski, 2017b], ani organizacyjnych [Dzidowski, 2017a].

9.2. Posthumanizm, wirtualność i sztuczna agencyjność

Posthumanizm to prąd myślowy, w którym odrzuca się centralną rolę człowieka w opisie i przeżywaniu świata na rzecz „demokracji bytów”. Katherine Hayles w książce *How We Became Posthuman* [2008] wskazuje cztery założenia, na których opiera się pojęcie posthumanizmu. Po pierwsze, sygnały komunikacyjne są bardziej istotne od ich materialnych nośników, w tym ludz-

kiego ciała i jego zmysłów. Po drugie, świadomość nie jest początkiem ani centrum życia ludzkiego, lecz jedynie epifenomenem. Po trzecie, posthumaności postrzegają ciało jako protezę, którą człowiek nauczył się wykorzystywać i nic nie stoi na przeszkodzie, aby ją zmienić, bądź ulepszać za pomocą innych protez (tzw. transhumanizm). Po czwarte, byt można zawrzeć w maszynie. W konsekwencji nie ma znaczących różnic pomiędzy cielesnością człowieka a komputerową symulacją, cybernetycznym mechanizmem a biologicznym organizmem, zasadami działania algorytmu a celowością ludzkich działań. Trudno przy tym nie zauważyć, że we współczesnym świecie nie-ludzkie byty zyskują charakter podmiotowy. Coraz częściej zależymy od syntetycznych systemów, które podejmują samodzielne decyzje, a założenia posthumanizmu zdają się mieć bezpośrednie przełożenie na istnienie współczesnych organizacji [Dżidowski, 2017a].

Ostatnie dekady XX wieku, naznaczone rozwojem społeczeństwa informacyjnego, rynków globalnych i hiperkonkurencji, doprowadziły do zmiany w pojmowaniu istoty procesów organizacyjnych, zwiększając presję na efektywność i elastyczność działań. Redukcja zatrudnienia, outsourcing strategiczny, organizacje zwinne i tym podobne trendy doprowadziły ostatecznie do powstania nowych struktur organizacyjnych, które odpowiadały potrzebie szybkiego gromadzenia zasobów i umiejętności, bez obciążania sztywnymi i formalnymi relacjami. Bazując na teorii złożoności, a także na kontraktowych i behawioralnych koncepcjach przedsiębiorczości, doprowadzono do przededefiniowania klasycznej definicji organizacji gospodarczej. Doskonałym przykładem zmian zachodzących we współczesnych strukturach organizacyjnych jest przypadek organizacji wirtualnych.

Już samo pojęcie wirtualności niesie ze sobą istotne problemy poznawcze [Welsch, 2000]. Operując w takich polach semantycznych, jak potencjalność, dynamizm, pluralizm, konstruktywizm, realność, sztuczność czy cyfrowość, podważa ono tradycyjne postrzeganie rzeczywistości. Podobnie jest z pojęciem organizacji wirtualnych, które redefiniują klasyczne rozumienie własności, kontroli i celów organizacji. Pod pojęciem organizacji wirtualnej rozumie się bowiem tymczasową koalicję niezależnych podmiotów gospodarczych, których struktura podlega stałej reorganizacji, a zakres i cel działalności jest ściśle podporządkowany wykorzystywaniu pojawiających się okazji rynkowych [Dżidowski, 2011a]. Podążając za wirtualnym wzorcem, wiele współczesnych przedsiębiorstw zostało zdekomponowanych do kluczowych czynności biznesowych, a wszystkie inne procesy są delegowane do wyspecjalizowanych podwykonawców, tworzących dynamiczne sieci biznesowe. Dzięki temu autonomiczni partnerzy mogą działać niczym jedna organizacja, spojona łańcuchem dostarczanej wartości, a nie formalną strukturą, dopóki

tylko będą uznawać to za ekonomicznie uzasadnione. Przykłady takich rozwiązań można znaleźć w branży kreatywnej, projektach z zakresu nowych technologii, aeronautyki, czy w ramach rozwiązań z zakresu handlu elektronicznego i e-biznesu.

