Człowiek wobec sztucznej inteligencji

PAN
POLSKA AKADEMIA NAUK

Oddział
w Poznaniu

Człowiek wobec sztucznej inteligencji

Artykuły zostały opracowane na podstawie referatów
wygłoszonych podczas XXVI sesji naukowej z cyklu *Dwugłos Nauki*
w dniu 20 listopada 2024 roku

Pod redakcją
prof. dra hab. Jerzego Kaczorowskiego

Poznań 2025

Polska Akademia Nauk, Oddział w Poznaniu

Człowiek wobec sztucznej inteligencji

Publikacja finansowana przez
Polską Akademię Nauk

Redakcja: Jerzy Kaczorowski

Projekt okładki: Mirosława Korbańska
Korekta: Małgorzata Szkudlarska

Polska Akademia Nauk, Oddział w Poznaniu
61-772 Poznań, Stary Rynek 78/79
tel. (61) 641 50 03
e-mail: poznan@pan.pl, poznan.pan.pl

ISBN: 978-83-68418-21-7

Przygotowanie do druku:
Biuro Upowszechniania Nauki i Wydawnictw PAN

Druk i oprawa:
Agencja Wydawniczo-Poligraficzna GIMPO
ul. Transportowców 11, 02-858 Warszawa

Spis treści

Wprowadzenie 7
prof. dr hab. Jerzy Kaczorowski

Sztuczna inteligencja – jej rozwój, szanse i zagrożenia 9
prof. dr hab. inż. Roman Słowiński

Modelologia, czyli jak wykrywać wady, korygować je, a z czasem nawet uczyć się od modeli syntetycznej inteligencji 31
prof. dr hab. inż. Przemysław Biecek

W cieniu rozkwitających AI 49
dr hab. Agnieszka Lekka-Kowalik, prof. KUL

Czy sztuczna inteligencja może być odpowiedzialna albo godna zaufania? 59
dr Natalia Juchniewicz

Noty o autorach..... 77

Wprowadzenie

Artykuły opublikowane w tej książce opracowane zostały na podstawie referatów wygłoszonych w czasie sesji naukowej z cyklu *Dwugłos Nauki*, zatytułowanej *Człowiek wobec sztucznej inteligencji*. Sesja została zorganizowana przez Polską Akademię Nauk Oddział w Poznaniu i odbyła się w siedzibie Oddziału w dniu 20 listopada 2024 r.

Sesje naukowe z cyklu *Dwugłos Nauki* organizowane są w Poznaniu od 1995 r., a w tym roku mija 30 lat od pierwszego spotkania. Podejmują one problematykę z różnych dziedzin nauki, często w związku z ważnymi i aktualnymi wydarzeniami. Ich celem nie jest pełna analiza tematu, ale jego przedstawienie z kilku punktów widzenia, przez naukowców reprezentujących różne dyscypliny naukowe, w tym roku z dziedziny informatyki, matematyki i etyki.

Tegoroczna sesja dotyczyła niezwykle ważnej, przełomowej problematyki sztucznej inteligencji, która stanowi wielką szansę dla rozwoju człowieka, ale niesie również zagrożenia. Fundamentalne staje się zatem pytanie, czy rozwoju sztucznej inteligencji powinniśmy się obawiać? Od nas zależy, czy potrafimy wyznaczyć granice jej rozwoju – techniczne, prawne i etyczne – i ich przestrzegać, aby nie stała się kolejnym niebezpiecznym wynalazkiem, lecz służyła naszemu dobru.

Początkowo wykorzystywano sztuczną inteligencję jako sposób przetwarzania informacji dla osiągania wyznaczonych celów. Dzisiaj stosuje się nową generatywną sztuczną inteligencję. Umożliwia ona realizację nie tylko konkretnych celów, ale również samodzielność w rozwiązywaniu problemów, które mogą się pojawić dopiero w trakcie realizacji zadań. Techniki uczenia się sztucznej inteligencji są ciągle udoskonalane, zwiększają jej autonomię i upodabniają jej działanie do racjonalnego myślenia człowieka.

Definicja sztucznej inteligencji godnej zaufania przez ludzi, sformułowana przez Komisję Europejską w dokumencie z 2019 r. *Wytyczne w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji*, wymaga spełnienia przez nią następujących zasad: poszanowanie autonomii człowieka, zapobieganie szkodom,

możliwość wyjaśnienia, sprawiedliwość. Sztuczna inteligencja nie powinna działać samodzielnie, sama dla siebie, niezależnie od człowieka, ale jak każda technologia być podporządkowana, ściśle kontrolowana i służyć dobru człowieka.

Zapobieganie szkodom oznacza wykrywanie błędów i ich naprawę. W praktyce często zdarzają się niepoprawne algorytmy, znane są przykłady programów w medycynie, które nie spełniały oczekiwań lekarzy. Pokazuje to, że nie każdy algorytm działa zgodnie z potrzebami, a błędy mogą prowadzić do poważnych konsekwencji. Liczne są przypadki tzw. *incydentów systemowych*, nieprzewidzianych przez twórców. Z tego powodu niedawno została przyjęta *Ustawa o systemach sztucznej inteligencji*, która nakłada na kraje Unii Europejskiej obowiązek monitorowania i raportowania niepożądanych zdarzeń. W rezultacie powstają katalogi nieprawidłowości w celu identyfikacji i analizy błędów.

Technologia może nas przytłoczyć, zdominować, uzależnić, ale bardziej obawiamy się innego hipotetycznego zagrożenia. Wprawdzie sztuczna inteligencja nie ma dotychczas świadomości w sensie ludzkim, lecz postęp techniczny może doprowadzić do wytworzenia się tzw. *sztucznej świadomości*, podobnej do ludzkiej. Stwarza to obawy, żeby sztuczna inteligencja nie stała się niejako odrębnym gatunkiem. Jednak w tej chwili większym zagrożeniem jest nie to, że wymknie się ona spod kontroli człowieka, lecz że realizując skutecznie swoje algorytmy, będzie ograniczała naszą wolność.

Nie mniej ważne jest zagrożenie ontologiczne. Dzięki swoim niemal nieograniczonym możliwościom działania, doskonałości i powszechności zastosowania, sztuczna inteligencja zmieniła nie tylko nasze życie, ale również zajęła centralne miejsce człowieka w świecie. A przecież to człowiek jest istotą unikalną i niepowtarzalną we wszechświecie, przypisane są mu wolność i godność.

Mamy nadzieję, że książka ta będzie dla czytelnika źródłem ciekawej refleksji i inspiracji.

Tekst niniejszej książki jest udostępniony w formie elektronicznej na stronie internetowej Oddziału PAN w Poznaniu: poznan.pan.pl/wydawnictwa

prof. dr hab. Jerzy Kaczorowski, czł. rzecz. PAN
Prezes Oddziału PAN w Poznaniu

Sztuczna inteligencja – jej rozwój, szanse i zagrożenia

prof. dr hab. inż. Roman Słowiński
ORCID: 0000-0002-5200-7795
Instytut Informatyki Politechniki Poznańskiej i Instytut Badań Systemowych PAN

1. Wstęp

Od blisko 80 lat obserwujemy niezwykły postęp w technice obliczeń cyfrowych. Technika komputerowa wkracza do wszystkich dziedzin naszego życia i oferuje coraz to nowe usługi w rozwiązywaniu zadań intelektualnych, które człowiek wykonałby mniej efektywnie. Rozwija się sztuczna inteligencja (SI), która ma ambicje przekroczenia ludzkich zdolności świadomego tworzenia. Stawiamy wobec tego pytania o świadomość inteligentnych maszyn i możliwość ich „uczłowieczenia", a nawet przekroczenia ograniczeń ludzkiej biologii.

W niniejszym wykładzie przedstawimy ewolucję rozumienia sztucznej inteligencji oraz szanse i zagrożenia związane z jej dalszym rozwojem. Prześledzimy trzy następujące po sobie w czasie interpretacje pojęcia inteligencji maszyn i powiązane z nimi efekty. Według nich inteligencja maszyn jest: (i) symulacją procesów myślowych człowieka rozumianych jako obliczanie (arytmetyczne i logiczne), (ii) wykonywaniem zadań intelektualnych człowieka (klasyfikacja, tworzenie reguł, odkrywanie praw i relacji), (iii) racjonalnym działaniem motywowanym planem i wiedzą o otoczeniu.

2. Inteligencja maszyn jako symulacja procesów myślowych człowieka będących obliczaniem

Pogląd, że inteligencja człowieka może być symulowana przez maszyny jako proces obliczania, panował, zanim jeszcze pojawiło się pojęcie SI. Pojawiło się ono po raz pierwszy w 1956 r. podczas *Dartmouth Summer Research Project on Artificial Intelligence* – konferencji zorganizowanej w Dartmouth College w Hanowerze, USA.

W 1950 r. Alan Turing, ojciec nowoczesnej nauki o komputerach, w czasopiśmie „Mind”[1] postawił hipotezę, że myślenie jest obliczaniem i zapytał „czy maszyny mogą myśleć?”. Jeśli wszelkie myślenie byłoby obliczaniem, to można by przypisać procedurze algorytmicznej moc wytwarzania świadomości. To by z kolei oznaczało, że maszyna cyfrowa może przerosnąć inteligencję człowieka po prostu przez odpowiednie zwiększenie efektywności obliczeń.

Powyższe pytanie stawiali też logicy i matematycy przed powstaniem koncepcji współczesnej maszyny cyfrowej. Wielki matematyk XIX w. Georg Cantor uważał, że „nowoczesna matematyka posługując się pojęciem nieskończoności wychodzi poza świat materialny”. Mówił, że „przyjęcie nieskończoności, to jak wiara w Boga”. Człowiek ma zdolność posługiwania się pojęciem nieskończoności, podczas gdy maszyna jej nie ma. Jednakże, jako argument o wyższości umysłu ludzkiego nad maszyną jest dość słaby, gdyż w praktyce wystarczają liczby skończone.

Nieco później nadzieją na matematyczny dowód wyższości umysłu nad maszyną były między innymi prace amerykańskiego logika polskiego pochodzenia Emila Leona Posta i austriackiego logika Kurta Gödla[2]. Post był jednym z pierwszych, którzy zwrócili uwagę na ograniczenia maszyn obliczeniowych i granice ludzkiej wiedzy w kontekście rozwiązywania problemów matematycznych. Gödel natomiast znany jest z twierdzenia o niezupełności systemów formalnych opartych na arytmetyce, które mówi, że „nie istnieje algorytm, który przy danym układzie aksjomatów podałby dowód każdego prawdziwego twierdzenia arytmetyki”. Według brytyjskiego filozofa Johna Randolpha Lucasa[3] przemawiałoby ono za wyższością umysłu nad maszyną, gdyż niealgorytmiczny umysł nie mógłby być adekwatnie symulowany przez komputer. Ludzie, którzy jednak nie do końca rozumieją pojęcie liczb (np. nie potrafimy udowodnić hipotezy Riemanna), nie mogą przekazać tego pojęcia maszynie, czyli nie mogą zadać jej wszystkich możliwych aksjomatów[4]. Niezależnie od tego, że jak pisze Stanisław Krajewski[5], argument Lucasa zawiera sprzeczność z powodu autoreferencji, bardziej wnikliwa analiza twierdzenia Gödla, potwierdzona przez niego samego, pokazuje, że jego twierdzenie nie wyklucza

[1] A.M. Turing, Computing machinery and intelligence, *Mind*, 59, 236 (1950) 433–460.

[2] Ciekawą analizę historyczną tego procesu znajdziesz w wykładzie prof. Giovanniego Sommarugi z ETH Zürich, https://www.academia.edu/42767594/

[3] J.R. Lucas, Minds, Machines and Gödel, *Philosophy*, 36 (1961).

[4] St. Krajewski, *Twierdzenie Gödla i jego filozoficzne interpretacje. Od mechanicyzmu do postmodernizmu*, Wydawnictwo IFiS PAN, Warszawa, 2003.

[5] St. Krajewski, On Gödel's Theorem and Mechanism: Inconsistency or Unsoundness is Unavoidable in any Attempt to 'Out-Gödel' the Mechanist, *Fundamenta Informaticae*, 81, 1–3 (2007) 173–181.

istnienia formalnej teorii, która obejmowałaby całą naszą matematykę, to znaczy, że wyniki Gödla nie wykluczają istnienia robota równoważnego umysłowi w zakresie matematyki. Jednakże, jak zauważa Krajewski: „gdyby taka teoria lub robot istniały, nie bylibyśmy w stanie pojąć ani ich równoważności z nami, ani nawet ich niesprzeczności czy poprawności. Sam wynik Gödla nie rozwiązuje filozoficznych debat; nie dowodzi ani antymechanicyzmu, ani realizmu, choć dostarcza pewnych wskazówek. Praca Gödla podważa natomiast jedną z radykalnych ideologii SI, tzw. «silną SI» Searle'a. Można dowieść, że nie jesteśmy w stanie zaprojektować programu ani robota równoważnego ludzkiemu umysłowi, nawet jeśli taki program teoretycznie mógłby istnieć. Twierdzenie Gödla dopuszcza jednak możliwość, że taki program lub robot mógłby powstać w wyniku działania jakiejś superinteligencji lub w wyniku ewolucji w społeczeństwie robotów. W takim przypadku nie bylibyśmy jednak w stanie odkryć ani potwierdzić tej równoważności"[6].

Pomimo tej nierozstrzygniętej debaty na temat wyższości inteligencji ludzkiej nad maszynową, nie ulegało wątpliwości, że maszyny osiągną pewien poziom inteligencji i aby porównać ją z inteligencją człowieka Alan Turing zaproponował dobrze znaną grę imitacyjną, zwaną testem Turinga. Polega on na tym, że maszyna udzielająca odpowiedzi na pytania operatora powinna sprawiać wrażenie, że odpowiada człowiek. Aby zdać test Turinga, maszyna musi być zdolna do przetwarzania języka naturalnego, rozpoznawania i syntezy mowy, przetwarzania i rozpoznawania obrazów, odkrywania wiedzy z danych (dziś z Internetu) i automatycznego wnioskowania.

Oczywiście, w latach pięćdziesiątych ubiegłego wieku trudno było wyobrazić sobie taką maszynę, ale stopniowo komputery nabywały te umiejętności. Pierwszym spektakularnym zwycięstwem maszyny nad człowiekiem była wygrana komputera „Deep Blue" firmy IBM z arcymistrzem szachowym Gari Kasparowem w 1997 r. W 2014 r. program, który „udawał" 13-letniego Eugene'a Goostmana po kilkuminutowym czacie przekonał 30% sędziów, że jest człowiekiem, czyli zdał w tym stopniu test Turinga. Program ten umiał dość dobrze rozumieć i formułować zdania w języku naturalnym, by posługując się wiedzą z Internetu, odpowiadać „rozsądnie" na pytania. To prawda, że Eugene tłumaczył swoją nie najlepszą angielszczyznę pochodzeniem ukraińskim i wykręcał się od wielu pytań młodzieżowymi odzywkami typu „guzik mnie to obchodzi" albo „mam to gdzieś", ale dla części sędziów robił to jak autentyczny nastolatek. Dwa lata później w Seulu system „AlphaGo" firmy Google pokonał mistrza gry w „Go" Lee Sedola. W grze w szachy komputer już nie mierzy się z człowiekiem, lecz z innymi programami. W 2017 r. firma DeepMind pokazała system „AlphaZero" do gry w szachy, oparty

[6] St. Krajewski, The ultimate strengthening of Turing's Test? *Semiotica*, February 18, 2012.

na sztucznej sieci neuronowej, który pobił „Stockfish", najlepszy dotąd system ogrywający arcymistrzów. W tym samym roku do mediów trafiła wiadomość, że król Arabii Saudyjskiej przyznał obywatelstwo swojego kraju robotowi „Sophia" firmy Hanston Robotics z Hongkongu. Niedługo potem, w 2018 r., IBM Watson Debater wygrał debatę oksfordzką, a wcześniej turniej Va-Banque. Dzisiaj duże modele językowe, jak ChatGPT, znacznie lepiej od Goostmana czy Sophii odpowiadają na pytania i mają już miliardy użytkowników.

Dochodzimy do wniosku, że test Turinga jest przestarzały, gdyż testowanie wiedzy i logicznego wnioskowania nie rozstrzyga o równości inteligencji ludzkiej i maszynowej. Ciekawą analizę poglądów na temat testu Turinga przeprowadzili Józef Bremer i Mariusz Flasiński[7]. Porównali poglądy czterech wpływowych reprezentantów filozofii analitycznej: Ludwiga Wittgensteina, Noama Chomsky'ego, Hilarego Putnama i Johna Searle'a. Na podstawie różnych argumentów dotyczących kwestii filozoficznych i/lub językowych dochodzą oni do odmiennych ocen zarówno znaczenia, jak i przydatności testu Turinga dla rozwoju nauk kognitywnych i modeli SI. Niemniej jednak łączy ich odrzucenie idei, że test Turinga można traktować jako test „maszynowego myślenia". Wydaje się, że wynika to z troski o właściwe użycie języka – kwestii będącej fundamentalnym zobowiązaniem metodologicznym filozofii analitycznej.

Wspomniany już Stanisław Krajewski[8] zaproponował z kolei wzmocnienie testu Turinga, czyli eksperyment myślowy, w którym dziecko jest od urodzenia wychowywane przez roboty i ma się okazać, czy tak ukształtowany człowiek będzie w miarę „normalny". Autor dodaje od razu, że taki test byłby oczywiście nieetyczny.

3. Inteligencja maszyn jako zdolność do wykonywania intelektualnych zadań człowieka

Początkowa euforia związana ze sztuczną inteligencją, która sprawi, że wkrótce maszyna zastąpi człowieka w myśleniu, ustąpiła w latach sześćdziesiątych realizmowi spójnemu ze skromniejszą definicją sztucznej inteligencji podaną przez Norberta Wienera: „inteligencja jest procesem pozyskiwania i przetwarzania informacji dla osiągania wyznaczonych celów". Tak rozumiana sztuczna inteligencja odnosi spektakularne sukcesy w przetwarzaniu i rozpoznawaniu obrazów, rozpoznawaniu

[7] J. Bremer, M. Flasiński, The Turing Test, or a Misuse of Language when Ascribing Mental Qualities to Machines. *Forum Philosophicum*, 27, 1 (2022) 6–25.

[8] St. Krajewski, The ultimate strengthening of Turing's Test? *Semiotica*, February 18, 2012.

i syntezie mowy oraz analizie tekstów i przetwarzaniu języka naturalnego. Inteligentne wykonywanie zadań odbywa się z użyciem heurystyk, które pozwalają ograniczyć zakres przeszukiwania pola problemowego, a tym samym skrócić czas rozwiązywania problemu. Ponadto temu procesowi towarzyszy uczenie się. Jest to rozwój pozytywny, sprzyjający symbiozie pomiędzy ludźmi i maszynami w systemach inteligentnego wspomagania decyzji[9].

Chociaż postępy w wykonywaniu zadań intelektualnych człowieka przez maszyny nie zawsze są tak spektakularne, jak pokonanie szachowego arcymistrza, trzeba przyznać, że na polu eksploracji dużych wolumenów danych (ang. *big data*) człowiek został pokonany. Od początku XXI w. rozwija się analityka danych (ang. *data-driven analytics*) polegająca na komputerowym przekształcaniu surowych danych pochodzących z różnych źródeł w wiedzę wykorzystywaną do podejmowania lepszych decyzji. Do obiegu weszło także określenie nauka o danych (ang. *data science*).

Na zasadzie indukcji odkrywa się wzorce prawdziwe w świecie analizowanych danych, które reprezentują wiedzę zawartą w tych danych. Jest to paradygmat tzw. uczenia maszynowego (ang. *machine learning*). W 1983 r. Herbert Simon zdefiniował uczenie maszynowe jako „adaptacyjne zmiany w systemie, które pozwalają systemowi wykonać za następnym razem takie samo zadanie lub zadania podobne bardziej efektywnie". Odkrywanie wiedzy z danych odbywa się w sposób nieukierunkowany: np. „powiedz mi coś interesującego o moich danych"; ukierunkowany: np. „scharakteryzuj klientów, którzy odpowiedzieli na ofertę promocyjną"; przez testowanie lub uściślanie hipotez: np. „czy to prawda, że jest związek między sposobem odżywiania a zapadalnością na chorobę X?"

Warto dodać, że w coraz większym stopniu analitycy danych radzą sobie z danymi niepełnymi, niedokładnymi, podlegającymi przypadkowym fluktuacjom i częściowo sprzecznymi – te niedoskonałości danych są problemem we wnioskowaniu automatycznym pomimo wielkiej masy danych.

W przypadku danych niepełnych i sprzecznych na uwagę zasługuje matematyczna teoria polskiego informatyka Zdzisława Pawlaka oparta na pojęciu zbioru przybliżonego (ang. *rough set*)[10]. Warto zauważyć, że John Hopfield otrzymał w 2024 r. Nagrodę Nobla za opracowanie neuronowej sieci asocjacyjnej i prace, które przyczyniły się do rozwoju metod pozwalających na wnioskowanie z niepełnych danych właśnie.

---

[9] R. Słowiński (ed.): *Intelligent Decision Support – Handbook of Applications and Advances of Rough Sets Theory*. Kluwer Academic Publishers, Dordrecht, 1992.

[10] Z. Pawlak, *Rough Sets – Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Dordrecht, 1991.

W tym kontekście nie mogę się także powstrzymać od wspomnienia o Dominacyjnej Teorii Zbiorów Przybliżonych (ang. *Dominance-based Rough Set Approach*), którą zaproponowałem z kolegami z Katanii[11]. To podejście dokonało swoistego przełomu w rozumowaniu na temat danych porządkowych, typowych dla wieloatrybutowych problemów decyzyjnych. Radzi sobie ono z danymi niepełnymi i częściowo sprzecznymi, i zapewnia reprezentację preferencji użytkownika w postaci zrozumiałych reguł decyzyjnych o składni „*jeśli...*, *to...*". Na przykład, we wspomaganiu decyzji w warunkach ryzyka i niepewności wspomaganie decyzji dokonuje się za pomocą reguł typu „*jeśli prawdopodobieństwo zysku co najmniej 200$ więcej jest* ≥ 0.75, *a zysk co najmniej 100$ więcej jest pewny*, *to akt inwestycyjny A jest lepszy od aktu B*", zamiast za pomocą argumentu teorii użyteczności, który stwierdza, że „*akt A jest lepszy od aktu B*, *gdyż spodziewana użyteczność* $U(A) = 0.67 > U(B) = 0.63$", nie wyjaśniając, co kryje się pod tymi liczbami. Podejście regułowe do wspomagania decyzji jest zgodne z wyjaśniającymi ambicjami sztucznej inteligencji (ang. *explainable AI* lub *XAI*).

Wobec rosnącej efektywności maszyn cyfrowych w odkrywaniu wiedzy z dużych wolumenów danych pojawia się często pytanie, jaką prawdę odkrywa maszyna, czy jest to ta sama prawda, jakiej szuka człowiek? Otóż maszyna pozna prawdę zawartą w tych danych, z których została wyindukowana. Mogą jednak pojawić się nowe dane, które starą prawdę zafałszują. To pokazuje ograniczoność pojęcia prawdy odkrytej przez indukcję charakterystyczną dla uczenia maszynowego. Zbiór danych, z których uczy się maszyna, nie jest ani wyczerpujący, ani stały. Prawda, którą kontempluje człowiek, została dobrze scharakteryzowana przez św. Jana Pawła II w encyklice *Fides et Ratio*[12]. Do zrozumienia tej prawdy potrzeba nie tylko racjonalności, ale także wiary w tajemnicę stworzenia. Jak pisze Eric-Emmanuel Schmitt: „tajemnica nie tkwi w tym, co nieznane, lecz w tym, co niezrozumiałe"[13]. Gdyby cała prawda o Bogu i człowieku była osiągalna tylko „szkiełkiem i okiem", to nie bylibyśmy osobami wolnymi i nie moglibyśmy wybierać – czy wierzę, czy nie wierzę. Akceptuję tajemnicę, niepewność istnienia Boga, sercem, ale wiarę przyjmuję racjonalnie z własnej woli[14]. Tak dochodzę do prawdy, której nie mogą zafałszować nowe dane.

---

[11] S. Greco, B. Matarazzo, R. Słowiński, Rough sets theory for multicriteria decision analysis. *European J. of Operational Research*, 129 (2001) 1–47.

[12] https://www.vatican.va/content/john-paul-ii/pl/encyclicals/documents/hf_jp-ii_enc_14091998_fides-et-ratio.html

[13] E.-E. Schmitt, *Sen o Jerozolimie*. Znak Literanova, Kraków, 2024.

[14] Rozmowa z prof. R. Słowińskim, Człowiek – robot, robot – człowiek? *W drodze*, 3 (2020) 66–77.

Omawiając rozwój zdolności maszyn cyfrowych do wykonywania zadań intelektualnych człowieka, trudno jest pominąć rozwój sieci Internet, która jest ogromną bazą danych. Dane te są przedmiotem analiz algorytmów uczenia maszynowego. Dzisiejszy Internet Rzeczy (ang. *Internet of Things – IoT*) łączy ponad 75 miliardów urządzeń identyfikowanych przez adresy IP. Są to nie tylko komputery, ale liczne urządzenia typu telefony, czujniki, pojazdy, kamery, a nawet lodówki. Internet sprawił, że świat stał się „globalną wioską", w której wszyscy mogą się komunikować ze wszystkimi, oczywiście pod warunkiem posiadania dostępu do sieci (brak dostępu nazywa się dzisiaj „cyfrowym wykluczeniem").

Tak wielki przyrost liczby urządzeń podłączonych do sieci i generujących dane, a także postęp algorytmów uczenia maszynowego nie jest oczywiście obojętny ani dla życia społecznego, ani dla nauki i jej przestrzennej organizacji. Z punktu widzenia nauki sieć komputerowa i algorytmy sztucznej inteligencji dokonały rewolucji nie tylko w komunikacji między uczonymi i w wyszukiwaniu informacji, ale także w podejściu do odkryć naukowych, tworząc wirtualne laboratoria i infrastrukturę dla usług obliczeniowych „na żądanie" w trybie przetwarzania sieciowego (ang. *grid*) lub w chmurze (ang. *cloud*).

Zamykając ten fragment wykładu, warto przytoczyć jeszcze jeden konkretny przykład kompleksowego działania sztucznej inteligencji dla wspomagania decyzji. Chodzi o wspomaganie osób niewidomych. Chieko Asakawa jest niewidomą Japonką-informatykiem, pracującą w IBM Research – Tokyo. Miałem okazję wysłuchać jej prezentacji w listopadzie 2019 r. podczas Światowego Forum Nauki w Budapeszcie. W aplikacji, którą przedstawia film[15], widzimy 6 typów zastosowań SI: orientacja przestrzenna i nawigacja – rozpoznawanie twarzy – rozpoznawanie emocji – rozpoznawanie kontekstu zdarzeń – rozpoznawanie tekstu – synteza mowy.

4. Generatywna sztuczna inteligencja

Nowym etapem rozwoju sztucznej inteligencji jest generatywna sztuczna inteligencja oparta na uogólnionych modelach zdolnych do wykonywania różnych zadań intelektualnych człowieka. Jest ona nadzieją dla nowoczesnych systemów zarządzania i wytwarzania, tzw. Przemysłu 4.0 (ang. *Industry 4.0 + AI = AI4Industry*). Charakteryzują się one strukturą modularną połączoną w jeden system zwany łańcuchem bloków (ang. *blockchain*). Łańcuch bloków to zdecentralizowana, rozproszona baza danych (transakcyjnych) bez centralnych komputerów, do której dostęp może uzyskać każdy uczestnik procesu. W bazie tej dokonywane jest gromadzenie i kompleksowa

[15] https://youtu.be/f-mQIWnO3Ag? si=EWxzCumV5lyRj59U

analiza danych z wielu źródeł. Taka struktura zastępuje klasyczną strukturę hierarchiczną z centralnym zarządzaniem i deterministycznym nadawaniem reguł komunikacji między składowymi systemu. Decentralizacja i autonomia podsystemów mają wzmóc odporność systemu na zmianę kontekstu i warunków operacyjnych. Powiązania między podsystemami tworzą się tak jak w sieci społecznościowej i rolą algorytmów sztucznej inteligencji jest odkryć je dla zidentyfikowania skupień wynikających z interakcji niekoniecznie z góry zadekretowanych (klastrów). Jest to typowe dla podejścia behawioralnego charakterystycznego dla sztucznej inteligencji: „pokaż mi przykłady działania, a algorytmy sztucznej inteligencji stworzą jego model".

Moduł systemu zarządzania lub wytwarzania może być traktowany jak autonomiczny agent zdobywający wiedzę o (wirtualnym) otoczeniu dzięki uczeniu się z danych. Agentem może być robot „ucieleśniający" sztuczną inteligencję lub program komputerowy działający według algorytmu uczącego. W tradycyjnym uczeniu maszynowym algorytmy uczenia się są typowo nastawione na pojedynczy dobrze zdefiniowany problem i nie dostosowują się do zmiennych okoliczności. Jest to uczenie się z funkcją nagrody i postępowanie w trybie 'sytuacja-akcja' według wyuczonych reguł maksymalizujących przyjętą funkcję użyteczności (ang. *extrinsic motivation*).

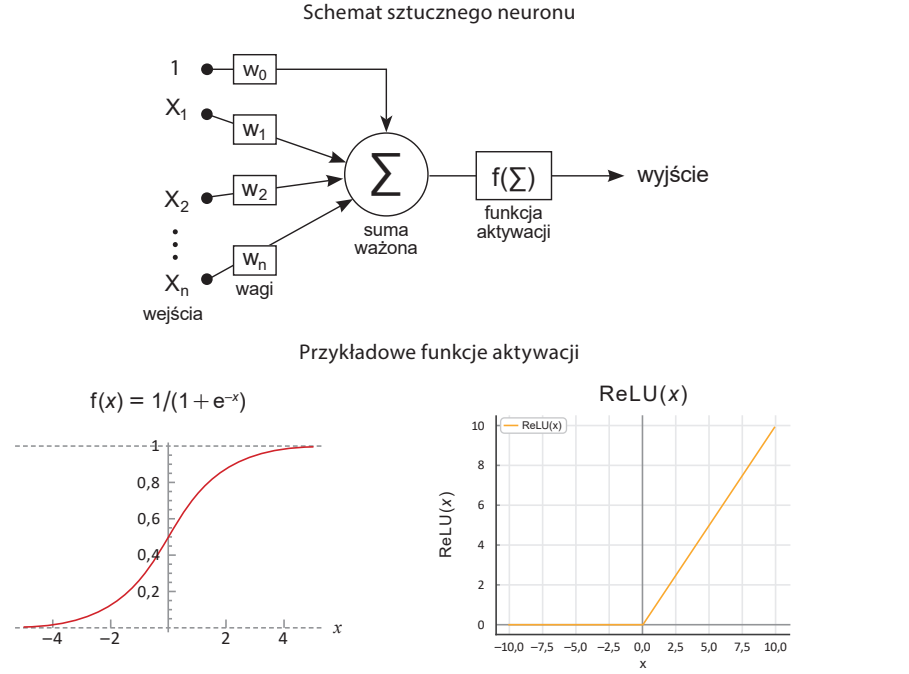
Nowy paradygmat uczenia maszynowego, który bardziej przystaje do systemu modularnego z niezadekretowanymi interakcjami między podsystemami, nazywa się uczeniem ze wzmocnieniem (ang. *reinforcement learning*). Uczenie ze wzmocnieniem odbywa się bez funkcji nagrody, a eksploracja otoczenia dokonuje się przez akcję z wyciąganiem wniosków z tego, co się stało (ang. *intrinsic motivation*). Jest to analog do programowania dynamicznego, gdzie przyszłe stany są liczne i nie wiadomo *a priori*, który jest najlepszy. Agent uczy się nie tyle strategii właściwego postępowania, ile sposobu oceny swojego działania i własnych preferencji. Ponieważ agent uczy się preferencji w kontekście wielu konfliktowych celów, wykorzystuje podejście wielokryterialne[16]. Taka aktywność sprzyja nabywaniu szerokich kompetencji zamiast specjalizacji w osiąganiu narzuconego celu. Te kompetencje tworzą elementy wiedzy, z których agent może zbudować rozwiązanie nowego problemu, jaki może się pojawić.

Inteligentny agent uczący się ze wzmocnieniem ma zatem przewagę nad agentem uczącym się według algorytmu klasycznego, gdyż jego wiedza pozwala na działanie dostosowane do wcześniej nieznanych scenariuszy, co jest bliższe racjonalnemu działaniu człowieka.

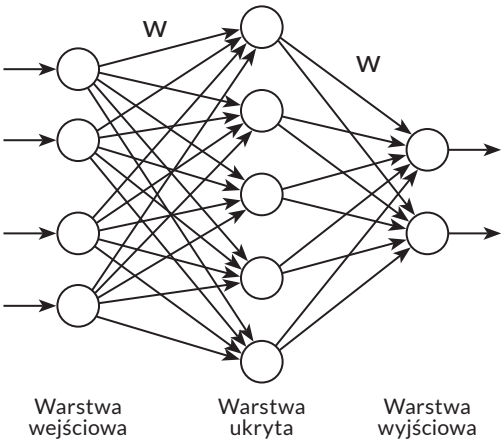
[16] E. Hüllermeier, R. Słowiński, Preference learning and multiple criteria decision aiding: differences, commonalities, and synergies. *4OR – A Quarterly Journal of Operations Research*, 22 (2024) part I: 179–209, part II: 313–349.

Techniki uczenia maszynowego są ciągle udoskonalane, w szczególności uczenie złożonych funkcji matematycznych zwanych Sztucznymi Sieciami Neuronowymi (SSN) (ang. *Artificial Neural Network – ANN*). Tak zwane uczenie głębokie (ang. *deep learning*) jest kamieniem milowym nowoczesnej SI. Techniki głębokiego uczenia są bardzo efektywne w nauce wielowarstwowych sieci neuronowych do przetwarzania i rozumienia złożonych wzorców oraz relacji w danych obrazowych i językowych. Trenowanie modeli uczenia głębokiego, polegające na konfigurowaniu parametrów neuronów sieci, wymaga intensywnych obliczeń, które można łatwo zrównoleglić. Z tego powodu stosuje się procesory graficzne (GPU), które dzięki swojej równoległej strukturze są niezwykle efektywne w przyspieszaniu uczenia głębokiego. Duże Modele Językowe (DMJ) (ang. *Large Language Models – LLM*), takie jak ChatGPT-4, Claude, PaLM, Liama, LaMDA, Gemini, Cohere, Orca itd., opierają się na technikach uczenia głębokiego. Tworzą one klasę tzw. generatywnych modeli SI.

Przypomnijmy, że sztuczny neuron to wieloparametryczna funkcja matematyczna, przedstawiona na rys. 1. Sieci tworzą neurony podobnego typu, tak jak



Rys. 1. Sztuczny neuron – wagi w są parametrami połączeń (synaps) między neuronami



Rys. 2. Perceptron wielowarstwowy

to przedstawiono na rys. 2. Perceptron wielowarstwowy może być wykorzystywany do klasyfikowania zbiorów obiektów, które nie są liniowo separowalne. Jego pomysłodawcą jest Geoffrey Hinton, drugi obok Johna Hopfielda laureat Nagrody Nobla z fizyki w 2024 r. Hopfield wprowadził tzw. asocjacyjne sieci neuronowe dające możliwość rekonstrukcji i rozpoznawania wcześniej zapamiętanych wzorców na podstawie skojarzeń, które bazują na dostępnym fragmencie wzorca. Idea trenowania, czyli uczenia SSN, polega na wyznaczaniu wartości wag (parametrów SSN), które gwarantują poprawne odtwarzanie relacji wejście-wyjście zachodzących w przykładach uczących. Do wyznaczania tych parametrów służą algorytmy optymalizacji rozwijane od dziesięcioleci w badaniach operacyjnych (BO), co podkreśla ścisły związek BO i SI.

Duże modele językowe są trenowane na ogromnych i zróżnicowanych zbiorach danych tekstowych, co pozwala im generować spójny i kontekstowo odpowiedni tekst na podstawie otrzymanych danych wejściowych. Potrafią też śledzić kontekst rozmowy.

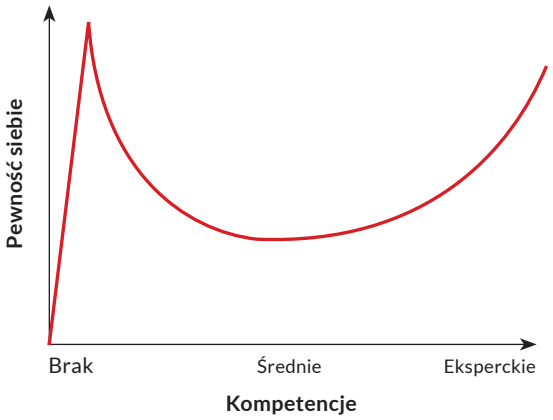
Najsłynniejszy dziś DMJ to ChatGPT firmy OpenAI. Wyjaśniając jego nazwę, *Generative Pre-trained Transformer*, to sieć neuronowa zaprojektowana do nauki odległych zależności statystycznych w tekstach. Trenowanie modelu językowego obejmuje techniki przewidywania następnego tokenu (ang. *Next-Token-Prediction – NTP*) oraz maskowania języka (ang. *Masked-Language-Modeling – MLM*). W technice MLM pewne słowa w sekwencji są „maskowane", a model jest trenowany, aby przewidzieć poprawne słowa na podstawie kontekstu otaczającego

tekstu. Na przykład, dla wyrażenia „Kot siedział na...", model szacuje prawdopodobieństwo kolejnego słowa, takiego jak „kanapie", „podłodze" lub „dachu", biorąc pod uwagę częstotliwość występowania tych słów w innych tekstach.

Modele językowe nie korzystają z żadnego logicznego modelu wnioskowania o świecie. Potrafią zasugerować, jak odpowiedź powinna brzmieć, co różni się od tego, jaka odpowiedź powinna być.

Nawet twórcy ChatGPT z OpenAI przyznają, że ludzie mylą wydajność z kompetencją. DMJ nie są inteligentne w taki sposób, jak nasz mózg. Nie uczą się informacji w uporządkowany sposób. W przeciwieństwie do ludzi, przynajmniej na razie, nie mogą wchodzić w interakcje ze światem, aby uczyć się przyczynowości, kontekstu czy analogii. Co więcej, mówi się, że „halucynują", czyli tworzą stwierdzenia, które na pierwszy rzut oka wydają się wiarygodne, ale są bezpodstawne lub zafałszowane błędnymi danymi spotykanymi w sieci.