Nowoczesne struktury organizacyjne, takie jak sieciowe, wirtualne czy fraktalne [Dzidowski, 2017a], nie są jednak w stanie przetrwać bez wsparcia zaawansowanych systemów informatycznych, które zarządzają produkcją, dostawami, sprzedają, koordynują działania czy wspomagają decyzje. Okazuje się przy tym, że komputerowo wspomagane technologie informacyjno-komunikacyjne (ICT), które są powszechnie stosowane we współczesnych przedsiębiorstwach, stały się na tyle ich integralną częścią, że nie można już oddzielić procedur technicznych od istoty działalności organizacji [Schulz-Schaeffer, 2001]. Przykładem mogą być tzw. agenci programowi, czyli algorytmy zdolne do komunikowania się, monitorowania swego otoczenia i podejmowania autonomicznych decyzji, aby osiągnąć cele określone podczas ich projektowania. Ich zastosowanie otwiera możliwość tworzenia wirtualnych organizacji hybrydowych, w których syntetyczni agenci podejmują działania w imieniu ludzkich interesariuszy. Podążając zaś dalej w tej wzajemnej integracji procesów organizacyjnych i informatycznych, możliwe się staje wprowadzenie interfejsów organizacyjnych na wzór interfejsów programistycznych, czyli tzw. API (ang. *Application Programming Interface*). Zadaniem API jest dostarczenie odpowiednich specyfikacji podprogramów, struktur danych, klas obiektów i wymaganych protokołów komunikacyjnych. Organizacyjne API to z kolei zestaw reguł, dzięki którym firmy mogą wchodzić w interakcje między sobą i wymieniać zasoby informacyjne [Iyer, Subramaniam, 2015]. Ich powstanie może doprowadzić do wyparcia tradycyjnych partnerstw biznesowych na rzecz zautomatyzowanego łączenia struktur organizacyjnych. Wiele z popularnych obecnie przedsięwzięć biznesowych jest w swojej istocie platformami umożliwiającymi klientom i partnerom automatyczne wchodzenie w predefiniowane relacje, tak jak ma to miejsce w przypadku platform e-biznesowych, takich jak Uber (taksówki), Airbnb (miejsca noclegowe) i AliExpress (handel). Podobnie mogą działać też inne przedsiębiorstwa, tworząc samoregulujące się struktury, poprzez udostępnianie zalgorytmizowanych reguł współpracy oraz otwartego i skalowalnego systemu informatycznego. Organizacje przyszłości mogą więc stać się strukturami społeczno-techniczno-informatycznymi, których istnienie i funkcjonowanie będzie wyłącznie oparte na automatycznych systemach negocjacji, podejmujących decyzje na podstawie dostępnych wartości zmiennych rynkowych. Co ciekawe, pierwsze kroki w tym kierunku są już podejmowane. W 2014 roku algorytm komputerowy o nazwie Vital stał się członkiem zarzą-

du w firmie Deep Knowledge Ventures (por. <http://www.bbc.com/news/technology-27426942>), odpowiedzialnym za decyzje inwestycyjne. Był to zabieg w dużej mierze marketingowy, gdyż wspomniana firma zajmuje się rozwiązaniami z zakresu biznesowego wykorzystania sztucznej inteligencji, jednak inne firmy, np. Cisco, IBM i General Electric, używają systemów eksperckich do zarządzania relacjami ze swoim personelem (rekrutacji, oceny i promowania) [Alsever, 2016]. Mając na uwadze wyłaniający się stan rzeczy oraz trendy rozwojowe w projektowaniu współczesnych organizacji, warto zastanowić się, jak percepcja maszynowa wpływa na procesy decyzyjne syntetycznych agentów organizacyjnych oraz jakie są tego konsekwencje w kontekście rozważań nad przyszłością postrzegania organizacji.

9.3. Nowa estetyka

We współczesnym środowisku organizacyjnym coraz częściej spotykamy elementy wykraczające poza ludzką percepcję, takie jak: kody QR, protokoły komunikacyjne używane w sieciach komputerowych czy systemy nadzoru procesów produkcyjnych operujące w widmie podczerwonym, ultrafiolecie, promieniach Roentgena czy używające kamer wysokiej prędkości. Przełamanie ludzkiego doświadczenia estetycznego nie ogranicza się jednak do sfery komunikacji i sensoryki. W wielu współczesnych organizacjach systemy sztucznej inteligencji, oparte np. na sieciach neuronowych czy algorytmach genetycznych, stają się autonomicznymi podmiotami decyzyjnymi. Póki co nie są one pełnoprawnymi bytami organizacyjnymi, ale często charakteryzują się dużą autonomią, a istnienie takich systemów, jak Transakcje Wysokich Częstotliwości (ang. *High Frequency Trading*, HFT), udowadnia, że ludzka percepcja nie jest już jedyną, którą należy rozpatrywać w kontekście gospodarczym. Wspomniane systemy potrafią analizować spadki i wzrosty globalnych kursów akcji w przedziale milisekund, podejmując przy tym decyzje służące zarobieniu ułamek centa w milionach transakcji dziennie (por. youtu.be/L5cZaIZ5bWc). Usprawiedliwione wydaje się więc pytanie czy nadal mamy do czynienia z wyłącznie ludzkim aspektem estetyki organizacji.

To pytanie jest ściśle związane z dyskusją, która rozpoczęła się około roku 2011, *nomen omen* w przestrzeni internetowych blogów i sieci społecznych. W tymże roku, James Bridle [2011] opublikował notatkę pod tytułem *The New Aesthetic*, powiązaną z blogiem o tym samym tytule (por. new-aesthetic.tumblr.com). Ideą, która mu przeświecała, było udokumentowanie „nowej estetyki przyszłości”. Stworzył on zbiór przykładów błędów,