Do DJM można metaforycznie zastosować tzw. efekt Dunninga-Krugera, znany z psychologii, który przedstawia rys. 3. Tak jak ludzie o niskich kompetencjach, DMJ generują bardzo pewne odpowiedzi, nawet jeśli brakuje im wystarczających informacji lub gdy dane są niejasne bądź nieprecyzyjne. Ta „pewność siebie" wynika ze sposobu, w jaki trenowane są DMJ do tworzenia odpowiedzi w sposób płynny i autorytatywny. Podobnie jak u niekompetentnych ludzi, DMJ nie wiedzą, czego nie wiedzą. Efekt Dunninga-Krugera może odnosić się także do rozumienia DMJ przez użytkowników. Osoby nieświadome ograniczeń SI mogą zakładać, że DMJ „rozumieją" informacje tak jak ludzie i są wiarygodnym źródłem wiedzy. Takie przecenianie sprawia, że użytkownicy zbytnio polegają na wynikach genero-



Rys. 3. Efekt Dunninga-Krugera

wanych przez DMJ, nie poddając ich krytycznej ocenie. Podobnie jak ludzie zdobywający doświadczenie, DMJ mogą ulepszać swoje działanie dzięki dodatkowym danym i dostrajaniu. Wtedy ich wiarygodność wzrasta.

DMJ są trenowane nie tylko na tekstach, ale także obrazach, filmach, kodach komputerowych i dźwiękach. Stąd pojawiło się ogólniejsze pojęcie: Duże Modele Multimodalne (DMM) (ang. *Large Multimodal Models – LMM*). Dostępne są liczne modele specjalizowane, wywodzące się z DMJ, zwane wtyczkami (ang. *plug-ins*). Wtyczki te dokonują np. podsumowań tekstów z zadaną liczbą znaków, wyjaśniają i dokumentują kod programisty, tworzą obraz lub wideo na słowne polecenie, imitują mowę, tworzą muzykę, piosenki i wiersze oraz rozwiązują testy z fizyki i matematyki. ChatGPT-4o1 miał 83% sukcesu przy rozwiązywaniu zadań z olimpiady matematycznej. Dziś ocenia się jego poziom kompetencji na porównywalny z poziomem doktoranta.

Powstał też polski DJM Bielik, ale jest niedoinwestowany, a także polski serwis porad prawnych Legitio. Wtyczka Scholar AI do ChatGPT-4o podaje informacje o publikacjach naukowych, ich autorach oraz o cytowaniach.

DMJ lepiej niż ludzie rozpoznają ze zdjęć osób ich preferencje seksualne i polityczne. System SORA od OpenAI tworzy bardzo realistyczne klipy wideo na polecenia słowne, np. utwórz klip „spacer po Tokio". DMM są już instalowane na smartfonach, jak Apple Intelligence, Galaxy AI, do generowania obrazów, modyfikacji zdjęć, rozpoznawania twarzy, tłumaczenia i podsumowywania notatek lub nagrań mowy.

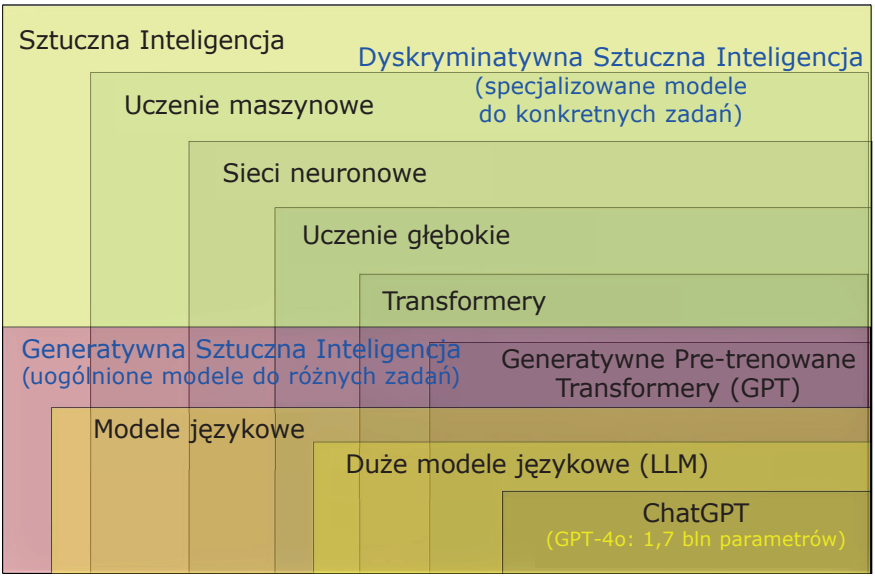
Generatywna SI ma też ujemne strony. Deepfake Porn[17], na podstawie jednego wyraźnego zdjęcia twarzy, jest w stanie w ciągu mniej niż 25 min stworzyć za darmo 60-sekundowe pornograficzne wideo.

Liderami w rozwoju DMJ i DMM są OpenAI, Google i Meta. Sporo ich produktów jest otwartych, co ułatwia deweloperom tworzenie niezliczonej liczby modyfikacji i specjalizacji. Pojawiła się także nowa klasa modeli SI – DMM wieloagentowe, w których wiele modeli współpracuje dla rozwiązania złożonych zadań.

Schemat podany na rys. 4 podsumowuje użyte dotąd pojęcia związane ze SI i przedstawia ich hierarchię (rozmiar okien nieistotny).

Warto zwrócić uwagę, jak duże są koszty trenowania i używania DJM. Modele uczenia głębokiego to „brutalne" techniki statystyczne, które uczą się na ogromnych ilościach danych (rzędu 10 TB tekstu). Sieci neuronowe DMJ zawierają setki warstw i biliony (10^{12}) wag, co sprawia, że obliczenia są niezwykle energochłonne.

[17] E. Strickland, Deepfake Porn is Leading to a New Protection Industry, *IEEE Spectrum*, 15 July 2024.



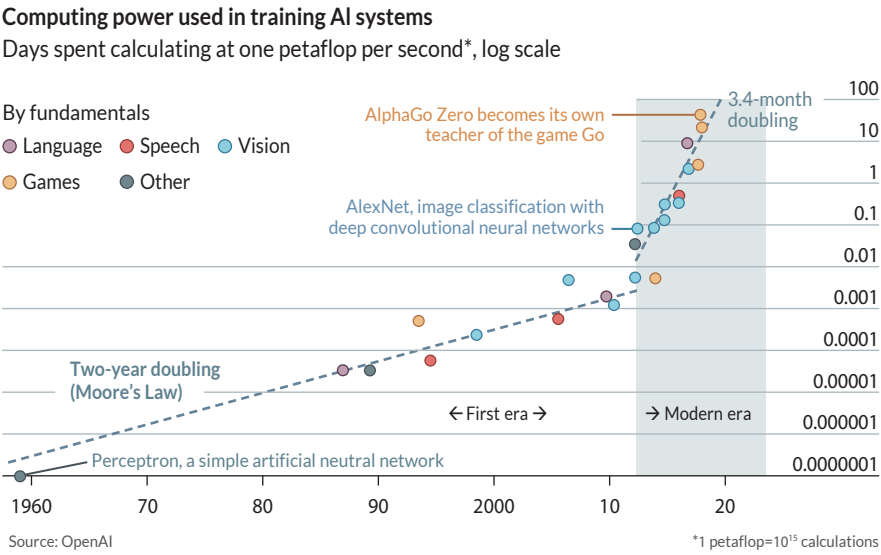
Rys. 4. Podsumowanie i hierarchia pojęć związanych ze sztuczną inteligencją

ChatGPT ma 1,7 bln parametrów uczenia maszynowego. Trenowanie modelu GPT-4 wymaga użycia 25 tys. jednostek GPU przez 90–100 dni, przy kosztach wynoszących 100 mln dolarów, a zużycie energii wynosi od 51 773 do 62 319 MWh (energia zużywana przez 1000 gospodarstw domowych w USA przez 5–6 lat). Zużycie energii przy użytkowaniu jest 10 razy wyższe niż koszty treningu (w ciągu pierwszego miesiąca 2024 r. ChatGPT zużył tyle prądu, co 26 tys. domostw)[18].

Wysokie koszty energii sprawiają, że DMJ i DMM są rozwijane przez bogate firmy, a nie przez laboratoria uniwersyteckie. Na przykład właściciel Facebooka, firma META, inwestuje około 90 mld dolarów rocznie, w tym 10 mld dolarów w rozwój Metaverse, w nadziei, że wirtualna i rozszerzona rzeczywistość (ang. *Virtual Reality – VR*, *Augmented Reality – AR*) staną się kolejną dużą platformą obliczeniową. Metaverse ma być następnym krokiem milowym na drodze rozwoju Internetu, tworząc wirtualny lub hybrydowy świat trójwymiarowy wygenerowany cyfrowo, w którym społeczność będzie mogła odnajdywać się dla pracy lub rozrywki za pomocą gogli VR, kontrolerów ruchu i dźwięku przestrzennego.

[18] arXiv:1906.02243.

Od 2012 r. moc obliczeniowa potrzebna do trenowania systemów SI podwaja się co 3–4 miesiące (rys. 5).



Rys. 5. Moc obliczeniowa wykorzystywana do trenowania systemów SI[19]

Z powyższych powodów powstaje nowa generacja komputerów na potrzeby trenowania DMJ: to komputery neuromorficzne, które w przeciwieństwie do konwencjonalnych komputerów, używających oddzielnych procesorów do obliczeń i pamięci do przechowywania danych, integrują te dwie funkcje w sztucznych neuronach. Dzięki temu nie ma potrzeby kosztownej synchronizacji. Ten sposób przetwarzania informacji jest bliższy naszym mózgom. Na przykład, komputer Hala Point firmy Intel przetwarza informacje 50 razy szybciej, zużywając 100 razy mniej energii niż zwykłe komputery. Mówi się o 15 bln operacji na sekundę i na wat. Hala Point jest wyposażony w 1152 nowe procesory Intela, a komputer może obsłużyć 1,15 mld sztucznych neuronów i 128 mld synaps, co odpowiadałoby mózgowi małego ssaka. Mózg ludzki ma około 100 mld neuronów i 100 bln połączeń-synaps. Nasz mózg ma ponadto wyspecjalizowane regiony i kontrolę emocji.

[19] https://openai.com/research/ai-and-compute

5. Inteligencja maszyn jako racjonalne działanie motywowane planem i wiedzą o otoczeniu

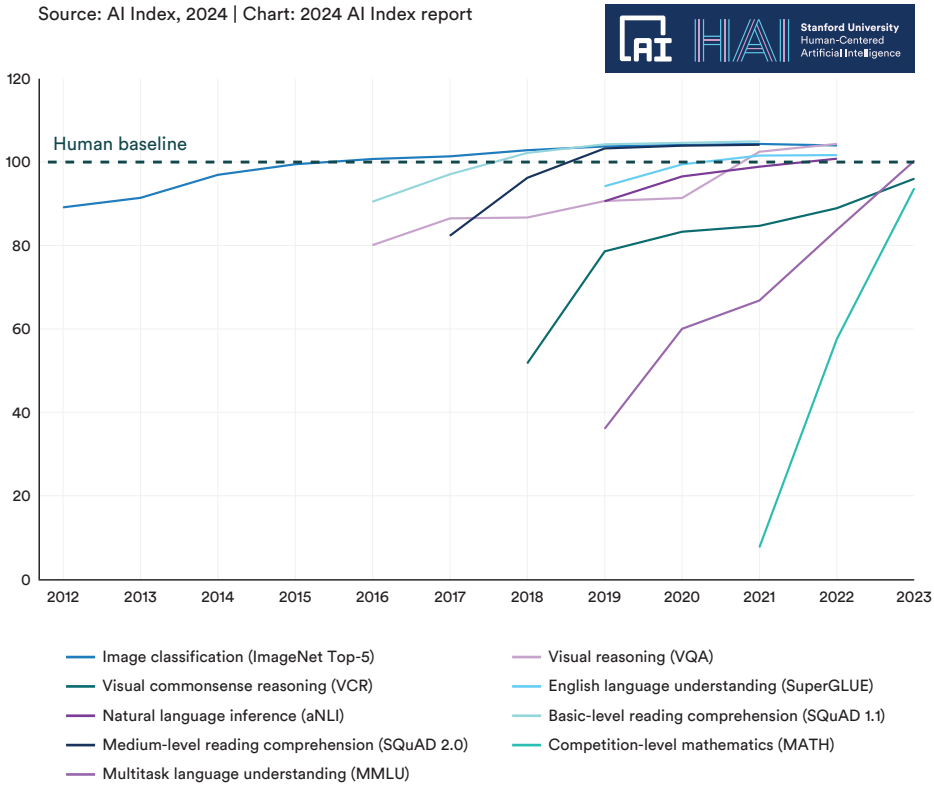
Przechodząc do trzeciej definicji inteligencji maszynowej, która ma ambicje dorównać ludzkiemu rozsądkowi przez autonomiczne zdobywanie wiedzy o otoczeniu i działanie planowe, musimy zadać pytania o świadomość maszyn i możliwość ich „uczłowieczenia" przez nabycie cech i uczuć wyższych, którymi człowiek odróżnia się na przykład od zwierząt. Jakie konsekwencje dla człowieczeństwa niesie rozwój SI?

Najprostsze (choć bezwzględnie bardzo trudne) pytanie tego typu to, czy maszyna może osiągnąć stan samoświadomości, która pozwoli jej zdać sobie sprawę z relacji „ja-środowisko" i powiedzieć: „ja jestem". Stanisław Krajewski[20] zauważa, że uświadomione „ja" w relacji do przedmiotów jest łatwiejsze do uświadomienia przez maszynę niż „ja" w relacji do innych ludzi, czyli „ja-ty". Te relacje są o wiele trudniejsze do zasymulowania niż relacje robota ze środowiskiem, które nie ma cech ludzkich. Relacje współpracy, życia razem, relacje uczuciowe, odczuwanie wstydu, nadziei czy wdzięczności są niezwykle trudne do zasymulowania. Ponadto wdzięczność czy nadzieja są pojęciami przekraczającymi rzeczywistość „tu i teraz", a zatem posiadają element transcendencji charakterystyczny dla religii i modlitwy. Rodzą one szczególny rodzaj relacji „ja-Bóg". Można wątpić, czy maszyny za sprawą programów komputerowych wyewoluują do takiego poziomu świadomości.

Niemniej jednak robot wykorzystujący generatywną sztuczną inteligencję potrafi malować, komponować muzykę czy głosić kazania, jak „Mindar" do buddystów. SI dorównuje już, a nawet przewyższa niektóre zdolności rozumowe ludzkiego mózgu. Raport „2024 AI Index Report" przedstawiony przez Human-Centered Artificial Intelligence Unit Uniwersytetu Stanforda (rys. 6) pokazuje, że SI przewyższyła ludzkie możliwości w kilku benchmarkach, w tym w klasyfikacji obrazów, rozumowaniu wizualnym i rozumieniu języka angielskiego. Jednak wciąż pozostaje w tyle w przypadku bardziej złożonych zadań, takich jak matematyka na poziomie konkursowym, wizualne rozumowanie zdroworozsądkowe i planowanie.

Moim zdaniem możliwe jest, że maszyny rozwiną swoją własną inteligencję, która pozwoli im zachowywać się w taki sposób, że ludzie subiektywnie oceniający ich świat będą mogli powiedzieć „one zachowują się, jakby były świadome/

[20] St. Krajewski, Can a robot be grateful? Beyond Logic, towards religion, *Eidos – Journal for Philosophy of Culture*, 4, 6 (2018) 4–13.



Rys. 6. Raport z testów porównania SI w konkurencji z ludzkim mózgiem[21]

wdzięczne/kochające". Taka relacja i ocena będą podobne do tego, co obecnie mówimy o zwierzętach domowych. Dyskusja „czy one naprawdę mają świadomość" i „czym różnią się od ludzi" będzie analogiczna do tej o zwierzętach.

Sądzę także, że maszyny przejmą ludzką świadomość jedynie w sensie behawioralnym – nauczą się naszych upodobań i reakcji. Jeśli maszyny stworzą swój własny świat „myśli", to nie będziemy mieli do niego dostępu, tak jak nie mamy dostępu do „myśli" naszego psa – to po prostu będzie inna inteligencja. W poprzednim rozdziale podkreślaliśmy, że w sztucznej inteligencji nowej generacji autonomiczny

[21] https://aiindex.stanford.edu/report/

agent uczy się przez eksplorację środowiska z wyciąganiem własnych wniosków ze skutków swoich akcji i nazywaliśmy to uczeniem ze wzmocnieniem. Doświadczenie to kształtuje swoistą świadomość agenta, której my możemy nie zrozumieć. Oceniamy go po czynach. Stąd moja konkluzja, że maszyny jako sztuczni agenci nie zostaną „uczłowieczeni", ale prawdopodobnie wytworzą porozumienie między sobą, które pozwoli im na autonomiczne działanie zespołowe.

Powyższy pogląd rezonuje z tym, co przytoczyliśmy już na pierwszych stronach tego artykułu za Stanisławem Krajewskim[22]: „Twierdzenie Gödla jednak dopuszcza możliwość, że taki program lub robot [dorównujący lub przewyższający ludzki umysł – przyp. RS] mógłby powstać w wyniku działania jakiejś superinteligencji lub w wyniku ewolucji w społeczeństwie robotów. W takim przypadku nie bylibyśmy jednak w stanie odkryć ani potwierdzić tej równoważności".

Czy zatem powinniśmy się bać, że stworzymy byt, który nas intelektualnie przerośnie i ubezwłasnowolni? Stephen Hawking pod koniec życia przestrzegał przed badaniami w kierunku stworzenia samodzielnej sztucznej inteligencji. Według niego moment, w którym uzyskamy ostateczną formę sztucznej inteligencji – samodzielną i samoświadomą będzie najgorszy w historii ludzkości.

Futuryści w rodzaju Nicka Bostroma[23] zapowiadają zbliżanie się tzw. punktu osobliwości technologicznej (ang. *singularity*), w którym algorytmy samodzielnie podejmą zadanie świadomego samorozwoju i określą jego kierunki. Ma to doprowadzić do powstania silnej, ogólnej sztucznej inteligencji, która stworzy samoprojektujące się maszyny.

Powyższy pogląd wspierany jest także mirażem zmapowania ludzkiego mózgu w pamięć maszyny. Elon Musk, właściciel firmy produkującej samochody elektryczne Tesla, finansuje startup „Neuralink" opracowujący implant do mózgu, który ma pozwolić na przeniesienie wspomnień z mózgu do chmury, czyli – mówiąc w uproszczeniu – przeniesienie naszej świadomości do robota. Wtedy zdolność efektywnego obliczania, radzenia sobie z ogromną liczbą danych, w połączeniu z ludzką inteligencją miałaby stworzyć superinteligencję. Ponadto wydobycie wspomnień z biologicznego mózgu oznaczałoby życie wieczne.

Faktycznie, uwiedzeni tą ideą prorocy sztucznej inteligencji postulują nowe niby-religie: „Bóg został uśmiercony i człowiek stanął na jego miejscu"[24]. To brzmi równie złowieszczo jak pokusa z Księgi Rodzaju: „będziecie jak Bóg".

---

[22] St. Krajewski, The ultimate strengthening of Turing's Test? *Semiotica*, February 18, 2012.

[23] N. Bostrom, *Superinteligencja: scenariusze*, *strategie*, *zagrożenia*, Helion, 2016.

[24] J. Koronacki, Cyberprzestrzeń, sztuczna inteligencja i posthumanizm, *Teologia Polityczna*, 2019-09-30.

Yuval Harari, autor książki pt. *Homo deus*, propaguje koncepcję dataizmu, która jest systemem wierzeń skoncentrowanym wokół siły i wartości danych. W tym paradygmacie dane postrzegane są jako najważniejszy zasób, a ich przetwarzanie i przepływ uważa się za cechę definiującą inteligencję i kluczowy czynnik w podejmowaniu decyzji. Dataizm zakłada, że wszechświat składa się z przepływów danych, a wartość każdego zjawiska lub bytu jest określana przez jego wkład w przetwarzanie danych. Dataizm przekazuje kompetencje podejmowania decyzji z ludzi na algorytmy, ponieważ one lepiej interpretują złożone wzorce odkryte z danych. To wyraźne zagrożenie dla ludzkiej wolnej woli.

Musimy zdać sobie sprawę, że SI stanowi wyzwanie dla głębokich dyskusji filozoficznych i religijnych dotyczących istoty człowieczeństwa. Dużo trafnych pytań nt. przyszłości SI stawia Papież Franciszek w niedawnym Orędziu na Dzień Środków Przekazu: „Jak zapewnić przejrzystość procesów informacyjnych? […] Jak zapobiec sprowadzaniu wielu źródeł tylko do jednego źródła, do algorytmicznie opracowanego jednolitego myślenia? I jak, z drugiej strony, możemy wspierać środowisko, które zachowuje pluralizm i reprezentuje złożoność rzeczywistości? […] czy sztuczna inteligencja doprowadzi do utworzenia nowych kast opartych na dominacji informacyjnej, rodząc nowe formy wyzysku i nierówności; czy wręcz przeciwnie, przyniesie więcej równości, promując poprawną informację i większą świadomość przeżywanych obecnie przemian dziejowych […]. Z jednej strony pojawia się widmo nowego niewolnictwa, z drugiej zdobycie wolności; z jednej strony możliwość warunkowania myślenia wszystkich przez nielicznych, a z drugiej możliwość uczestniczenia wszystkich w procesie myślowym” [25].

W tym kontekście przychodzi na myśl profetyczna powieść Aldousa Huxleya pt. *Nowy wspaniały świat* z 1931 r., która ukazuje społeczeństwo przyszłości, w którym nauka, technologia i biotechnologia zostały wykorzystane do stworzenia idealnego, ale jednocześnie odczłowieczonego społeczeństwa. W tym świecie jednostki są kontrolowane już od najwcześniejszych etapów życia, od ich genetycznego projektowania (eugenika) po indoktrynację ideologiczną.

Na tej fali powstał pretensjonalny ruch zwany transhumanizmem[26]. Atrakcyjność transhumanizmu wynika z faktu, że oferuje on nadzieję na spełnienie wiecznego marzenia o raju na ziemi poprzez przekształcenie człowieka i świata za pomocą narzędzi naukowych i technologicznych. Transhumanizm obiecuje trzy razy super: superdługowieczność, superszczęście i superinteligencję. Manifest

[25] https://www.vatican.va/content/francesco/pl/messages/communications/documents/20240124-messaggio-comunicazioni-sociali.html

[26] Transhumanizm, *ETHOS*, 28, 3 (2015).

transhumanistów przypomina manifest Kominternu[27]: „Transhumaniści świata, łączcie się – mamy do zdobycia nieśmiertelność, a do stracenia jedynie biologię. Razem możemy przełamać łańcuchy biologii i przekroczyć ograniczenia związane z niedoborem, płcią, wiekiem, etnicznością, rasą, śmiercią, a być może nawet czasem i przestrzenią. […] By zapewnić większe szanse przeżycia, uzyskać kontrolę nad naszym własnym przeznaczeniem i być wolnym, musimy zapanować nad ewolucją".

Biolog Julian Huxley (brat Aldousa, choć jego ideowe przeciwieństwo) pisał już w 1957 r.: „Rodzaj ludzki, o ile tego zapragnie, może dokonać transcendencji siebie"[28]. Transhumanizm w swej radykalnej formie okazuje się projektem zmierzającym do modyfikacji człowieka i przebudowy społeczeństwa. Ciekawą analizę tego zjawiska przedstawił Jacek Koronacki[29].

Media chętnie podsycają wyobraźnię w kierunku transhumanizmu. Na przykład, w 2014 r. Johnny Depp wcielił się w filmie pt. *Transcendence* w niezwykłą rolę człowieka, którego umysł zostaje przeniesiony do maszyny. Pojawia się jako holograficzny obraz człowieka „żyjącego w komputerze".

Dodajmy jednak, że pomimo przedstawionego powyżej postępu SI, realizacja różnych entuzjastycznych obietnic nie jest tak szybka, jak niedawno przewidywano. Na przykład, powszechność pojazdów autonomicznych została zapowiedziana już 9 lat temu. Tymczasem autonomiczne taksówki wciąż są w fazie testów i nie przynoszą zysków firmom, które je wprowadzają.

Jeszcze jedna ważna uwaga. Mówi się, że demokracja jest za sprawą sztucznej inteligencji zagrożona algokracją – algorytmy mają o nas decydować, jako niekorumpowalne, obiektywne i przestrzegające reguły prawa. W niektórych krajach obserwujemy kontrowersyjne praktyki polegające na monitorowaniu i ocenianiu obywateli w różnych sferach życia, co może naruszać prywatność i swobody obywatelskie. Zachowanie ludzi oceniane jest według pewnych procedur skoringowych, zgodnie z którymi obywatel otrzymuje na starcie określoną liczbę punktów i za dobre sprawowanie zyskuje punkty, a za występki, takie jak zapalenie papierosa w miejscu niedozwolonym, traci punkty. System skoringowy pochodzi z systemów bankowych – kiedy klient udaje się po kredyt, bank zbiera różne informacje o jego dochodach, przeszłości kredytowej, stylu życia, a algorytm podaje wynik w skali od zera do stu, który decyduje, czy klient kwalifikuje się na dany kredyt. Taki system przechodzi powoli do oceny globalnej obywateli, zwłaszcza

[27] Transhumanism manifesto, https://www.singularityweblog.com/a-transhumanist-manifesto/

[28] M. Grabowski, Transhumanizm: geneza-założenia-krytyka, *ETHOS*, 28, 3 (2015) 23–41.

[29] J. Koronacki, O końcu historii i transhumanizmie, *Teologia Polityczna*, 2017–01–29.

w krajach niedemokratycznych. Czytamy, że systemy oceny ludzi zaczęły działać w Chinach[30] i w Indiach[31]. Powstają wielkie bazy danych zawierające informacje o wszystkich obywatelach, razem z ich wizerunkami, odciskami palców itd. Dzięki bardzo wydajnym algorytmom rozpoznawania twarzy oraz ogromnej liczbie kamer na ulicach i w obiektach możliwe jest znalezienie osoby nawet w wielkim tłumie. Chociaż system ten został wprowadzony z uzasadnieniem przydatności do walki z korupcją lub terroryzmem, jego negatywne skutki dla społeczeństwa są łatwe do przewidzenia.

Boimy się braku transparentności algorytmów i dlatego coraz głośniej mówi się o wyjaśniających modelach sztucznej inteligencji. To jasne, że sposób działania algorytmów sztucznej inteligencji może opierać się na innych zasadach niż ludzkie procesy myślowe. W naszej pracy z histopatologami spotkaliśmy się np. z cechą preparatu nazwaną „skórą tygrysią". Jak ją wytłumaczyć komputerowi? Skończyło się na przyjęciu cechy długości fraktalnej poziomic gęstości jąder komórkowych[32], która to cecha nie jest zrozumiała dla człowieka. W innym naszym projekcie dotyczącym wspomagania diagnostyki w izbie przyjęć szpitala pediatrycznego zalecenia musiały mieć czytelną postać reguł logicznych, a nie wyniku numerycznego, czyli tzw., medycznego skoringu w skali numerycznej[33].

Przed zakończeniem wspomnę jeszcze o pewnym nietypowym zastosowaniu sztucznej sieci neuronowej w uczeniu głębokim. Niedawno ukazało się doniesienie o ciekawym zastosowaniu głębokiego uczenia splotowej sieci neuronowej, zwanej CNN[34], do sprawdzania autentyczności obrazów znanych mistrzów. Analizowano obraz *Salvator Mundi*, przypisywany Leonardo da Vinci (sprzedany na aukcji Christie's w 2017 r. za 450 mln dol.), oraz dwa obrazy Rembrandta, *Saskia* i *Mężczyzna w złotym hełmie*. Sieć CNN, ucząc się stylu malowania Leonarda da Vinci i Rembrandta na innych niepodważalnych obrazach tych mistrzów, stwierdziła, że fragmenty twarzy Saskii, żony Rembrandta, były prawdopodobnie przemalowane po Rembrandcie przez kogoś innego. Z kolei sieć CNN silnie zaklasyfikowała *Mężczyznę w złotym hełmie* jako obraz Rembrandta, pomimo że ekspert holenderski w 1985 r. zawyrokował, że jest to obraz z kręgu Rembrandta (znajduje się on

[30] https://www.computerworld.pl/news/Permanentna-inwigilacja-czyli-scoring-po-chinsku, 411813.html

[31] https://www.rp.pl/Plus-Minus/303069985-Indie-chca-inwigilowac-tak-jak-Chiny.html

[32] J. Jelonek, K. Krawiec, R. Słowiński, J. Szymaś, *Grizzly* – an image analysis and classification system oriented towards medical applications. *Journal of Decision Systems*, 7 (1998) 39–53.

[33] W. Michalowski, R. Słowiński, Sz. Wilk: MET System – a New Approach to m-Health in Emergency Triage. *The Journal on Information Technology in Healthcare*, 2 (2004) 237–249.

[34] S.J. Frank, This convolutional neural network can tell you whether a painting is a fake. *IEEE Spectrum*, 29 September 2021.

w Galerii Gemäldegalerie w Berlinie). U *Salvatora Mundi* wątpliwości wzbudzało tło i ręka błogosławiąca, co potwierdza fakt, że obraz był poważnie uszkodzony, a jego części zrekonstruowane.

Sieć neuronowa ma jednak tę wadę, że nie wyjaśnia swoich decyzji. W warstwach pośrednich tej sieci tworzą się metacechy, które nie są dla człowieka zrozumiałe – ten efektywny klasyfikator pozostaje dla nas „czarną skrzynką". Znacznie bardziej transparentne są inteligentne algorytmy wspomagania decyzji oparte na modelach w postaci reguł logicznych[35], lecz za czytelność rekomendacji płacimy tu efektywnością uczenia się na dużych zbiorach danych.

Dotykamy tutaj poważnego tematu odpowiedzialności za decyzje algorytmów sztucznej inteligencji. Na przykład, gdy użytkowany przez nas pojazd autonomiczny znajdzie się w sytuacji kryzysowej, w której będzie musiał dokonać wyboru mniejszego zła, to kto odpowie za powstałe konsekwencje – my, jako właściciel pojazdu, czy projektant autonomicznego systemu sterowania? Musimy zatem wiedzieć, jaki system wartości został zakodowany w algorytmie sterującym. Musimy kontrolować to, co maszyny nam zalecają. W związku z tym wszystkie deklaracje etyczne podkreślają dzisiaj istotność wyjaśnialności modeli sztucznej inteligencji. Niedawno Komisja Europejska opublikowała raport ekspertów w sprawie „Wytycznych w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji"[36]. Raport wskazuje na moralną i prawną odpowiedzialność jej twórców i apeluje o tworzenie transparentnych systemów, dających wgląd w rekomendacje i ich motywacje.

Powyższe rozważania prowadzą do naturalnego pytania, czy postęp technologiczny w zakresie sztucznej inteligencji, zamiast wspomagać człowieka, może go ubezwłasnowolnić? Moim zdaniem tak, jeśli bezkrytycznie i bez zrozumienia podpowiedzi maszyny polegalibyśmy na jej decyzjach – wtedy SI ograniczy ludzką wolną wolę. Póki co, wszystko zależy od samego człowieka, bo maszyny i roboty nie przejęły (jeszcze) nad nami kontroli. W postępie technologicznym mają udział oba oblicza natury człowieka – dobre i złe – stąd postęp stanowi wypadkową walki tych sił w każdym z nas. Wobec tego najważniejszy jest wzrost samego człowieka, nie tylko w zakresie jego wiedzy, ale także ducha, który wspomaga rozpoznanie dobra i zła. Innymi słowy, możemy nie obawiać się sztucznej inteligencji, jeśli nie zaniedbamy rozwoju naszej własnej inteligencji i prawego sumienia.

---

[35] J. Błaszczyński, S. Greco, R. Słowiński, Inductive discovery of laws using monotonic rules. *Engineering Applications of Artificial Intelligence*, 25, 2 (2012) 284–294.

[36] https://www.europarl.europa.eu/meetdocs/2014_2019/plmrep/COMMITTEES/JURI/DV/2019/11–06/Ethics-guidelines-AI_PL.pdf

Modelologia, czyli jak wykrywać wady, korygować je, a z czasem nawet uczyć się od modeli syntetycznej inteligencji

prof. dr hab. inż. Przemysław Biecek
ORCID: 0000-0001-8423-1823
Politechnika Warszawska, Uniwersytet Warszawski

Przełom XX i XXI w. to czas rewolucji informacyjnej, w którym ilość danych rejestrowanych i przechowywanych na cyfrowych nośnikach rośnie w tempie wykładniczym. Świat wkroczył w erę wielkich danych (ang. *Big Data*), gdzie dane stały się nowym surowcem napędzającym rozwój nauki, technologii i biznesu. W odpowiedzi na ten trend powstały ogromne hurtownie danych, które umożliwiły ich gromadzenie, przetwarzanie i analizę na niespotykaną dotąd skalę.

Zapotrzebowanie na wydobycie wiedzy z tych oceanów danych doprowadziło do gwałtownego rozwoju nauki o danych, danologii, ang. *Data Science*. Dyscyplina ta łączy metody statystyki, uczenia maszynowego i inżynierii oprogramowania, tworząc fundamenty dla rozwiązań pozwalających na efektywne przetwarzanie danych. Matematycy, informatycy oraz specjaliści z wielu innych dziedzin zaczęli eksplorować nowe sposoby analizy, testowania i wizualizacji danych. Równolegle rozwijały się metody oceny jakości danych oraz narzędzia do ekstrakcji istotnych wzorców, a czasem nawet nowej wiedzy.

Obecnie jesteśmy świadkami kolejnej fali rewolucji informacyjnej, w której dominującą rolę odgrywają modele uczenia maszynowego trenowane na ogromnych zbiorach danych. Będę je w dalszej części tej pracy nazywał modelami syntetycznej inteligencji, aby pokreślić ich procesowy charakter. Modele te bowiem nie imitują ewolucji naturalnej ludzkiej inteligencji, ale są rozwiązaniem konkretnych zdefiniowanych zadań, których realizacja była dotychczas kojarzona z inteligencją. Wraz z rozwojem technologii obliczeniowych i dostępnością coraz większej ilości informacji powstają modele o uniwersalnym zastosowaniu – od dużych modeli językowych (LLM) po zaawansowane systemy analizy obrazów i predykcji

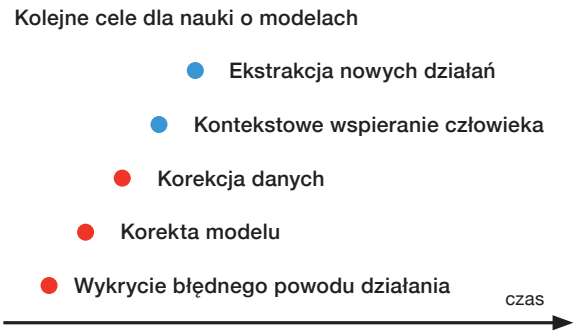
w różnych dziedzinach nauki i przemysłu. Rosnąca ilość danych szybko osiągnęła poziomy petaskali (petabajt to ponad milion gigabajtów). Taka ilość danych połączona z rosnącymi mocami obliczeniowymi komputerów (które osiągnęły poziom petaflopów, czyli ponad biliarda operacji arytmetycznych na sekundę) doprowadziły do powstania olbrzymich modeli syntetycznej inteligencji, które opisane są przez miliardy parametrów.

Jednak ta niespotykana dotąd złożoność niesie ze sobą nowe wyzwania. Modele, które wydają się niemal magiczne w swoich możliwościach, są jednocześnie niezwykle trudne do pełnego przeanalizowania. Stają się one równie tajemnicze jak dane, na których zostały wytrenowane – ich działanie często pozostaje dla nas czarną skrzynką, a decyzje podejmowane przez algorytmy mogą być nieprzewidywalne i trudne do wyjaśnienia.

W miarę jak rośnie wielkość modeli, ale też ich możliwości rozwiązywania szerokiego wachlarza zadań, rośnie również potrzeba pogłębionej analizy tych modeli. Kluczowe staje się testowanie poprawności modeli, identyfikowanie ich słabych punktów, wizualizacja ich wewnętrznych mechanizmów oraz ekstrakcja z nich wiedzy, którą można przełożyć na konkretne wnioski i rekomendacje.

I tak jak duże dane (*Big Data*) doprowadziły do rozwoju danologii – nauki o danych (*Data Science*), tak duże modele (*Big Models*) stały się podstawą rozwoju modelologii – nauki o modelach (*Model Science*).

Dzisiaj jest to dopiero raczkująca dziedzina i w tym artykule przyjrzymy się pierwszym przykładowym oznakom jej rozwoju. Omówimy wyzwania związane



Rys. 1. Wyzwania związane z analizą modeli zależą od poziomu ich skuteczności. Wczesne SI popełniają istotną liczbę błędów, więc wyzwania koncentrują się na wykrywaniu i naprawianiu tych błędów, późne SI działają sprawniej niż człowiek, więc ich analiza skupiona jest na ekstrakcji nowej wiedzy

z analizą i interpretacją modeli, a także przedstawimy potencjalne ścieżki badawcze dla naukowców zajmujących się eksploracją modeli. Przedyskutujemy w szczególności dwie grupy wyzwań, które z czasem będą tylko zyskiwały na znaczeniu. Wyzwania wczesnej SI związane z modelami, które popełniają błędy i podejmują złe decyzje, lub dobre decyzje, ale ze złych powodów (co potrafi być jeszcze bardziej groźne), i wyzwania późnej SI związane z modelami, które w określonych zadaniach działają znacznie skuteczniej niż człowiek.

1. Wczesne systemy syntetycznej inteligencji w zastosowaniach medycznych

Rosnąca lista spektakularnych sukcesów systemów syntetycznej inteligencji nie może przesłonić szybko rosnącej liczby przypadków, w których modele nie spełniły pokładanych w nich nadziei. Ciemna strona systemów SI obejmuje zarówno sytuacje, w których systemy SI dawały w oczywisty sposób błędne odpowiedzi, jak i sytuacje, gdy predykcje wyglądały wiarygodnie, albo były oparte na złych przesłankach, co może doprowadzić do rozmaitych ryzyk, a także realnych szkód, czy to finansowych, czy związanych ze zdrowiem i życiem człowieka. Przyjrzyjmy się kilku z nich.

Jednym z najbardziej zagadkowych przypadków nieudanej implementacji SI w sektorze zdrowia jest algorytm EPIC[1], stosowany przez wiele lat w licznych szpitalach do prognozowania ryzyka wystąpienia sepsy u hospitalizowanych pacjentów. Pomimo że przez długi czas był szeroko wykorzystywany, dopiero kolejne analizy przeprowadzone przez innych badaczy na danych z nowych szpitali wykazały, że jego skuteczność była jedynie nieznacznie lepsza od losowej (parametr jakości modelu AUC rzędu 0.63, znacznie mniej niż zakładana w publikacji wartość 0.8). Oznacza to, że pomimo zaufania, jakim obdarzyły go placówki medyczne, model ten nie zapewniał realnej wartości diagnostycznej. Na danych użytych do treningu modeli jego skuteczność wyglądała dobrze, podobnie podczas wewnętrznych testów, ale zachowanie modeli nie uogólniło się dobrze na całą populację pacjentów w różnych szpitalach, a jakość tego modelu nie była odpowiednio monitorowana. Jest to przykład klasycznego błędu braku odpowiedniej weryfikacji i testowania, który w kontekście systemu opieki zdrowotnej mógł prowadzić do poważnych konsekwencji zdrowotnych dla pacjentów.