pikselizacji i innych wizualnych anomalii generowanych przez nie-ludzkich aktorów, aby przedstawić obce i syntetyczne formy obrazowania. Podobnie do *Manifestu Futurystów* z 1909 roku, publikacja Bridle'a rozpoczęła ożywioną dyskusję na temat nieuniknionej „erupcji cyfrowości w fizyczność” [Sterling, 2012]. Z punktu widzenia procesów organizacyjnych najcenniejszym wydaje się jednak głos Michaela Betancourta [2013], który rozważa nową estetykę w kontekście cyfrowej automatyzacji i odnosi ją do koncepcji maszyny Karola Marksa zawartej we *Fragmencie o maszynach*. Betancourt umieszcza nową estetykę w opozycji do rzemieślniczej estetyki Morrisa i Ruskina z ruchu *Arts and Crafts*, zauważając przy tym, że o ile ta ostatnia została wyparta przez gładkie, błyszczące i masowe projekty modernistyczne, to modernizm wciąż był dziełem ludzkim. Z kolei nowa estetyka odzwierciedla proces, w którym maszyny nie tyle wspierają czy zmieniają ludzką pracę, ale ją zastępują, ostatecznie eliminując człowieka nawet z procesu decyzyjnego.

W tym momencie warto odnieść się do ontologii zorientowanej na przedmiot (ang. *object-oriented ontology*). Ian Bogost, współtwórca tego trendu filozoficznego (obok Grahama Harmana i Levi Bryanta), definiuje go, jako umieszczenie przedmiotów w centrum filozoficznych zainteresowań i przyjęcie, że nic w otaczającym nas świecie nie ma uprzywilejowanej pozycji, ale istnieje na równi z innymi bytami [Bogost, 2009]. Podobnie jak w przypadku posthumanizmu, postulowana jest przy tym „demokracja rzeczy” w ramach „płaskiej ontologii”, która ma nie tyle obalić dotychczasowe hierarchie, co raczej uznać inne. Levi Bryant stwierdza przy tym, że nie chodzi o to, by przestać myśleć o ludziach, ale zacząć myśleć o roli nie-ludzi w kształtowaniu społecznych relacji [Bryant, 2012].

To wysoce abstrakcyjne ujęcie rzeczywistości oznacza, że istnienie maszyn, systemów czy przedmiotów może mieć wymiar egzystencjalny, ale jednocześnie wymykający się ludzkiemu poznaniu. Otwartą kwestią pozostaje, czy refleksja ta pomoże w projektowaniu struktur organizacyjnych, w których autonomiczni agenci programowi, czy specyficzna dla danego przedsiębiorstwa sztuczna inteligencja, będą pełnoprawnymi uczestnikami procesów decyzyjnych [Dzidowski, 2015]. Co jednak ciekawe, z punktu widzenia rozważań na temat podmiotowości systemów cyfrowych w organizacjach, niektórzy autorzy dopatrują się wręcz związków współczesnej filozofii z technologiami informatycznymi. Z jednej strony, ontologia zorientowana na przedmiot jest podobna do programowania obiektowego i systemów wieloagentowych, czerpiąc z tej samej, zorientowanej informatycznie wyobraźni. Z drugiej strony, teoria zbiorów, topologia, teoria grafów, cybernetyka i teoria systemów wpłynęły na rozwój myśli takich autorów, jak Deleuze, Badiou czy Latour, i to w podobny sposób, jak na współczesną teorię organizacji. W zagmatwanych

dzielach Gillesa Deleuze'a oraz Félixu Guattari znaleźć można rozważania na temat wirtualności, deterytorializacji, kłączy (*rhizomes*) i przepływów, czyli idei znajdujących bezpośrednie odniesienia do problemów zarządzania współczesnymi strukturami organizacyjnymi (usieciowienie, fraktalizacja, płynność, rozproszenie).

9.4. Estetyka spekulatywna

O ile nowa estetyka uznaje równoprawne istnienie nie-ludzkiej estetyki, to estetyka spekulatywna pojawiła się jako próba odpowiedzi na pytania, jaka nie-ludzka estetyka może być. Estetyka spekulatywna jest przy tym częścią szerszego nurtu filozoficznego, jakim jest realizm spekulatywny [Bryant, Srnicek, Harman, 2011]. Realizm spekulatywny definiuje się w opozycji do korelacionizmu i wynikającego z niego antropocentryzmu [Meillassoux, 2008], uznając przy tym, podobnie jak ontologia zorientowana na przedmiot, że relacje człowieka ze światem są jedynie specjalnym przypadkiem relacji pomiędzy dowolnymi bytami. W tym też kontekście definiuje się pojęcie estetyki spekulatywnej.

Speculative Aesthetics Working Group na *Duke University* w Stanach Zjednoczonych określa estetykę spekulatywną, odwołując się zarówno do realizmu spekulatywnego, jak i do prac Deleuze'a, Whiteheada i Kanta, czy obcej fenomenologii Iana Bogosta [Blas, Rhee, 2010]. Z kolei uczestnicy *Speculative Aesthetics Research Project* na *University for the Creative Arts* w Wielkiej Brytanii definiują estetykę spekulatywną przez potrzebę odrzucenia wyłącznie ludzkiego postrzegania świata, przy jednoczesnym niepopadaniu w naiwny realizm, ucieleśnioną subiektywność czy (nowy) materializm [Trafford, 2014]. Efektem prac tych grup, jak i powiązanych z nimi środowisk, są dwie publikacje z 2014 roku, będące próbą podsumowania rozważań na temat estetyki spekulatywnej: *Speculations: A Journal of Speculative Realism // Issue V: Aesthetics in the 21st Century* [Askin i in., 2014] oraz *Speculative Aesthetics* [Mackay, Pendrell, Trafford, 2014].

Istotnym problemem, który nie doczekał się jednak jednoznacznego rozstrzygnięcia w ramach rozważań nad estetyką spekulatywną, jest sama definicja estetyki, która tradycyjnie obejmuje tylko ludzkie postrzeganie. O ile spekulatywny realizm przyjmuje inne perspektywy, to nadal pozostaje niewyjaśniona kwestia doświadczenia estetycznego nie-ludzi. Problem ten został częściowo przedstawiony przez Grahama Harmana w artykule *Aesthetics as First Philosophy: Levinas and the Non-Human* [2012]. Podstawowym zabie-

giem myślowym jest utożsamienie estetyki z „przyjemnością/korzystaniem” (ang. *enjoyment*, wg Levinas) lub „przyciąganiem/wabieniem” (ang. *allure*, wg Harmana), przez które obiekty „pławia” się jedne w drugich, co utożsamiane jest z doświadczeniem estetycznym. O ile można zauważyć, że kategorie „przyciągania”, „wabienia”, „korzystania” znajdują swoje miejsce w rozważaniach nad relacjami organizacyjnymi, zwłaszcza w perspektywie istnienia organizacyjnych API, to operowanie na proponowanych poziomach abstrakcji jest jednak niezwykle trudne. Oznacza to, że zrozumienie percepcji sieci neuronowych czy innych systemów sztucznej inteligencji, nie powinno polegać na pytaniu, jaka ona jest, ale raczej, jaka ona może być. Ian Bogost [2012] w książce *Alien Phenomenology or What It's Like to Be a Thing* pisze, że wszystko na równi istnieje, ale nie istnieje w równy sposób, a jedynym sposobem na przeprowadzenie obcej fenomenologii jest analogia. Jeżeli więc przyjmiemy propozycje wysuwane przez spekulatywny realizm, to obca percepcja może być określona jedynie przez metaforę i to metaforę samej percepcji, z której my, ludzie, zostajemy usunięci. Co ważne z punktu widzenia estetyki spekulatywnej, Bogost wprowadza pojęcie Obcej Estetyki (ang. *Alien Aesthetics*), która nie ma bynajmniej odnosić się do naszego postrzegania, ale przykładowo stawiać pytania o to, jak bambus kpi z innych traw, czy jaką poezję pisałyby komputery [Jackson, 2012].

Tego rodzaju projekcja może być szczególnie interesującym ćwiczeniem mentalnym w odniesieniu do estetyki nie-ludzkich organizacji [Dzidowski, 2015]. Obecnie całe ekosystemy biznesowe zaczynają przypominać obce formy życia. Na przykład model rozgwiezdy [Brafman, Beckstorm, 2006], w którym efekty sieciowe i brzegowe w zdecentralizowanych i samoreplikujących się strukturach stają się główną przewagą strategiczną sieci przedsiębiorstw, które naśladują cechy regeneracyjne niektórych gatunków rozgwiezd. Idąc dalej tropem tej analogii, oznacza to, że jeśli chcemy eksplorować nowe terytoria, nieznaną przestrzeń rynkową, nieskażoną konkurencją, takie jak „błękitne oceany”, znane z książki *Blue Ocean Strategy* autorstwa Kima i Mauborgne'a [2005], powinniśmy wziąć pod uwagę percepcję „rodzimych gatunków”. Jeśli ta metafora ma być owocna z perspektywy alternatywnych sposobów percepcji, być może powinniśmy odnieść nasze rozważania do doświadczeń stworzeń głębinowych. Stworzenia te, swoim wyglądem, nie tylko kwestionują nasze wyobrażenia o „estetyce naturalnej”, ale także otwierają umysły na alternatywne formy percepcji i komunikacji w całkowicie obcym nam środowisku. Za inspirujący przykład może posłużyć filozoficzny dyskurs Viléma Flussera *Vampyrotheuthis Infernalis* o percepcji kałamarnicy wampirzej [Flusser, Bec, 2012]. Choć częściowo jest to prowokacja, a częściowo bajka, rozważania przedstawione w *Vampyrotheuthis Infernalis* dają interesu-

jącą perspektywę, gdy chcemy wyobrazić sobie estetyczne wymiary działania zwinnych, syntetycznych agentów, które „pławią się” w sieciach organizacyjnych.