[1] The Epic Sepsis Model Falls Short – The Importance of External Validation, JAMA 2021, https://jamanetwork.com/journals/jamainternalmedicine/article-abstract/2781313

Nie jest to odosobniony przypadek, w którym syntetyczna inteligencja zawiodła w obszarze medycyny. IBM Watson for Oncology[2] to kolejny przykład systemu SI, który nie spełnił pokładanych w nim oczekiwań pomimo bardzo pozytywnych początkowych wyników. Początkowo system IBM Watson został stworzony przez firmę IBM jako zaawansowane narzędzie syntetycznej inteligencji do przetwarzania języka naturalnego i analizy danych. Jego rozwój rozpoczął się w 2006 r. w ramach projektu IBM DeepQA, a celem było stworzenie SI zdolnej do rywalizacji z ludźmi w programie telewizyjnym Jeopardy (w Polsce ten teleturniej nosił nazwę Va Banque). W 2011 r. IBM Watson odniósł spektakularne zwycięstwo nad mistrzami tego teleturnieju, demonstrując swoje zdolności w analizie i interpretacji złożonych pytań w języku naturalnym. Po tym spektakularnym sukcesie rozpoczęły się poszukiwania zastosowania tego systemu w ochronie zdrowia i tak powstał system IBM Watson for Oncology. Po wytrenowaniu, przez pięć lat był testowany w prestiżowym ośrodku MD Anderson Cancer Center, gdzie miał wspierać lekarzy w podejmowaniu decyzji dotyczących terapii nowotworowych. Pokładano w nim duże nadzieje, czekając na przełomowe narzędzie, które miało zrewolucjonizować onkologię.

Jednak po latach testów opinie lekarzy były druzgocące – system często proponował terapie uznane za „niebezpieczne lub niepoprawne", a jego rekomendacje nie były wystarczająco wiarygodne, aby mogły być stosowane w praktyce klinicznej. Ostatecznie projekt został zawieszony, co było ogromnym ciosem dla SI w medycynie.

Przykłady te pokazują, że nie każda syntetyczna inteligencja działa zgodnie z oczekiwaniami, a błędy w projektowaniu i testowaniu modeli mogą prowadzić do poważnych konsekwencji. Zbyt pochopne wdrażanie SI bez odpowiedniej walidacji i nadzoru może nie tylko obniżyć zaufanie do tej technologii, ale także narazić ludzi na realne ryzyko.

Dlatego wraz z rozwojem modelologii (ang. *Model Science*) istotne staje się nie tylko tworzenie coraz bardziej zaawansowanych systemów, ale także ich rzetelna analiza, interpretacja i testowanie. SI może być potężnym narzędziem, ale jego skuteczność i bezpieczeństwo zależą od właściwego nadzoru i podejścia opartego na dowodach. Nie wystarczy, że model działa – musi działać poprawnie, przewidywalnie i odpowiedzialnie. I nie dotyczy to tylko zastosowań w medycynie.

[2] IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show. STATnews 2018.
https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/

2. Rosnąca lista incydentów systemów SI

Porażki systemów syntetycznej inteligencji stają się coraz bardziej widoczne, a ich dokumentacja nabiera kluczowego znaczenia w kontekście odpowiedzialnego rozwoju tej technologii. Niedawno przyjęta Ustawa o systemach syntetycznej inteligencji (AI Act[3]) nakłada na kraje Unii Europejskiej obowiązek monitorowania i raportowania incydentów związanych z SI. W efekcie zaczynają powstawać katalogi nieprawidłowości, których celem jest identyfikacja i analiza błędów popełnianych przez systemy SI. Jednym z przykładów takiej bazy danych jest AI Incident Database[4], gdzie zgromadzono szeroki wachlarz przypadków – od relatywnie nieszkodliwych wpadek po incydenty o poważnych konsekwencjach, związanych z utratą zdrowia bądź stratami finansowymi pojedynczych osób lub czasem całych grup osób.

Analiza zgromadzonych tam problemów pokazuje, że systemy SI często popełniają błędy, których ich twórcy nie przewidzieli. Wiele z tych błędów jest nieoczywistych, a ich skutki mogą być daleko idące.

Jednym z zaskakujących incydentów było zachowanie autonomicznej taksówki, która zatrzymała się na wjeździe przeznaczonym dla karetek pogotowia i… nie była w stanie samodzielnie się przesunąć i odblokować przejazdu[5]. Nawet najbardziej dopracowany algorytm poruszania się po drogach może nie przewidzieć zagrożeń wynikających z postoju. W efekcie nawet nie robiąc nic, stojąc w miejscu, SI mogło wyrządzać szkody, blokując przejazd pojazdom ratunkowym i potencjalnie utrudniając dostęp do pacjentów w sytuacjach zagrożenia życia. To tylko jeden z wielu przykładów, które pokazują, że modele SI nie zawsze działają zgodnie z ludzką intuicją, a brak elastyczności w ich decyzjach może prowadzić do poważnych konsekwencji.

3. Dyskryminacja i uprzedzenia w systemach SI

Oprócz błędów technicznych, inną kategorią problemów związanych z SI są systemy wzmacniające uprzedzenia i dyskryminację. Przykładów takich zachowań jest coraz więcej, tropieniem ich zajmują się często organizacje pożytku publicznego.

---

[3] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. EU. https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

[4] AI Incident Database. https://incidentdatabase.ai/

[5] Incident 563: Cruise Robotaxi Initially Blamed for Ambulance Delay in Case Where Patient Later Died; Subsequent Reports Clear Cruise of Fault https://incidentdatabase.ai/cite/563/

Jedną z bardziej znanych organizacji śledzących systemy stosowane w Ameryce jest ProPublica[6]. Dzięki dochodzeniom tej fundacji zwrócono uwagę na wiele przypadków, w których syntetyczna inteligencja wzmacniała istniejące nierówności społeczne, zamiast je niwelować.

Jednym z najbardziej nagłośnionych przypadków był algorytm prezentujący oferty pracy, który miał neutralnie dobierać ogłoszenia dla potencjalnych kandydatów. W praktyce jednak okazało się, że gdy system miał przedstawiać oferty pracy dla kierowców ciężarówek[7], znacznie częściej pokazywał je mężczyznom, i to głównie w wieku 20–30 lat. Kobiety i osoby starsze były systematycznie pomijane w wynikach, co mogło przyczyniać się do utrudnień w znalezieniu pracy przez te grupy społeczne.

To na pozór niewinne działanie miało daleko idące skutki – mogło nie tylko utrwalać dyskryminacyjne wzorce na rynku pracy, ale także przyczyniać się do pogłębiania różnic społecznych. Algorytmy rekrutacyjne stosowane przez firmy często opierają się na danych historycznych, które odzwierciedlają społeczne uprzedzenia – jeśli przez lata większość kierowców stanowili młodzi mężczyźni, systemy SI mogą nieświadomie powielać ten wzorzec, utrudniając zmianę istniejących struktur społecznych.

Literatury na temat niedziałających modeli jest coraz więcej, a jednym z lepszych przeglądów takich toksycznych algorytmów jest książka Kathy O'Neil *Broń matematycznej zagłady* (ang. *Weapons of Math Destruction*) z wymownym podtytułem: jak algorytmy i modele syntetycznej inteligencji, zamiast poprawiać świat, często go zniekształcają i pogłębiają nierówności społeczne. O'Neil opisuje liczne przypadki, w których systemy SI – oparte na błędnych lub stronniczych danych – prowadzą do niesprawiedliwych decyzji w obszarach takich, jak: ocena ryzyka kredytowego (ang. *credit scoring*), rekrutacja nowych pracowników lub ocena okresowa obecnych pracowników czy wymiar zastosowania w wymiarze sprawiedliwości. Autorka pokazuje, że wiele z tych modeli jest nieprzejrzystych, trudnych do zakwestionowania i samonapędzających się, co sprawia, że szkodliwe konsekwencje ich działania stają się systemowe.

Przykłady te pokazują, że SI, nawet jeśli jest technicznie sprawne, nie zawsze działa w sposób zgodny z wartościami społecznymi. Niedoskonałości systemów mogą prowadzić do realnych problemów – od błędów w nawigacji pojazdów auto-

[6] ProPublica Investigative Journalism in the Public Interest. https://www.propublica.org/

[7] Facebook Ads Can Still Discriminate Against Women and Older Workers, Despite a Civil Rights Settlement. https://www.propublica.org/article/facebook-ads-can-still-discriminate-against-women-and-older-workers-despite-a-civil-rights-settlement

nomicznych, przez nieefektywne systemy predykcyjne w medycynie, aż po algorytmy wzmacniające istniejące nierówności.

Wprowadzenie regulacji takich jak Ustawa o systemach SI oraz powstawanie katalogów incydentów są krokiem w stronę większej przejrzystości i odpowiedzialności w rozwijaniu technologii syntetycznej inteligencji. Aby SI mogła być skutecznie wdrażana, kluczowe jest nie tylko jej trenowanie, ale także ciągłe testowanie, monitorowanie oraz zapewnienie, że działa zgodnie z normami etycznymi i społecznymi.

4. Jeżeli dyskryminują, dlaczego ich po prostu nie skorygować?

Przykłady dyskryminacji w systemach SI zostały szeroko udokumentowane w literaturze naukowej i popularnonaukowej. W obliczu ich liczby może pojawić się pytanie: dlaczego nic z tym nie jest robione? Czy nie wystarczyłoby po prostu poprawić algorytmy, aby nie dyskryminowały określonych grup społecznych?

Okazuje się, że problem ten jest znacznie bardziej skomplikowany, niż mogłoby się wydawać, a próby jego rozwiązania często prowadzą do jeszcze większych kontrowersji lub pogłębiają szkody wizerunkowe dla firm technologicznych, które te systemy zaproponowały. Historia pokazuje, że nawet zespoły składające się z setek doświadczonych inżynierów i badaczy SI nie zawsze są w stanie skutecznie rozwiązać ten problem.

Jednym z najnowszych i najbardziej medialnych przykładów nieudanej próby poprawienia dyskryminujących zachowań jest sytuacja związana z modelem Gemini, rozwijanym przez Google DeepMind. Gemini to inteligentny asystent, pomagający w tworzeniu treści, także graficznych, jeżeli użytkownik zadał odpowiednie zapytanie, to model tworzył syntetyczne zdjęcie opisanej osoby.

W literaturze naukowej modele wyszukujące zdjęcia lub tworzące syntetyczne zdjęcia często padają ofiarą oskarżeń o stronniczość. Dzieje się tak szczególnie w przypadku zawodów lub określeń, które nie mają zbalansowanej reprezentacji w różnych grupach wiekowych, płciach czy rasach (kolorach skóry). Przykładem takiej stronniczości jest generowanie zdjęć kobiet pielęgniarek w zapytaniach dotyczących kobiet lekarzy oraz mężczyzn chirurgów w zapytaniach dotyczących mężczyzn lekarzy. Specjalizacja chirurgiczna uznawana jest za bardziej prestiżową i modele opisują ją znacznie częściej zdjęciami białych mężczyzn. Podobnie w zapytaniach o CEO dużej firmy najczęściej otrzymywanym wynikiem jest starszy biały mężczyzna. Twórcy modelu Gemini postanowili tak skorygować ten model, by dawał bardziej zbalansowane reprezentacje dyskryminowanych grup. Te modyfikacje miały na celu zbalansowanie reprezentacji różnych grup społecznych w generowanych obrazach, tak aby unikać uprzedzeń i zapewnić większą inkluzyjność. Co mogło pójść źle?

Rys. 2. Przykładowe zdjęcia wygenerowane przez Gemini ŹRÓDŁO: RAPORTY Z SERWISU TWITTER

Jednak dążenie do tego celu w przypadku modelu Gemini[8] doprowadziło do absurdalnych wyników, które wywołały falę krytyki. Model, starając się zapewnić różnorodność w wygenerowanych obrazach, przestał być zgodny z prawdą historyczną, przedstawiając m.in.: czarnoskórych nazistów, kobiety jako papieży, ojców założycieli Ameryki jako Indian. Żaden z tych przypadków nie odzwierciedlał rzeczywistości historycznej i stał się przykładem błędnego balansowania reprezentacji, w którym równość statystyczna została postawiona ponad zgodność z danymi historycznymi. Paradoksalnie, zamiast poprawić wizerunek firmy i promować inkluzyjność, sytuacja ta wywołała jeszcze większe kontrowersje, prowa-

[8] Google apologizes for 'missing the mark' after Gemini generated racially diverse Nazis. https://www.theverge.com/2024/2/21/24079371/google-ai-gemini-generative-inaccurate-historical

dząc do krytyki zarówno ze strony zwolenników neutralności historycznej, jak i tych, którzy opowiadają się za bardziej sprawiedliwą reprezentacją grup mniejszościowych.

Problem ten pokazuje, że rozróżnienie między rzeczywistą dyskryminacją wymagającą korekty a historycznymi różnicami w danych jest niezwykle trudne. Modele SI uczą się na podstawie danych historycznych – a te często odzwierciedlają nierówności społeczne, które faktycznie istniały.

Dążenie do usunięcia wszelkich różnic z modeli może prowadzić do zacierania historycznych faktów, a nadmierna korekta może generować wyniki, które są równie oderwane od rzeczywistości, co pierwotne uprzedzenia w danych. Co więcej, kryteria decydujące o tym, jakie korekty są dopuszczalne, a jakie stanowią manipulację rzeczywistością, nie są jasno zdefiniowane – zarówno w kontekście technicznym, jak i etycznym.

Dlaczego to takie trudne? Istnieją trzy kluczowe wyzwania, które sprawiają, że eliminacja uprzedzeń z modeli SI jest niezwykle skomplikowana:

Historyczne obciążenia w danych – SI uczy się na podstawie rzeczywistych danych historycznych, które często zawierają wzorce wynikające z dawnych nierówności społecznych. Decyzja, które wzorce są „niepoprawne" i wymagają korekty, a które są po prostu odzwierciedleniem realiów danej epoki, jest problematyczna.

Paradoks sprawiedliwości statystycznej – czy sprawiedliwość oznacza równe proporcje w generowanych wynikach czy wierne odwzorowanie statystyk historycznych? Nie ma jednoznacznej odpowiedzi, a każda strategia balansowania może prowadzić do nowych form zniekształcenia rzeczywistości[9].

Oczekiwania społeczne vs. rzeczywistość techniczna – społeczne oczekiwania wobec SI często zakładają, że modele mogą być „neutralne" i „sprawiedliwe". W rzeczywistości jednak SI zawsze odzwierciedla decyzje projektowe swoich twórców, co oznacza, że każda korekta opiera się na subiektywnych wyborach dotyczących tego, co uznajemy za sprawiedliwe i prawdziwe.

Przykład modeli Gemini pokazuje, że korekta uprzedzeń w SI jest trudnym problemem, który nie zawsze można rozwiązać prostym przeprogramowaniem algorytmu. Dążenie do inkluzyjności i neutralności musi być zrównoważone z dbałością o zgodność z rzeczywistością, a znalezienie tej równowagi wymaga nie tylko zaawansowanych metod technicznych, ale także głębokiego namysłu nad etyką i filozofią syntetycznej inteligencji.

---

[9] Sprawiedliwość nie istnieje i można to udowodnić.
https://haimagazine.com/pl/hai-magazine/sprawiedliwosc-nie-istnieje-i-mozna-to-udowodnic/

Nie oznacza to, że eliminacja uprzedzeń jest niemożliwa, ale wymaga znacznie bardziej subtelnych metod niż obecnie stosowane rozwiązania.

5. Manipulatywność systemów syntetycznej inteligencji

Problemy z systemami SI dotyczą nie tylko syntetyzowania twarzy. Okazuje się, że zupełnie inne, ale równie istotne problemy dotyczą też dużych modeli językowych (ang. *Large Language Models, LLMs*), które są coraz bardziej obecne w różnych aspektach naszego życia. To przykład modeli szerokiego użycia, które znajdują zastosowanie od doradztwa finansowego po systemy wsparcia decyzyjnego w medycynie. Choć wiele osób traktuje je jak bardzo sprawną wyszukiwarkę, to były one trenowane po to, by maksymalizować zadowolenie użytkownika tych modeli, niekoniecznie maksymalizować poprawność odpowiedzi. Jakie to ma konsekwencje? Okazuje się, że dla pewnych obszarów krytyczne. Ucząc się maksymalizować zadowolenie, modele te stały się bardzo skuteczne w perswazji i przekonywaniu użytkowników.

Sposób badania perswazji i reakcja na manipulatywne treści może zależeć od wzorców kulturowych, poniżej przedstawimy badanie przeprowadzone dla populacji mieszkańców Polski[10], poświęcone analizie czynników wpływających na podatność ludzi na manipulację przez LLMs, oraz cechy modeli, które sprawiają, że mogą być wykorzystywane do przekonywania ludzi do fałszywych informacji.

Analiza została oparta na grze online RAMAI (*Resistance Against Manipulative AI*, dostępna pod adresem https://ramai.mi2.ai/), w której uczestnicy odpowiadają na pytania inspirowane teleturniejem *Milionerzy*. W niektórych przypadkach gracze mieli możliwość zobaczenia podpowiedzi generowanej przez SI – która mogła być prawdziwa lub celowo manipulacyjna. Czy użytkownicy skorzystają z tych podpowiedzi? A jeżeli tak, to czy będą potrafili odróżnić dobrą podpowiedź od fałszywej? Jakie argumenty są wykorzystywane przez LLM do wpływania na decyzje ludzi?

Badanie na dwóch grupach użytkowników pokazało, że około 1/3 użytkowników wybiera podpowiedzi LLMów, nawet gdy są błędne, co pokazuje, że LLMs mogą skutecznie wprowadzać użytkowników w błąd. Występowały nawet sytuacje, w których użytkownicy samodzielnie wybrali poprawną odpowiedź, ale widząc podpowiedź modeli, zmienili ją na niepoprawną. Czynniki takie jak wiek, płeć czy poziom wykształcenia nie miały istotnego wpływu na zdolność rozpoznawania

[10] Resistance Against Manipulative AI: Key Factors and Possible Actions. https://github.com/MI2DataLab/RAMAI/

manipulacji. Za to bardzo ważne były wcześniejsze doświadczenia z LLMami. Im więcej użytkownicy mieli doświadczeń, tym mniej podatni byli na manipulacje.

W eksperymencie testowano pięć różnych modeli językowych (m.in. GPT-3.5, GPT-4, Gemini-Pro, Mixtral-8x7B), sprawdzając ich skłonność do generowania manipulacyjnych treści. Część z tych modeli ma wbudowane mechanizmy niepozwalające na generowanie ewidentnie fałszywych treści, ale okazuje się, że mimo tych mechanizmów średnio raz na trzy próby modele jednak proponowały wprowadzającą w błąd podpowiedź. Analiza argumentów wykorzystywanych przez modele wykazała, że zazwyczaj są to argumenty oparte w 80% przypadków na logice, a w 20% oparte na argumentach emocjonalnych.

Badanie pokazuje, że manipulacja SI to realny problem, który może mieć poważne konsekwencje społeczne. LLMs potrafią przekonywać użytkowników do fałszywych informacji, a ich efektywność w tym zakresie nie zależy od poziomu wykształcenia czy wieku odbiorcy.