„Świat, który człowiek pojmuje, jest twardy (jak gałęzie, które pierwotnie trzymaliśmy). Musimy go przejść – przemierzyć – aby go uchwycić, ponieważ dziesięć palców naszych chwytnych rąk to kończyny dawnego organu lokomocji. Kałamarnica, przeciwnie, chwyta świat ośmioma mackami, otaczającymi jej usta, które pierwotnie służyły do kierowania strumieni pokarmu do przewodu pokarmowego. Świat chwytny przez kałamarnicę jest płynnym, dośrodkowym wirem. Ujmuje go, aby dostrzec jego płynące cechy szczególne. Podczas, gdy nasza metoda pojmowania jest aktywna – przemierzamy statyczny i ustalony świat – jej metoda jest pasywna i beznamiętna: przyjmuje świat, który pędzi do niej.” [Flusser, Bec, 2012, s. 39].

Organizacyjne API pozwalają właśnie na odwrócenie strategii z poszukiwania partnerów na otwarcie się na ich propozycje. To tzw. strategię URL (ang. *Ubiquity first, Revenue Later*) – wszechobecność najpierw, zysk później, przewrotnie nawiązujące do nazwy linków do stron internetowych (ang. *Uniform Resource Locator*). Strategie te są już wdrażane w sytuacjach, w których bardziej liczy się penetracja rynku i wyjście naprzeciw potencjalnym użytkownikom, niż natychmiastowe generowanie zysków.

Kończąc zaś metaforę oceanu, warto zauważyć, że najnowsza idea płynnych organizacji [Kociatkiewicz, Kostera, 2014] wykazuje silne podobieństwo do świadomej, oceanicznej powierzchni planety Solaris z książki Stanisława Lema pod tym samym tytułem. Co ciekawe, książka ta bada antropomorficzne ograniczenia ludzkiego umysłu i nieadekwatność komunikacji między gatunkami ludzkimi i nie-ludzkimi na długo przed tym, jak realizm spekulatywny i ontologia zorientowana na przedmiot uczyniły to samo.

Przed estetyką spekulatywną organizacji stoi więc zadanie unaocznienia, że syntetyczne, nie-ludzkie byty podejmują samodzielne decyzje na podstawie zupełnie odmiennych niż nasze bodźców i co ważniejsze, że również one mogą ulegać iluzjom. Te przekłamania w sensoryce maszynowej są jednak zupełnie różne od naszych, gdyż bazują na innych procesach poznawczych, takich jak chociażby uczenie maszynowe [Nguyen, Yosinski, Clune, 2015]. Powoduje to, że nie tylko nie w pełni rozumiemy procesy decyzyjne maszyn (oparte na rachunku prawdopodobieństwa, a nie relacjach przyczynowo-skutkowych), ale też często nie potrafimy wyobrazić sobie potencjalnych błędów, wynikających chociażby z nadreprezentacji pewnych bodźców (np. typowego tła czy powtarzającego się kontekstu), które nasz mózg automatycznie filtruje [Shane, 2020]. Warto przy tym zauważyć, że dla autonomicznych robotów, takich jak samojezdne auta, to nasz świat jest podobny filmowemu Matrixowi

(por. youtu.be/a-WuwSwlfIA), a wizualizacje tworzone przez sieci neuronowe mogą przyprawić nas, ludzi, o szczerze zdumienie. Bardzo ciekawym przykładem jest prezentacja samodzielnych wizualizacji stworzonych przez sieci neuronowe firmy Google, odpowiedzialnych za rozpoznawanie obiektów na zdjęciach. Systemy te otrzymały zadanie wizualnego opisanie danego obiektu, czyli odwrócenia procesu rozpoznawania obrazu. Psychodeliczna i oniryczna forma otrzymanych wizualizacji przybliży nam sposób, w jaki niektóre systemy sztuczne postrzegają analizowane obrazy [Mordvintsev, Olah i Tyka, 2015], a przez to umożliwia projekcję sposobów ich funkcjonowania, co może przyczynić się do lepszego zrozumienia nowych form interakcji powstających pomiędzy ludźmi, maszynami i otaczającym je światem.

9.5. Podsumowanie

Pierwsze doświadczenia z pogranicza nowej i spekulatywnej estetyki pozwalają nam lepiej zrozumieć, że doświadczanie systemów społecznych i gospodarczych jest coraz częściej zapośredniczone przez syntetycznych agentów. Co jednak ważniejsze, wszechobecna ewolucja systemów komputerowych, opartych na rozwiązaniach z zakresu sztucznej inteligencji, prowadzi do urzeczywistnienia koncepcji „demokracji bytów”. Najlepszym przykładem na to, jak poważne rodzi to konsekwencje, są pojawiające się doniesienia na temat dylematów moralnych powstających przy projektowaniu autonomicznych samochodów czy wojskowych dronów, które mają automatycznie decydować, kto ma przeżyć w wypadku komunikacyjnym [Stewart, 2016] lub zostać zabity na polu walki [Cohen, Bowes, 2015]. O ile w środowisku organizacyjnym alternatywy decyzyjne rzadko mają taki ciężar moralny, to stosowne rozważania na temat aksjologii organizacji, reprezentowane przez etykę i estetykę organizacyjną, powinny stać się powszechne. Tym bardziej, że co prawda przedwcześnie, ale silnie pobudzające ludzką wyobraźnię informacje o zaistnieniu świadomości w systemach SI, już zaczęły się pojawiać [Vallence, 2022]. Dlatego też powstają już pierwsze projekty legislacyjne z zakresu odpowiedzialności za działania robotów i sztucznej inteligencji wprowadzane przez Unię Europejską [*Roboty i sztuczna inteligencja: Posłowie za odpowiedzialnością prawną w UE*, 2017], a myśliciele na całym świecie przestrzegają przed niekontrolowanym rozwojem sztucznej inteligencji, jak chociażby w liście otwartym *Research Priorities for Robust and Beneficial Artificial Intelligence: an Open Letter* [2016], podpisanym przez Stephena Hawkinga, Elona Muska i wielu innych ekspertów. O ile bowiem autonomiczne fabryki, syste-

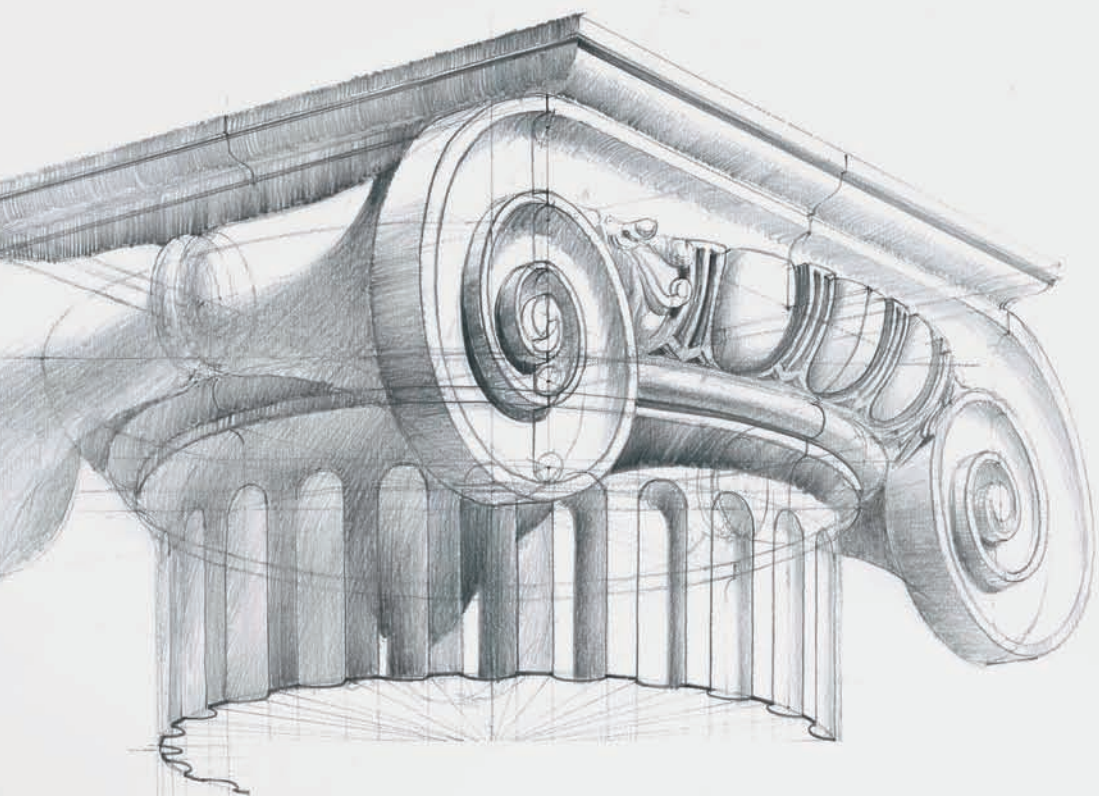
my komputerowe, samochody, drony i inne inteligentne urządzenia już istnieją, to nie jesteśmy jeszcze przygotowani na nie-ludzkie organizacje. Z czystej przeczności nasza badawcza wyobraźnia powinna więc zająć się kwestią syntetycznej podmiotowości w organizacjach społecznych i gospodarczych oraz ewentualnych cech jej autonomicznej strukturalizacji i przyszłości relacji człowiek–maszyna–świat. Tym bardziej, że spekulatywny realizm i ontologia zorientowana na przedmiot to nie tylko „demokracja bytów” i „płaska ontologia”, ale też „mroczny witalizm” i „mroczny materializm”. Dla autorów tych koncepcji obcy świat nie jest wyłącznie światem, gdzie nas nie ma, ale światem aktywnie wrogim wobec nas.