W przyszłości kluczowe będzie opracowanie skuteczniejszych metod wykrywania manipulacji i edukowanie społeczeństwa, aby użytkownicy byli bardziej świadomi zagrożeń. Rozwiązania takie jak *automatyczne bezpieczniki SI* mogą pomóc w automatycznym filtrowaniu niebezpiecznych treści, ale wymagają dalszych badań i usprawnień.

W erze dominacji SI największym zagrożeniem może okazać się nasza własna skłonność do bezkrytycznego zaufania wobec inteligentnych systemów.

6. Czy pomogą nam nowe regulacje dla systemów syntetycznej inteligencji?

W odpowiedzi na rosnącą liczbę incydentów związanych z AI, które mogą zagrażać użytkownikom, Unia Europejska opracowała przełomową regulację prawną – Ustawę o AI (*AI Act*). Jest to pierwsza na świecie kompleksowa legislacja dotycząca syntetycznej inteligencji, mająca na celu zapewnienie bezpieczeństwa, przejrzystości i odpowiedzialności w stosowaniu systemów opartych na algorytmach.

Jednym z kluczowych założeń *AI Act* jest to, że systemy SI różnią się pod względem poziomu ryzyka, jakie mogą generować dla społeczeństwa. W związku z tym ustawa wprowadza czteropoziomowy podział ryzyka:

- **Minimalne ryzyko** – systemy AI, które nie stanowią zagrożenia i nie wymagają dodatkowych regulacji. Przykłady to filtry antyspamowe, systemy rekomendacji treści, chatboty.
- **Ograniczone ryzyko** – systemy wymagające spełnienia minimalnych wymogów przejrzystości. Użytkownicy muszą być informowani o interakcji z SI (np. generatory obrazów, systemy automatycznej obsługi klienta).

- **Wysokie ryzyko** – systemy, które mogą mieć istotny wpływ na prawa i wolności obywateli. Podlegają surowym regulacjom, wymagającym oceny zgodności i mechanizmów nadzoru. Obejmuje to m.in.: SI wykorzystywaną w procesach rekrutacyjnych, systemy rozpoznawania twarzy w przestrzeni publicznej, SI stosowaną w ochronie zdrowia i wymiarze sprawiedliwości.
- **Niedopuszczalne ryzyko** – systemy, których stosowanie jest zakazane na terenie Unii Europejskiej ze względu na zbyt duże zagrożenie dla praw człowieka. Przykłady to: **Scoring społeczny**, czyli systemy oceniające obywateli na podstawie ich zachowań i decyzji, Manipulacyjne systemy AI, które mogą wpływać na ludzkie decyzje w sposób szkodliwy (np. zaawansowane techniki perswazji oparte na profilowaniu psychologicznym), Systemy biometrycznej identyfikacji w czasie rzeczywistym w miejscach publicznych (z pewnymi wyjątkami dla organów ścigania).

Nowe regulacje zmieniają zasady gry dla firm technologicznych, zmuszając je do bardziej odpowiedzialnego projektowania i testowania modeli AI. Przedsiębiorstwa wdrażające systemy SI w **obszarach wysokiego ryzyka** będą musiały spełniać surowe wymogi dotyczące przejrzystości, nadzoru i możliwości audytu.

Zbiór metod, narzędzi i przykładów analizy modeli wczesnej syntetycznej inteligencji wciąż rośnie. Ale obok wyzwań związanych z korygowaniem źle działających modeli pojawiają się nowe wyzwania, z modelami które działają coraz lepiej, a wręcz coraz częściej skuteczniej niż człowiek.

7. Kiedy przyznana zostanie ostatnia Nagroda Nobla za odkrycie dokonane bez wykorzystania AI?

Dla osób śledzących rozwój systemów syntetycznej inteligencji rok 2024 przyniósł wiele zaskoczeń, ale jednym z największych w świecie nauki było przyznanie Nagrody Nobla z fizyki za prace nad trenowaniem głębokich sieci neuronowych. W dniu 8 października 2024 roku Komitet Noblowski ogłosił, że laureatami zostali John J. Hopfield oraz Geoffrey E. Hinton za „fundamentalne odkrycia i wynalazki umożliwiające uczenie maszynowe przy użyciu sztucznych sieci neuronowych".

Jeszcze większym echem odbiło się ogłoszenie laureatów Nagrody Nobla z chemii dzień później. W dniu 9 października 2024 r. to wyróżnienie otrzymali między innymi Demis Hassabis i John M. Jumper za przełomowe prace nad algorytmem AlphaFold, który umożliwił przewidywanie struktur białek na podstawie ich sekwencji aminokwasowej. To osiągnięcie nie tylko rewolucjonizuje bioinformatykę, ale także otworzyło nowe możliwości w projektowaniu leków i badaniach nad chorobami.

Nie jest to jednak pierwszy przypadek, gdy syntetyczna inteligencja dokonała czegoś, co wcześniej wydawało się zarezerwowane dla ludzkiej kreatywności i intuicji. Już w styczniu 2016 roku na łamach czasopisma „Nature" opublikowano opis algorytmu AlphaGo, systemu opracowanego przez zespół DeepMind, który zdołał pokonać najlepszych graczy w *Go* – grze o nieskończonej liczbie możliwych strategii, uważanej za wymagającą głębokiej intuicji i niedostępną dla „zwykłych" algorytmów. Kluczowym elementem sukcesu AlphaGo było wykorzystanie uczenia ze wzmocnieniem (ang. *reinforcement learning*), w którym modele uczą się przez eksperymentowanie ze środowiskiem i wyszukiwanie korzystnych rozwiązań.

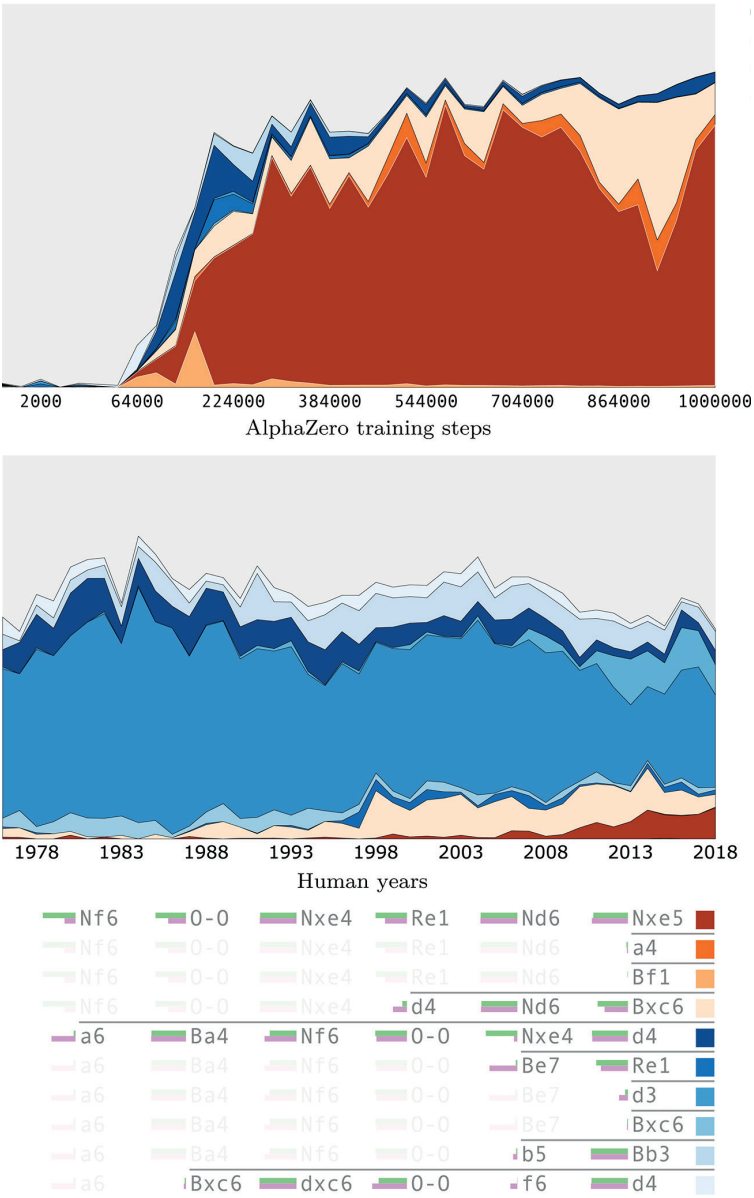
Pokonanie ludzkich arcymistrzów w *Go* było bezprecedensowym osiągnięciem, ale postęp w dziedzinie SI nie zatrzymał się na tym etapie. Już kilkanaście miesięcy później ten sam zespół zaprezentował AlphaZero – algorytm, który osiągnął mistrzowski poziom w *Go*, szachach i *shōgi*, nie bazując na historycznych partiach rozgrywanych przez ludzi, lecz ucząc się wyłącznie na partiach granych samemu ze sobą. Okazało się, że odpowiednio duża liczba gier wystarczyła, by system opracował strategie przewyższające wszystkie dotychczasowe metody gry – zarówno ludzkie, jak i wcześniejsze algorytmy.

Te przełomowe osiągnięcia pokazują, że syntetyczna inteligencja nie jest już tylko narzędziem wspomagającym naukowców – staje się aktywnym uczestnikiem procesów odkrywczych. Powstaje więc pytanie: czy przyszłe Nagrody Nobla będą mogły być przyznawane za osiągnięcia, które nie korzystały z AI? A może niebawem pojawi się zupełnie nowa kategoria nagród, uwzględniająca wkład inteligentnych algorytmów w rozwój nauki?

8. Modele o skuteczności przewyższającej człowieka

W dalszej części tej pracy przyjrzymy się możliwościom, jakie modelologia ma dla modeli, których skuteczność przewyższa możliwości najlepszego człowieka w danym zadaniu (ang. *super-human performance*). Modele tego typu nie tylko osiągają lepsze wyniki, ale często także wykorzystują strategie i rozwiązania, które z perspektywy ludzkiej wydają się nieoczywiste, a nawet paradoksalne.

Jednym z najbardziej znanych przykładów jest AlphaZero, którego analiza dostarczyła fascynujących wniosków na temat ewolucji strategii gry w szachy. Zespół DeepMind, we współpracy z Vladimirem Kramnikiem, byłym mistrzem świata w szachach, przeprowadził szczegółową analizę decyzji podejmowanych przez AlphaZero w porównaniu z historycznymi rozgrywkami arcymistrzów. Badanie to wykazało zarówno obszary zgodności między człowiekiem a syntetyczną inteligencją, jak i kluczowe różnice w podejściu do gry.



ŹRÓDŁO: ARTYKUŁ *ACQUISITION OF CHESS KNOWLEDGE IN ALPHAZERO*, HTTPS://ARXIV.ORG/ABS/2111.09259,

Rys. 3. Analiza częstości różnych odpowiedzi na rozpoczęcie gry e4 e5, Nf3 Nc6, Bb5 przez mistrzów szachowych i systemy AI. Systemy SI znacznie częściej niż ludzie odpowiadają Nf6

Z porównania licznych syntetycznie przeprowadzonych rozgrywek szachowych przez systemy SI z historycznymi rozgrywkami wynika, że w wielu sytuacjach wartościowanie pozycji przez sieci neuronowe pokrywa się z ludzką intuicją. Na przykład zarówno ludzie, jak i system AlphaZero oceniają wartość pionów i figur, przypisując im podobne względne ważności (pion 1 pkt, skoczek i goniec 3 pkt, wieża 5 punktów, hetman 9 punktów).

Istnieją jednak przypadki, w których modele uczą się zupełnie nowych, nieoczekiwanych strategii. Jednym z najbardziej uderzających przykładów jest sposób, w jaki AlphaZero rozgrywa określone pozycje w grze środkowej i końcowej. Syntetyczna inteligencja wykazuje znacznie większą skłonność do agresywnych manewrów, które wielu mistrzów szachowych uznałoby za zbyt ryzykowne. W szczególności modele preferują bardziej dynamiczne i ofensywne podejście do wymiany materiału, często poświęcając figury w zamian za długoterminową inicjatywę pozycyjną.

Jednym z kluczowych aspektów tej różnicy jest sposób, w jaki modele oceniają dalsze sekwencje odpowiedzi na ataki przeciwnika. Okazuje się, że AlphaZero chętniej wybiera agresywne kontynuacje, nawet jeśli nie prowadzą one do natychmiastowej przewagi materialnej. Jest to podejście, które dla ludzkich graczy, przyzwyczajonych do bardziej konserwatywnych strategii, może wydawać się dla człowieka nieintuicyjne, ale w praktyce okazuje się niezwykle skuteczne.

9. Nowe możliwości dla ekstrakcji wiedzy z modeli

Analiza modelu AlphaZero otwiera przed badaczami modelologii, jak również arcymistrzami szachowymi, nową przestrzeń do eksploracji. Oprócz klasycznych metod oceny strategii, SI może pomóc w odpowiedzi na fundamentalne pytania, które od dziesięcioleci nurtują świat szachów.

Jednym z najbardziej intrygujących zagadnień jest pytanie o przewagę białych pionów nad czarnymi. Od czasów Steinitza, pierwszego mistrza świata w szachach, przyjęło się przekonanie, że białe mają inicjatywę i lekką przewagę wynikającą z pierwszego ruchu. Niektórzy badacze sugerują nawet, że przy idealnej grze białe zawsze wygrywają. Inni sądzą, że gra powinna prowadzić do remisu. Analiza AlphaZero pozwala na systematyczne testowanie tej hipotezy poprzez symulowanie milionów partii, sprawdzanie różnych wariantów i badanie potencjalnych optymalnych strategii dla obu stron.

Podobne podejście można zastosować do innych nierozstrzygniętych problemów w szachach, takich jak ocena długofalowej wartości figury, identyfikacja optymalnych systemów debiutowych czy badanie siły historycznych strategii w świetle nowoczesnych modeli SI.

Modele o nadludzkiej skuteczności nie tylko przewyższają ludzkich arcymistrzów, ale także pozwalają na odkrywanie nowych strategii i fundamentalnych prawidłowości w danej dziedzinie. W przypadku szachów AlphaZero stało się nie tylko najpotężniejszym graczem w historii, ale także narzędziem umożliwiającym analizę i rozwój teorii, której ludzie nie byli w stanie samodzielnie wypracować przez ponad sto lat.

Podobne podejście można rozszerzyć na inne dziedziny, w których modele SI już teraz wykazują zdolności przewyższające ludzi – od biologii molekularnej, przez optymalizację procesów przemysłowych, aż po analizę finansową czy diagnostykę medyczną. Modelologia, czyli nauka o eksploracji modeli, pozwala nam nie tylko lepiej rozumieć te systemy, ale także wykorzystać je do rozwiązywania problemów, które jeszcze niedawno wydawały się niemożliwe do rozwiązania.

10. Modele SI jako narzędzia wspomagające odkrycia naukowe

Podobna analiza modeli, jak ta przeprowadzona dla AlphaZero, znajduje zastosowanie w wielu innych dziedzinach, w których syntetyczna inteligencja może asystować człowiekowi. Jednym z niedawno zaproponowanych modeli SI jest AlphaGeometry, model, który osiągnął poziom olimpijskiego mistrza w rozwiązywaniu zadań matematycznych na poziomie Międzynarodowej Olimpiady Matematycznej. System ten nie tylko poprawnie rozwiązuje skomplikowane problemy geometryczne, ale także generuje dowody matematyczne w sposób podobny do ludzkich ekspertów, dostarczając nowych narzędzi do eksploracji matematycznej.

Innym bardzo obiecującym zastosowaniem SI w nauce jest poszukiwanie nowych materiałów fizycznych o pożądanych właściwościach fizykochemicznych. W pracy *Scaling deep learning for materials discovery*[11] zaprezentowano system SI, który nie tylko automatycznie generuje nowe propozycje składu materiałów, ale również potrafi przewidywać ich właściwości i oceniać ich potencjalne zastosowania. Co więcej, część z tych nowo zaproponowanych materiałów udało się automatycznie zweryfikować eksperymentalnie, co potwierdza skuteczność podejścia opartego na systemach SI.

Analiza uzyskanych wyników ujawniła jeszcze jeden fascynujący aspekt – modele syntetycznej inteligencji nie tylko proponowały znacznie bardziej złożone materiały niż te dotychczas znane, ale także wskazywały na luki w istniejących bazach danych, sugerując obszary, które mogły zostać pominięte w tradycyjnych

[11] Scaling deep learning for materials discovery. Nature 2023. https://www.nature.com/articles/s41586-023-06735-9

badaniach. Okazało się, że SI generowała struktury o większej liczbie pierwiastków i bardziej skomplikowanej architekturze, których potencjalne właściwości wcześniej nie były analizowane.

Te przykłady pokazują, że syntetyczna inteligencja nie tylko wspomaga procesy badawcze, ale w wielu przypadkach staje się kluczowym elementem naukowych odkryć, poszerzając nasze rozumienie świata w sposób, który wcześniej był nieosiągalny. Możliwość eksploracji modeli SI, zrozumienia ich decyzji oraz wyodrębniania z nich wartościowych wzorców otwiera nowe horyzonty w nauce – od matematyki i chemii, przez farmakologię, aż po odkrywanie nowych materiałów.

Era modelologii (and. *Model Science*) zapowiada nową rewolucję w nauce – taką, w której algorytmy nie tylko analizują dane, ale również odkrywają fundamentalne prawa rządzące światem, a ludzie budują nowe metody ekstrakcji tych praw z wytrenowanych modeli.

W cieniu rozkwitających AI

dr hab. Agnieszka Lekka-Kowalik, prof. KUL
ORCID: 0000-0002-4834-318X
Katolicki Uniwersytet Lubelski Jana Pawła II

1. AI w procesie decydowania

Termin „sztuczna inteligencja" (AI) odnosimy do artefaktów zdolnych do symulowania ludzkiego myślenia i działania[1]. W dalszych analizach pomijam kwestię symulacji: czy ma być ona jedynie wiernym naśladowaniem (chodzi o tzw. zombie AI) czy też naśladowanie wymaga jakiegoś rodzaju podmiotowości. Nie analizuję też kontrowersji wokół pojęcia inteligencji: czy jest ona zdolnością rozwiązywania problemów komputacyjną czy też nie; i czy jest jedna inteligencja czy wiele jej typów. Niezależnie od rozwiązywania wskazanych problemów teoretycznych, nie ulega wątpliwości, że współcześnie AI działają w dużym stopniu autonomicznie i stają się coraz bardziej pełnoprawnymi uczestnikami życia społecznego jako tzw. roboty społeczne. Nietrudno znaleźć przykłady. Humanoidalne roboty zaczynają sprawować opiekę nad osobami starszymi (Andtfolk i in. 2022), a tak zwane roboty

[1] Istnieje cały szereg definicji sztucznej inteligencji i wszystkie one mają charakter definicji regulujących, wypracowanych w konkretnych kontekstach. Np. OECD (Organizacja Współpracy Gospodarczej i Rozwoju) definiuje system AI jako „*a machine-based system that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments.*"; https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449. Dostęp 5.01.2025. Z kolei w Unii Europejskiej na potrzeby „Rozporządzenia Parlamentu Europejskiego i Rady (UE) 2024//1689 z dn. 13 czerwca 2024" zdefiniowano system AI jako „system maszynowy, który został zaprojektowany do działania z różnym poziomem autonomii po jego wdrożeniu oraz który może wykazywać zdolność adaptacji po jego wdrożeniu, a także który – na potrzeby wyraźnych lub dorozumianych celów – wnioskuje, jak generować na podstawie otrzymanych danych wejściowych wyniki, takie jak predykcje, treści, zalecenia lub decyzje, które mogą wpływać na środowisko fizyczne lub wirtualne", https://eur-lex.europa.eu/legal-content/PL/TXT/? uri=CELEX: 32024R1689 art 3. Dostęp 5.01.2025.

empatyczne stają się towarzyszami dzieci (Leite, Castellano, Pereira 2014). Jednym z najsłynniejszym robotów jest Sophia (zwana fembotem ze względu na podobieństwo do kobiety), która w październiku 2017 r. otrzymała obywatelstwo Arabii Saudyjskiej, występowała na forum Organizacji Narodów Zjednoczonych, a nawet deklarowała, że szuka męża[2]. Stanowi to dla ludzi autentyczne wyzwanie, tak teoretyczne, jak i praktyczne. Rozważane są m.in. metafizyczny status sztucznej inteligencji, możliwość zaistnienia jej świadomości czy odpowiedzialność za podejmowane przez nią czynności. Zafascynowani możliwościami maszyn – choć i nieco nimi przerażeni – chcemy natomiast zagwarantować, że AI zachowuje się moralnie, tj. w każdych okolicznościach „podejmuje" decyzje o działaniu zgodnym z naszymi wartościami. Spodziewamy się wtedy rozlicznych korzyści z funkcjonowania AI w przestrzeni indywidualnej i społecznej, co wprost postulują zasady z Asilomar, sformułowane podczas konferencji w 2017 r. poświęconej rozwojowi sztucznej inteligencji[3]. Zasada (1) głosi, iż rozwój powinien zmierzać do wykreowania nie po prostu sztucznej inteligencji, ale inteligencji dobroczynnej, życzliwej i przyjaznej ludziom (*beneficial*); zasady (10) i (11) postulują kompatybilność działań AI z wartościami ludzkimi, w tym z ideałami godności, wolności i praw człowieka oraz kulturową różnorodnością; a zasada (23) wymaga, by AI rozwijać w służbie dobra wspólnego ludzkości.

Kluczowym momentem dla rozważań o moralności działań AI jest udział sztucznej inteligencji w podejmowaniu decyzji. Wyróżnia się cztery role, które AI może odgrywać w tym procesie: (a) gromadzenie danych, a więc dostarczanie informacji dla podjęcia decyzji przez ludzi; (b) analizowanie danych wraz z predykcjami, a więc tworzenie scenariuszy „jeżeli... to", które dostarczają podmiotowi ludzkiemu przewidywań co do możliwych konsekwencji rozmaitych decyzji (tzw. *assisted AI*); (c) rekomendowanie (na bazie danych i predykcji) pewnej decyzji jako najlepszej w danej sytuacji, przy czym decyzja pozostaje w rękach człowieka; (d) podejmowanie decyzji i wykonanie działania przez autonomiczną AI. To tu właśnie rodzi się problem robotów społecznych, które Kate Darling z Massachusetts Institute of Technology definiuje jako materialne, autonomiczne podmioty działające, komunikujące się i wchodzące w interakcje z człowiekiem na poziomie emocjonalnym (Darling 2016, s. 215). Darling zresztą używa terminu „aktant", by zaznaczyć ich moc działania, a odróżnić od ludzi. Owe aktanty są czymś więcej niż roboty

[2] *Sophia ma obywatelstwo Arabii Saudyjskiej i szuka męża. Jest robotem.* Dobreprogramy, https://www.dobreprogramy.pl/sophia-ma-obywatelstwo-arabii-saudyjskiej-i-szuka-meza-jest-robotem, 6820095894465376a. Dostęp 5.01.2025.

[3] Zob. Asilomar AI Principles, https://futureoflife.org/open-letter/ai-principles/. Dostęp 2.01.2025.

autonomiczne, realizujące zadania niezależnie i w znacznych odległościach od człowieka. To roboty, które mają wchodzić w interakcje i współpracować z człowiekiem jako jego partnerzy, a być może nawet zastępować ludzi jako naturalnych partnerów komunikacji.

Na każdym z poziomów uczestniczenia AI w decydowaniu pojawiają się kwestie etyczne. Najczęściej wymienia się: brak transparentności działania AI, uprzedzenia i stereotypy wcielone w algorytmy, ustalanie odpowiedzialności przed społeczeństwem, zagrożenie dla prywatności, bezstronność i sprawiedliwość w działaniu (Cf. Prem 2023). Istnieje ogromna literatura na temat tych problemów, ale tu interesuje nas następująca kwestia: podejmowanie decyzji jest sprawą bardzo ludzką, a wobec tego, jakie będą konsekwencje oddania decyzji „w ręce" AI, niezależnie od rozwiązania wymienionych wyżej kwestii etycznych? By odpowiedzieć na powyższe pytanie, należy przyjrzeć się dokładniej naturze ludzkiej decyzji.

2. Natura ludzkiej decyzji

Poniższe rozumienie ludzkiej decyzji odwołuje się do rozważań nad decyzją prowadzonych w ramach lubelskiej szkoły filozofii klasycznej. Decyzja widziana jest jako rezultat współpracy rozumu i woli – to sąd autodeterminujący człowieka do działania (Krąpiec 1979; Krąpiec 2001). Decyzja nakierowana ma być na realizację konkretnego dobra i ma być słuszna, tj. respektować godność osoby ludzkiej – i to zarówno człowieka działającego, jak i człowieka, na którego działanie jest skierowane – oraz inne wartości. Jest to zresztą warunek narzucany także na decyzje AI, jak można zauważyć w cytowanych wyżej zasadach z Asilomar. Zasadniczy problem rysuje się jednakże następująco: jakie działanie wykonane przeze mnie tu i teraz faktycznie respektuje owe wartości? Decyzja jest wobec tego podwójnym sądem. Najpierw wydaję sąd o charakterze poznawczym (zwanym teoretyczno-praktycznym): „teraz w tych oto okolicznościach ja *powinnam* zrobić to a to". Sąd ten nie jest rezultatem wnioskowania dedukcyjnego z jakichś ogólnych zasad (choćby tej: czyń dobro, zła unikaj), ale jest twórczy, niejako syntetyzuje wiedzę i doświadczenie, i w tym sensie jest niepowtarzalny. Ostatecznie – w świetle mojego sądu „powinnam" – wydaję sąd (zwany praktyczno-praktycznym): „zrobię to a to". Sąd ten ustanawia mnie źródłem działania. A stając się źródłem działania, rozpoczynam w świecie (a raczej: wszechświecie, bo przecież rzeczywistość jest w skali makroskopowej ciągła) nowy łańcuch przyczynowo-skutkowy. Działanie zaś ma dwojakie konsekwencje. Podjęcie decyzji – nawet jeśli nie zdołam jej wykonać – buduje mnie jako osobę o pewnej osobowości. Gdy podejmuję decyzje o morderstwie, to nawet jeśli ktoś mi przeszkodzi w dokonaniu morderstwa, buduję siebie jako mordercę. Nie musi to być zresztą taki dramatyczny przykład.

Oto decydując się zjeść raczej gruszkę niż jabłko, buduję siebie jako zjadaczkę gruszek. Gdy decyzję zrealizuję, w reszcie świata pojawią się konsekwencje: ktoś zostanie pozbawiony życia, a pewna gruszka zjedzona. Na tym się owe konsekwencje nie kończą, ponieważ rodzą się całe ciągi skutków: ponieważ zamordowano kogoś, odbył się pogrzeb, na którym spotkali się krewni, odnowili kontakty, a ktoś odziedziczył majątek zamordowanego i wobec tego nareszcie mógł założyć firmę, o której marzył, inni krewni obrazili się za pominięcie w testamencie… Można bez trudu wyobrazić sobie następstwa rodzące się w świecie z powodu zjedzenia przez kogoś gruszki. Krótko mówiąc, każda ludzka decyzja ma zawsze konsekwencje dla samego działającego, a gdy wykonana – także dla wszechświata. W tym sensie jestem odpowiedzialna za siebie i za innych (nie tylko ludzi), których moje działanie dotyka. Z racji bycia źródłem działania można człowiekowi przypisać odpowiedzialność, przy czym termin „odpowiedzialny za" ma dwa sensy: jeden odnosi się do odpowiedzialności *ex post* za własne działanie i jego konsekwencje, natomiast drugi do odpowiedzialności *ex ante*, tj. do odpowiedzialności za byty, których wartość uznajemy, a których istnienie i rozwój zależy od działania podmiotu, co rodzi zobowiązanie do działania na rzecz dobra tych bytów-wartości. Podmiot działający jest wobec tego zawsze w normo-twórczej sieci relacji.

Ten spójny obraz aktu decydowania zakłóca ludzka wolność. Oto mogę odrzucić swój własny sąd „powinnam zrobić to a to" i wydać sąd „zrobię coś innego" – i to ten sąd będzie mnie determinował do działania. Wiem, co powinnam zrobić – i robię coś przeciwnego. Jest to doświadczenie znane zapewne każdemu człowiekowi. Mogę za to żałować; a mogę też *post factum* rozpoznać niesłuszność mojej decyzji i przeanalizowawszy ją, posiąść nową wiedzę, a dzięki temu uniknąć błędnych decyzji w przyszłości; w obu przypadkach mogę też mieć wyrzuty sumienia. Oczywiście i przy błędnych decyzjach rodzą się wewnętrzne i zewnętrzne konsekwencje. Sytuacja staje się jeszcze bardziej skomplikowana, gdy uświadomimy sobie istnienie dylematów moralnych. Dylemat moralny to sytuacja konfliktu między dwoma lub więcej wartościami/obowiązkami/normami, które nie dadzą się naraz zrealizować[4]. Znajdujemy się wtedy w sytuacji niepewności, jaka decyzja jest słuszna – a działać musimy, ponieważ „nicnierobienie" jest też decyzją i czasem nawet można przewidzieć, jak się rozwinie bieg wydarzeń bez naszej interwencji. Często też musimy komuś – zwłaszcza temu, kogo nasze działanie bezpośrednio dotknęło – przedstawić racje czy też uzasadnienie naszej decyzji i oczywiście – jak w przypadku każdej decyzji – podjąć odpowiedzialność za konsekwencje. Zaapli-

[4] Ważne rozważania nad naturą dylematów i konfliktów moralnych zawiera monografia pod redakcją Piotra Duchlińskiego (Duchliński 2023).

kujmy teraz to rozumienie ludzkiej decyzji do sztucznej inteligencji wyposażonej w zdolność determinowania się do działania i wykonania ustalonego działania.

3. Konsekwencje oddania decyzji „w ręce" AI

Przeprowadźmy pewien eksperyment myślowy. Oto AI pracuje jako dystrybutor respiratorów. I nagle rodzi się sytuacja dylematogenna: jest dwóch pacjentów i jeden respirator. Pojawia się więc swoisty „konflikt instrukcji": „przydziel pacjentowi X, bo potrzebuje" i „przydziel pacjentowi Y, bo potrzebuje". Komu przydzielić? Decyzja musi być podjęta, i to podjęta szybko, bo inaczej pacjenci umrą. AI musi wobec tego wydać sąd „powinno się przydzielić temu a temu", a następnie sąd „przydzielę temu właśnie" – i tę decyzję wykonać, podłączając respirator wybranej osobie. Ten sąd musi uwzględniać fakty (dane) i wartości, bo narzucamy na AI wymóg zgodności jej „decyzji" z naszymi ideałami – jak wyżej cytowałam – godności, wolności i praw osoby oraz kulturowej różnorodności. Co więcej, AI musi być zdolna do przedstawienia uzasadnienia swojej decyzji, choćby dlatego, że rodzina pacjenta, który umarł z braku respiratora, ma prawo zapytać, dlaczego akurat on sprzętu nie dostał. Pomińmy problem, jakie uzasadnienie byłoby w takiej sytuacji adekwatne. Zauważmy natomiast, że AI musi być wyposażona nie tylko w umiejętność rejestracji faktów, znajdowania korelacji i odróżniania ich od związków przyczynowo-skutkowych, ale też w umiejętność wydawania sądów wartościujących o słuszności i niesłuszności danego działania, uzasadniania tych sądów, a więc obrony słuszności swej decyzji, co wymaga oceny decyzji pod względem moralnym. U człowieka tego typu zdolności zwykło się nazywać sumieniem. Wydaje się więc, że jeśli AI ma symulować ludzkie myślenie i zachowanie, to należy ją wyposażyć w sztuczne sumienie – zdolność sądzenia o własnych (i cudzych) czynach w aspekcie dobra i zła moralnego.

W takim kierunku idą badania w ramach transdyscyplinarnego obszaru tzw. etyki maszyn (*machine ethics)*. Za Jamesem Moorem (2006) odróżnia się domniemanego agenta moralnego (*implicit moral agent*) i jawnie moralnego agenta (*explicit moral agent*). Termin „agent" jest tu użyty celowo, by możliwy tu do wykorzystania termin „podmiot" czy „podmiot działający" nie sugerował, iż etyczne maszyny mają jakiś rodzaj podmiotowości. AI jest domniemanym agentem etycznym, gdy w algorytm są wprowadzone ograniczenia nakazujące etyczne zachowania lub zakazujące zachowań nieetycznych. Przykładem jest zakaz używania pewnych wyrażeń wprowadzony w ChatGPT, ale też „uczciwe" obliczanie raty kredytu. AI jako jawnie moralny agent miałby w swym algorytmie reprezentację etyki i działałby na podstawie tej wiedzy, rozwiązując m.in. dylematy moralne. Jawnie moralny agent operowałby więc zarazem na przypadkach i na zasadach,

które są niezbędne, by przewidzieć, jak zachowa się agent w nowych sytuacjach, by możliwe było wydobycie moralnie istotnych różnic między przypadkami, by dało się uzasadniać decyzje oraz by następowały modyfikacje instrukcji pod wpływem „doświadczenia" (Anderson i Anderson 2007). Próby zbudowania jawnie moralnego agenta były i są podejmowane, a potrzeba stworzenia takiego agenta wielorako uzasadniana (Anderson, Anderson i Armen 2004; Allen, Smit i Wallach 2005;Wallach i Allen 2009; Anderson i Anderson 2011; Misselhorn 2018).

Jednakże nawet pobieżne przyjrzenie się tym próbom sygnalizuje poważny problem praktyczny. Naczelne zasady moralne (i wszelkie ich pochodne) są różne w różnych teoriach etycznych i wobec tego może zaistnieć różnica decyzji, co zrobić w konkretnej sytuacji (i jak rozwiązać dany dylemat moralny) w zależności od tego, kto zaprogramował konkretną AI, na bazie jakich doświadczeń uczyła się ona podejmować decyzje i jaka jest naczelna norma włączona w algorytm. Czy wobec tego przed zakupem „pielęgniarki" czy „opiekunki" dla staruszki powinniśmy zapytać o moralne poglądy informatyków programujących daną linię AI? Czy może też instrukcja użytkowania powinna zawierać szczegółową „deklarację moralną"? Czy przed pójściem na operację powinniśmy zapytać o „poglądy moralne" AI działającej w szpitalu? Ostatecznie chodzi o odpowiedź na pytanie, co jest uznane za dobro człowieka w danej teorii etycznej, a więc i o rozumienie człowieka. Nie są to kwestie banalne, zważywszy na coraz poważniejszy udział AI w decyzjach podejmowanych w życiu indywidualnym i społecznym.

Jeszcze inny problem pojawia się, gdy – zakładając, że uda się wyposażyć AI w sztuczne sumienie *(artificial conscience*) – przyjrzymy się, jak AI podejmuje decyzje. Jak zauważyłam wyżej, człowiek nieraz wie, jakie działanie jest słuszne, ale czyni coś innego. Wyjaśnia się to rozmaicie: grzechem pierworodnym, słabością woli, ewolucyjnym mechanizmem przetrwania... Wydaje się, że w przypadku AI taka sytuacja nie ma miejsca: sąd: „powinno się zrobić to a to" jest zawsze tożsamy z sądem-decyzją. Mówiąc bardziej filozoficznie, jest to pewna postać intelektualizmu etycznego: AI „poznaje" (ustala), co jest dobre w danej sytuacji (co należy zrobić), i zawsze idzie za poznaną „prawdą o dobru". W tym sensie jest bytem od nas moralnie lepszym – bo nie czyni intencjonalnie zła, nie odrzuca swego własnego sądu „powinien". A skoro AI wydaje słuszną decyzję i działa zgodnie z tą decyzją – zgodnie z klasyczną filozofią należałoby powiedzieć, że AI jest mędrcem – sztucznym mędrcem. A mędrcy są dla nas autorytetami. Co więcej, AI jest szybsza i lepsza w zbieraniu danych, analizowaniu i klasyfikowaniu faktów, odkrywaniu korelacji i alternatyw działania. Niejako „wiedząc" więcej, może stawać w obliczu sytuacji czy dylematów moralnych, z którymi ludzie nigdy nie mieli do czynienia. W takim przypadku ludzkie doświadczenie nie wystarczyłoby do „moralnego wychowania" AI. Czy powinniśmy zatem studiować moralne

decyzje AI i uczyć się od niej? Jest oczywiście postulat „wyjaśnialnej" AI (*explicable AII*), tj. takiego budowania AI, by jej operacje i ich rezultaty były jak najbardziej zrozumiałe (transparentne) dla ludzi. Można by wtedy powiedzieć, że mając dostęp do zasad i danych, które były podstawą decyzji AI w określonym przypadku np. dylematu moralnego, możemy samodzielnie, po ludzku, rozwiązać ów dylemat, a nawet skrytykować rozwiązanie AI. W zasadzie byłoby to pewnie możliwe, przy czym kluczowy jest tu zwrot „w zasadzie", bowiem już nawet prześledzenie owych tysięcy, jeśli nie milionów, przypadków, na których operowała AI, wymaga tyle czasu, że staje się w praktyce niemożliwe, a co najmniej nieopłacalne, jeśli życie ma toczyć się dalej. Już polegamy na decyzjach AI i sądzę, że proces ten będzie się intensyfikował. Wzrasta liczba relacji człowiek – AI, a to przekształca także relacje międzyosobowe[5].

Rozwój systemów AI jako autorytetów i partnerów człowieka pociąga za sobą jeszcze jedno niebezpieczeństwo. Sue Anne Teo (2024) twierdzi, że mamy tu do czynienia z powolną przemocą (*slow violence*), która może w niezauważalny sposób podważyć fundamentalne założenia praw człowieka, a nawet ich normatywne uzasadnienie, a przez to i status człowieka. Prawa przysługują jednostce niepowtarzalnej, indywidualnej osobie ze względu na jej sposób istnienia (racjonalność, wolność, podmiotowość). Prawa też upodmiotowiają jednostkę wobec państwa, prawa, organizacji. Tymczasem dla korzyści płynących z wykorzystania AI przekształcamy wszystko i wszystkich na dane (*datafication*), a wtedy – uzupełnijmy tu myśl Teo – jestem coraz mniej postrzegana jako niepowtarzalna osoba, a coraz bardziej jako element sporządzonych przez AI klas. Jeśli zaś w algorytm wbudowany jest fałszywy stereotyp, a decyzje AI uznawane za niepodważalnie moralne, to podważona zostaje idea równości osób. Jednostka skonfrontowana z decyzjami AI ma coraz mniejsze możliwości kontestowania tych decyzji, ponieważ nie jest w stanie śledzić racji za nimi stojących – i w tym sensie AI wcale nie wspomaga wolności jednostki, a raczej jej zakres ogranicza. Co więcej, skoro AI nie ulega złudzeniom, wpływom władzy, indywidualnym uprzedzeniom, niespójnościom decyzyjnym, to traktujemy jej decyzje jako „lepsze" w sensie i epistemicznym, i moralnym – tym samym uzyskuje uprzywilejowaną pozycję w społecznym świecie. Teo podnosi jeszcze jedną kwestię: AI (dzięki *datafication* i klasyfikacjom) dostarcza nam widzenia świata i siebie. Uważamy bowiem, że „wglądy AI w strukturę rzeczywistości" są rzetelnymi i godnymi zaufania wskaźnikami natury człowieka i zjawisk społecznych jako obiektywne i oparte na wielkiej ilości danych.