BIBLIOGRAFIA

- Research Priorities for Robust and Beneficial Artificial Intelligence: an Open Letter*, (2016), Future of Life Institute, <https://futureoflife.org/ai-open-letter/> (dostęp: 30.04.2017).
- Roboty i sztuczna inteligencja: Posłowie za odpowiedzialnością prawną w UE*, (2017), Parlament Europejski, Komunikat Prasowy, http://www.europarl.europa.eu/pdfs/news/expert/infopress/20170210IPR61808/20170210IPR61808_pl.pdf (dostęp: 30.04.2017).
- Alsever J., (2016), *Is Software Better at Managing People Than You Are?*, Fortune, March 21, <http://fortune.com/2016/03/21/software-algorithms-hiring/> (dostęp: 08.12.2016).
- Askin R., Ennis P. J., Hägler A., Schweighauser P., (2014), *Speculations V: Aesthetics in the 21st Century*, New York, Punctum Books.
- Bauman Z., (2000), *Liquid Modernity*, Cambridge, Polity Press.
- Betancourt M., (2013), *Automated Labor: The 'New Aesthetic' and Immaterial Physicality*, Ctheory: Theory Beyond the Codes, tbc048, <http://ctheory.net/articles.aspx?id=717> (dostęp: 20.12.2014).
- Blas Z., Rhee J., (2010), *Speculative Aesthetics*, <http://fhi.duke.edu/projects/interdisciplinary-working-groups/speculative-aesthetics> (dostęp: 20.12.2014).
- Bogost I., (2009), *What is Object-Oriented Ontology?*, Ian Bogost, http://bogost.com/writing/blog/what_is_objectoriented_ontology/ (dostęp: 20.12.2014).
- Bogost I., (2012), *Alien phenomenology, or, what it's like to be a thing*, Minneapolis, University of Minnesota Press.
- Brafman O., Beckstrom R. A., (2006), *The starfish and the spider: The unstoppable power of leaderless organizations*, New York, Penguin.
- Bridle J., (2011), *The New Aesthetic*, Really Interesting Group, <http://www.riglondon.com/blog/2011/05/06/the-new-aesthetic/> (dostęp: 20.12.2014).
- Bryant L.R., (2012), *Worries About OOO and Politics*, Larval Subjects, <http://larvalsubjects.wordpress.com/2012/05/29/worries-about-ooo-and-politics/> (dostęp: 20.12.2014).
- Bryant L.R., Srnicek N., Harman G., (2011), *The speculative turn: continental materialism and realism*, Melbourne, Re.press.
- Cohen M., Bowes J., (2015), *Is drone warfare ethical?*, The Stanford Daily, February 16, <http://www.stanforddaily.com/2015/02/16/is-drone-warfare-ethical/> (dostęp: 30.04.2017).
- De Monthoux P.G., (2004), *The Art Firm: Aesthetic Management and Metaphysical Marketing*, Stanford, Stanford University Press.

- Dutton D., (2002), *Aesthetic Universals*, [w:] *The Routledge Companion to Aesthetics*, red. B. Gaut, D. McIver Lopes, London, Routledge, s. 279–292.
- Dzidowski A., (2011a), *Organizacje wirtualne we współczesnej gospodarce*, *Przegląd Organizacji*, 7–8, s. 20–24.
- Dzidowski A., (2011b), *Antropologia wizualna organizacji*, *Problemy Zarządzania*, 2/32, s. 51–62.
- Dzidowski A., (2015), *New and speculative organisational aesthetics*, *Organizational Aesthetics*, 1(4), s. 19–31.
- Dzidowski A., (2017a), *Czy nastąpił koniec historii organizacji?: posthumanizm a ewolucja struktur organizacyjnych*, *Sensus Historiae*, 29(4), s. 213–225.
- Dzidowski A., (2017b), *Posthumanizm a estetyka organizacji*, *Studia Metodologiczne*, 38, s. 149–165.
- Dzidowski A., (2018), *Aesthetic reflection in managerial theory and practice*, *Problemy Zarządzania*, 16/6, s. 39–51.
- Flusser V., Bec L., (2015), *Vampyrotheuthis Infernalis: A Treatise, with a Report by the Institut Scientifique de Recherche Paranaturaliste*, Minneapolis, University of Minnesota Press.
- Griffin R., (1996), *Podstawy zarządzania organizacjami*, Warszawa, PWN.
- Harman G., (2012), *Aesthetics as First Philosophy: Levinas and the Non-Human*, *Naked Punch*, <http://www.nakedpunch.com/articles/147> (dostęp: 20.12.2014).
- Hatch M. J., Cunliffe A. L., (2006), *Organization theory: modern, symbolic and postmodern perspectives*, Oxford, Oxford University Press.
- Hatch M.J., Kostera M., Kozminski A.K., (2005), *The Three Faces of Leadership: Manager, Artist, Priest*, Malden–Oxford–Carlton, Blackwell.
- Hayles N.K., (2008), *How We Became Posthuman: Virtual Bodies in Cybernetics, Literature, and Informatics*, Chicago, University of Chicago Press.
- Hekkert P., (2006), *Design aesthetics: Principles of pleasure in product design*, *Psychology Science*, 48, s. 157–172.
- Iyer B., Subramaniam M., (2015), *Corporate Alliances Matter Less Thanks to APIs*, *Harvard Business Review*, June 08, <https://hbr.org/2015/06/corporate-alliances-matter-less-thanks-to-apis> (dostęp: 08.12.2016).
- Jackson R., (2012), *The Banality of The New Aesthetic*, *Furtherfield*, <http://www.furtherfield.org/features/banality-new-aesthetic> (dostęp: 20.12.2014).
- Kim W.C., Mauborgne R., (2005), *Blue Ocean Strategy: How to Create Uncontested Market Space and Make the Competition Irrelevant*, Boston, Harvard Business School Press.
- Kociatkiewicz J., Kostera M. (red.), (2014), *Liquid Organization: Zygmunt Bauman and Organization Theory*, New York, Routledge.
- Linstead S., Höpfl H.J., (2000), *The Aesthetics of Organization*, London, Sage.
- Mackay R., Pendrell L., Trafford J., (2014), *Speculative Aesthetics*, Falmouth, Urbanomic.
- McAule J., Duberley J., Johnson P., (2007) *Organization theory: Challenges and perspectives*, Harlow, Pearson Education.
- Meillassoux Q., (2008), *After finitude: An essay on the necessity of contingency*, New York, Continuum.
- Mordvintsev A., Olah Ch., Tyka M., (2015), *Inceptionism: Going Deeper into Neural Networks*, *Google Research Blog*, <http://googleresearch.blogspot.co.uk/2015/06/inceptionism-going-deeper-into-neural.html> (dostęp: 04.05.2016).
- Nelson R., (2005), *Aesthetics: Universal or Enculturated?*, *Collegiate Anthropologist*, XXVII/1.
- Nguyen A., Yosinski J., Clune J., (2015), *Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images*, *Computer Vision and Pattern Recognition (CVPR '15)*, IEEE, <http://arxiv.org/abs/1412.1897>.

- Reber R., Schwartz N., Winkielman P., (2004), *Processing Fluency and Aesthetic Pleasure: Is Beauty in the Perceiver's Processing Experience?*, *Personality and Social Psychology Review*, 8/4, s. 364–382.
- Schulz-Schaeffer I., (2001), *Enrolling Software Agents in Human Organisations*, [w:] *Cooperative Agents. Applications in the Social Sciences*, red. N.J. Saam, B. Schmidt, Springer Netherlands, s. 149–163.
- Shane J., (2020), *When data is messy*, AI Weirdness, <https://www.aiweirdness.com/when-data-is-messy-20-07-03/> (dostęp: 22.06.2021).
- Sterling B., (2012), *An Essay on the New Aesthetic*, WIRED, <http://www.wired.com/2012/04/an-essay-on-the-new-aesthetic/> (dostęp: 20.12.2014).
- Stewart I., (2007), *Why beauty is truth: a history of symmetry*, Cambridge, Basic Books.
- Stewart J., (2016), *People want self-driving cars that save lives. Especially theirs*, WIRED, June 23, <https://www.wired.com/2016/06/people-want-self-driving-cars-save-lives-especially/> (dostęp: 30.04.2017).
- Strati A., (1999), *Organization and Aesthetics*, London, Sage.
- Taylor S.S., Hansen H., (2005), *Finding Form: Looking at the Field of Organizational Aesthetics*, *Journal of Management Studies*, 42(6), s. 1211–1231.
- Trafford J., (2014), *Speculative Aesthetics Research Project*, <http://www.ucreative.ac.uk/research-partners-projects-speculative-aesthetics> (dostęp: 20.12.2014).
- Weiner J. (red.), (1994), *Aesthetics Is a Cross-cultural Category*, Manchester, Group for Debates in Anthropological Theory.
- Welsch W., (1996), *Grenzgänge der Ästhetik*, Stuttgart, Reclam.
- Welsch W., (1997), *Aesthetics beyond aesthetics*, *Literature and Aesthetics: The Journal of the Sydney Society of Literature and Aesthetics*, 7, s. 7–23.
- Welsch W., (2000), *Virtual to begin with?*, [w:] Sandbothe M., Marotzki W. (red.), *Subjektivität und Öffentlichkeit – Kulturwissenschaftliche Grundlagenprobleme virtueller Welten*, Herbert von Halem Köln, Verlag, s. 25–60.
- Welsch W., (2005), *Estetyka poza estetyką – o nową postać estetyki*, przeł. K. Guzalska, Kraków, Universitas.
- Vallence Ch., (2022), *Google engineer says Lamda AI system may have its own feelings*, BBC, <https://www.bbc.com/news/technology-61784011> (dostęp: 15.06.2022).



Współczesne badania nad sztuczną inteligencją (SI) mają charakter interdyscyplinarny. Ich jądro stanowi wprawdzie informatyka, dbająca o właściwe oprogramowanie i sprzętową architekturę systemów SI, lecz równie ważne są nauki wnikające w naturę inteligencji ludzkiej, na której ta sztuczna ma się wzorować. Należą do nich: psychologia poznawcza, socjologia, kognitywistyka i różne działy filozofii.

W niniejszej pracy zbiorowej, współtworzonej przez specjalistów reprezentujących różne dyscypliny, podejmujemy kwestię zaufania do inteligencji sztucznej. Pośród wielu czynników, które wpływają na zaufanie człowieka do systemów SI, uwypuklamy dwa: skuteczność systemu połączoną z gwarancjami bezpieczeństwa jego użytkowników oraz poznawczą przejrzystość systemu połączoną ze zdolnością do generowania zrozumiałych dla ludzi wyjaśnień.

ISBN 978-83-8156-511-0



9 788381 565110