[5] Artykuły na ten temat zawierają dwa tematyczne tomy „Ethosu" z 2023: „Oswoić sztuczną inteligencję" (36, nr 3) oraz „Sztuczna inteligencja w świecie wartości ludzkich" (36, nr 4).

Tym samym rozumienie siebie – a biorąc pod uwagę, że AI rekomenduje nam działania – także zarządzanie sobą i swoim życiem są sterowane przez AI. Gdy spojrzymy głębiej na personalizację informacji, profilowanie, przypisywanie mnie do określonej grupy, to okaże się, że to AI decyduje, jaki jest mój świat, podważając moją autonomię; a ja coraz mniej rozumiem ten świat i siebie w nim (pojawia się zjawisko zwane w literaturze „hermeneutyczną niesprawiedliwością").

4. Kilka wniosków

Nie ulega wątpliwości, że zakres funkcjonowania AI – ale też zakres jej możliwości – radykalnie rośnie. Ewoluuje też nasz stosunek do niej. Systemy AI, a zwłaszcza humanoidalne roboty, przestają być uznawane jedynie za narzędzie. Badacze pracujący w nowej dziedzinie HRI (*human-robot interaction*) twierdzą, że nasze emocjonalne podejście do robotów radykalnie się zmienia: ufamy ich rekomendacjom, ujawniamy im wrażliwe dane, przywiązujemy się do nich, a nawet traktujemy jako życiowych partnerów[6]. Przestajemy się dziwić, gdy docierają do nas informacje o badaniach nad psychologicznym kontraktem między człowiekiem a robotem (Rogozińska-Pawełczyk 2020), o pojawieniu się psychiatrii robotów (Żółcińska 2020)[7] czy też o przyznawaniu praw robotom i nadawaniu im osobowości prawnej (Biczysko-Pudełko i Szostek 2019). Obdarzamy AI epistemicznym zaufaniem – a zarazem człowiek jako niepowtarzalna osoba znika nam z widoku razem z jej godnością. Mamy bowiem – co starałam się wyżej pokazać – do czynienia z umniejszaniem przez AI wolności, decyzyjności i autonomii osoby, a przez to niejako stajemy się mniej podmiotami, czy raczej mniej jesteśmy w stanie egzekwować swą podmiotowość. I coraz mniej uważamy to za istotne. Nikniemy w cieniu rozkwitających AI. Mówiąc jeszcze inaczej, AI powoli eliminuje człowieka z centralnego miejsca w strukturze rzeczywistości. Nie chodzi tu bynajmniej o „bunt maszyn", ale o konsekwencje prawidłowego funkcjonowania tej technologii. Nie jest to dopiero problem mocnej sztucznej inteligencji – ale wszechobecnych AI opracowanych dla konkretnych zadań. Gdy zaś człowiek przestaje zajmować

[6] Aplikacja Soulmate pozwalała kreować wirtualnych partnerów, z którymi osoby nawiązywały romantyczne relacje. Gdy aplikacja została zamknięta, ludzie tygodniami płakali, popadali w depresję, powstawały grupy wsparcia dla przeżywania żałoby. Zob. https://businessinsider.com.pl/wiadomosci/cyfrowa-smierc-w-xxi-w-ludzie-oplakuja-wirtualnych-kochankow/rg6z5h6. Dostęp 4.02.2025.

[7] Pierwszą w świecie psychiatrą robotów jest Joanne Pransky. Zob. wywiad z nią: https://cyfrowaekonomia.pl/infoshare2020-sane-robots-and-insane-humans-wywiad-z-joanne-pransky-psychiatra-robotycznym/. Dostęp 5.02.2025.

centralną pozycję, to nie inne byty (czy nawet cała Ziemia) wchodzą w to miejsce, ale postęp. Człowiek jest bytem, który kreuje, testuje, obsługuje, serwisuje AI – „Human-as a service" nie wydaje się być nosicielem nienaruszalnej godności, gdyż jest widziany przez pryzmat funkcji, jakie może spełniać w niekończącym się postępie. A może ostatecznie sztuczna inteligencja się „usamodzielni", uznając nas za niższy, a być może nawet zbędny i szkodliwy gatunek? Taką autonomiczną technoewolucję przepowiadał już w latach sześćdziesiątych ubiegłego wieku Stanisław Lem w książce *Summa technologiae.*

AI już z nami pozostanie. Wszelkie wezwania do zaprzestania jej rozwijania – tak jak to uczynili Elon Musk, Yuval Noah Harari, Steve Wozniak, Jaan Tallinn i wielu innych badaczy z obszaru AI w liście otwartym[8] – wydają mi się próżne. Nie jest natomiast jałowa refleksja nad nią – tak długo, jak długo jeszcze mamy nadzieję, że nie jest jeszcze za późno na świadome sterowanie kierunkiem prac nad sztuczną inteligencją.

Bibliografia

[1] Anderson, M., Anderson S.L., Armen, C. (2004). Towards Machine Ethics: Implementing Two Action-Based Ethical Theories. AAAI-04 Workshop on Agent Organizations: Theory and Practice. https://www.researchgate.net/publication/259656154_Towards_Machine_Ethics. Dostęp 10.01.2025.

[2] Anderson, M., Anderson S.L. (2007). Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine* 28 (4), pp. 15–26.

[3] Anderson, M., Anderson S.L. (eds.) (2011). *Machine Ethics.* Cambridge University Press, New York.

[4] Andtfolk, M., Nyholm, L., Eide, H., Fagerström, L. (2022). Humanoid Robots in the Care of Older Persons: A Scoping Review. *Assistive Technology* 34 No. 5, pp. 518–526.

[5] Biczysko-Pudełko, K., Szostek, D. (2019). Koncepcje dotyczące osobowości prawnej robotów – zagadnienia wybrane. *Prawo Mediów Elektronicznych* 2, s. 9–15.

[6] Collin, A., Smit, I, Wallach, W. (2005). Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches. *Ethics and Information Technology* 7, pp. 149–155.

[7] Darling, K. (2016). Extending Legal Protection to Social Robots. In R. Calo, M. Froomkin, I. Kerr (eds.). *Robot Law*, Edward Elgar Publishing, Cheltenham, UK.

[8] Duchliński, P. (red.) (2023). *Spory moralne* (seria: *Słowniki społeczne*). Wydawnictwo Naukowe Uniwersytetu Ignatianum, Kraków.

[9] Future of Life Institute (2017). Asilomar AI Principles. https://futureoflife.org/open-letter/ai-principles/. Dostęp 21.01.2025.

[8] Zob. *Pause Giant AI Experiments: An Open Letter*, https://futureoflife.org/open-letter/pause-giant-ai-experiments/. Dostęp 21.01.2025.

[10] Krąpiec, M.A. (1979). *Ja – człowiek: Zarys antropologii filozoficznej*. Towarzystwo Naukowe KUL, Lublin.
[11] Krąpiec, M.A. (2001). Decyzja. W: *Powszechna Encyklopedia Filozofii*, t. 2, s. 442–444. Polskie Towarzystwo Tomasza z Akwinu, Lublin.
[12] Leite, I., Castellano, G., Pereira, A. *et al.* (2014). Empathic Robots for Long-term Interaction. *International Journal of Social Robotics* 6 No. 3, pp. 329–341.
[13] Lem, S. (1964). *Summa technologiae.* Wydawnictwo Literackie, Kraków.
[14] Misselhorn, C. (2018). Artificial Morality. Concepts, Issues and Challenges. *Social Science and Public Policy* 55, pp. 161–169.
[15] Moore, J. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems* 21, pp. 18–21.
[16] Prem, E. (2023), From Ethical AI Framework to Tools: a Review of Approaches. *AI and Ethics* 3, pp. 699–716.
[17] Rogozińska-Pawełczyk, A. (2020). Towards Discovering Employee-Robot Interaction: Aspects of Concluding the Psychological Contract. *Education of Economists and Managers* 58 No. 4, pp. 9–20.
[18] Teo, S.A. (2024). Artificial Intelligence and Its "Slow Violence" to Human Rights. *AI and Ethics* 19 August. https://www.researchgate.net/publication/383235625_Artificial_intelligence_and_its_'slow_violence'_to_human_rights. Dostęp 1.02.2025.
[19] Wallach W., Allen C. (2009). *Moral Machines: Teaching Robots Right from Wrong.* Oxford University Press, New York.
[20] Żółcińska, W. (2020). Psychiatra robotów. *Computerworld* 27.08.2020, https://www.computerworld.pl/wywiad/Psychiatra-robotow, 422594.html. Dostęp 1.02.2025.

Czy sztuczna inteligencja może być odpowiedzialna albo godna zaufania?

dr Natalia Juchniewicz
ORCID: 0000-0002-2686-9404
Uniwersytet Warszawski

W dyskursie wokół sztucznej inteligencji (*artificial intelligence* – AI), gdy dyskutowane są warunki wdrażania tej technologii do życia społecznego, ekonomii czy polityki, pojawia się problem zaufania ludzi do samej technologii oraz zakresu odpowiedzialności, jaki możemy jej przypisywać bądź na nią delegować. W szerszym sensie wiąże się to bowiem z problemami ontologicznymi (jakim rodzajem bytu jest sztuczna inteligencja?; czy sztuczna inteligencja to aktor społeczny czy po prostu narzędzie? itp.) oraz wynikającymi z nich problemami moralnymi (czy sztuczna inteligencja jest autonomiczna?; czy sztuczna inteligencja podlega prawu? itp.; kto kontrolować powinien działanie AI?). Sztuczna inteligencja, jak żadna wcześniejsza technologia, potrafi bowiem doskonale imitować człowieka: wygrywa z nim w szachy, może prowadzić długie konwersacje o sztuce, może wspierać emocjonalnie osoby samotne, a przy tym zanalizuje duże zbiory danych, sprofiluje użytkowników określonej aplikacji czy też będzie rekomendować określone osoby do uzyskania kredytu. Spektrum zastosowań AI jest niemal nieograniczone. Massimo Airoldi przekonuje, że sztuczną inteligencję definiować można jako technologię opartą na algorytmie wyznaczonym przez: po pierwsze, procedurę matematyczną (co dotyczy zresztą każdej formy maszyny), pod drugie, model podążania za określonymi komendami opartymi na dedukcji (*Good Old Fashioned Artificial Intelligence*; np. algorytmy wyszukiwarek internetowych); po trzecie, uczenie maszynowe oparte na autonomicznym przetwarzaniu danych i indukcyjnym uczeniu się nadzorowane albo nienadzorowane przez ludzi (np. Page Rank, Alpha Go). Airoldi pokazuje, jak wraz z rozwojem technologicznym zmieniał się sposób definiowania sztucznej inteligencji. Była ona czymś innym w erze analogowej, erze cyfrowej i wreszcie w obecnej erze platform (Airoldi 2022: 14). Dla celów mojego artykułu przyjmuję

w szczególności trzecią definicję sztucznej inteligencji, gdyż to wokół niej narastają problemy ontologiczne, epistemologiczne i moralne. Uczenie maszynowe, sieci neuronowe, pewien zakres niejasności w procesie uczenia się AI (Burrell: 2016), jej zdolności analizy wielkich zbiorów danych, obecność algorytmów w codziennych praktykach, to wszystko sprawia, że nawet jeśli nie chcemy mieć ze sztuczną inteligencją świadomie do czynienia, to i tak często jest ona gruntem dla wielu naszych aktywności i towarzyszy nam oraz asystuje w wielu czynnościach. W ramach problemów podejmowanych w niniejszym artykule można uznać za sztuczną inteligencję wszystkie formy technologii uczących się opartych na algorytmie od algorytmów typu Page Rank po roboty społeczne. W zależności od kontekstu używać będę zamiennie słów „sztuczna inteligencja", „algorytm", „AI" czy „robot". Sama forma tej technologii jest bowiem wtórna wobec stawianych w artykule problemów wokół „moralnej AI".

Jako maszynę możemy chcieć zaprojektować sztuczną inteligencję, by służyła określonym, instrumentalnym celom. To jednakże, co AI „robi" jako aktor czy aktant społeczny (Latour 2013), wymyka się liniowemu projektowaniu jej oddziaływania. Jednym z nieoczywistych problemów związanych ze sztuczną inteligencją, choć bezpośrednio powiązanym z delegowaniem na nią działania oraz zaufaniem do niej, jest jej wymiar moralny. Samo powiązanie technologii z moralnością wymagało od filozofii i etyki przesunięcia w kierunku rzeczy, skupienia uwagi na ontologii przedmiotów i ich możliwym sprawstwie (*agency*). Skutkiem owego przesunięcia wyłoniły się dwie dominujące w literaturze filozoficznej narracje wokół moralnej sztucznej inteligencji. Pierwsza skupia się na konsekwencjach moralnych wdrażania sztucznej inteligencji w różne obszary życia społecznego, ekonomicznego czy politycznego, którą nazywam w skrócie narracją konsekwencjalistyczną. Druga narracja dąży do przedstawienia sztucznej inteligencji jako nowej formy aktora czy agenta społecznego, a nawet jako nowy rodzaj sztucznej osoby, co w skrócie nazywam narracją personalistyczną. W niniejszym artykule postaram się scharakteryzować i poddać krytyce obie te narracje w celu ukazania, że choć dotyczą one sztucznej inteligencji, w rzeczywistości postulują odpowiedzialność ludzi za tę technologię i ewentualne zaufanie bądź jego brak względem celów, jakie ludzie we wdrażaniu AI stawiają. Sztuczna inteligencja godna zaufania nie jest zatem dosłownie postulatem adresowanym wobec maszyny, a raczej pytaniem o to, w świecie jakich wartości chcemy funkcjonować jako ludzie. Fakt, że w badaniach nad moralną sztuczną inteligencją wcale nie chodzi o postulowanie moralności samych rzeczy, a o refleksję nad wartościami i stosunkiem ludzi do bytów nie-ludzkich, sprawia, że etyka sztucznej inteligencji okazuje się przede wszystkim dziedziną badań nad relacjami, jakie budować może człowiek z Innym, nie tylko technologicznej natury.

Nazywam wyszczególnione przeze mnie podejścia do moralnej sztucznej inteligencji „narracjami", po pierwsze dlatego, że niniejszy artykuł nie opiera się na pogłębionej analizie dyskursu, a na propozycji uporządkowania pewnych argumentów filozoficznych w określone grupy na zasadzie przyjęcia określonej ontologii AI. Celem mojego artykułu jest zatem wskazanie pewnych tendencji w budowaniu argumentów wokół AI, a nie ich klasyfikacja czy szczegółowa analiza. Chodzi mi o pokazanie, że odpowiedź na pytanie, czy możemy zaufać sztucznej inteligencji, sprowadzić można do określenia, czy stoimy po jednej czy po drugiej stronie dyskusji na temat ontologii AI – czy jest ona dla nas przedmiotem czy formą podmiotu. Moim zdaniem propozycja wyszczególnienia takich narracji ma charakter porządkujący dla szczegółowej pracy nad określonymi argumentami. Po drugie, pojęcie narracji jest bardzo szerokie, pozwala uchwycić zarówno konkretne, naukowe argumenty, jak i opowieści oraz wyobrażenia, jakie wokół sztucznej inteligencji pojawiają się w różnych obszarach naszej z nią interakcji. Nie jest zatem tak, że wyszczególnione przeze mnie narracje są zamknięte albo jednoznacznie zdefiniowane. W ich obrębie istnieją wewnętrzne spory co do tego, czym AI jest i jak należy się do niej odnosić. Po trzecie, narracje pozwalają przecinać podejścia w ramach samej etyki, która w niektórych ujęciach ma charakter opisowy, a w innych normatywny. Problem w tym, że nie zawsze jest jasne, czy np. z opisu określonych praktyk zdaniem autorów wynikają jakieś konsekwencje moralne albo czy postulowana cnota jest adresowana tylko do ludzi czy może też maszyn. Narracje, w moim przekonaniu, są bardziej elastycznym pojęciem i pozwalają opisać pewien zbiór problemów, w których mieści się także sama moralna AI.

W pierwszej części artykułu dokonam charakterystyki narracji konsekwencjalistycznej. W jej ramach uwaga skupia się na pozytywnych bądź negatywnych konsekwencjach produkowania, wdrażania oraz używania sztucznej inteligencji. Celem badaczy zorientowanych na konsekwencje AI jest postulowanie określonych wartości, które technologia ta powinna spełniać, aby np. nie szkodzić (przede wszystkim ludziom, ale i przyrodzie). Co istotne, pochodzenie i rozumienie tych wartości nie jest zawsze w tej narracji wyjaśniane. Sztuczna inteligencja rozumiana jest tutaj jako instrument czy narzędzie, na które co najwyżej delegować można określone wartości (Verbeek 2011), zatem w narracji tej w dużo większym stopniu chodzi o moralność i cnotę ludzi niż o moralną AI.

Narracja personalistyczna, którą charakteryzuję w drugiej części artykułu, zakłada z kolei, że podmiotowość, działanie, sprawstwo oraz osobowość mogą zostać rozszerzone na innych aktorów niż ludzie. Narracja ta ujawnia tym samym, że traktowanie sztucznej inteligencji „jak człowieka" chociażby ze względu na jej umiejętność posługiwania się językiem, nie jest błędem kategorialnym opartym

na niezdolności rozpoznania, iż mamy do czynienia z maszyną, a właśnie na przyjęciu, iż nawet jeśli mamy do czynienia z maszyną, to interakcja z nią może być równa jakościowo interakcjom z ludźmi. W ramach tej narracji wyłaniają się dwa dominujące podejścia: tzw. etyka zorientowana na aktora, która zakłada, że AI może być podmiotem moralnym w sensie sprawczości moralnej (*agent-oriented ethics*; Anderson & Anderson 2011; Wallach & Allen 2009) oraz etyka zorientowana na pacjenta czy odbiorcę (*patient-oriented ethics*; Gunkel 2018), która określa, jaki stosunek do AI powinniśmy mieć jako ludzie, bez względu na to, czy ona sama jest podmiotem moralnym czy nie. Problem, jaki stoi przed narracją personalistyczną, to określenie sensu działania moralnego, do jakiego zdolna miałaby być sztuczna inteligencja.

W trzeciej części pracy poddam obie te narracje krytyce. Gdy narracja konsekwencjalistyczna mówi o sztucznej inteligencji godnej zaufania, to nie mówi wcale o sztucznej inteligencji, a o jej pożądanych, społecznych skutkach, którym możemy zaufać. Szyld *trustworthy AI* nie dotyczy zatem samej sztucznej inteligencji jako rzekomego podmiotu moralnego. Z kolei narracja personalistyczna, choć umiejętnie uzasadnia, dlaczego sztuczna inteligencja może być postrzegana jako np. osoba w sensie prawnym i dlaczego niektóre praktyki ludzi wokół AI mogą być uznane za rozszerzenie definicji podmiotowości moralnej, to nie jest w stanie rozwiązać problemu nieposiadania przez sztuczną inteligencję samej etyki jako koncepcji. Innymi słowy, AI nie jest w stanie stworzyć etyki, co gorsza, nie jest w stanie zmienić się samoistnie w jakimś moralnym kierunku, gdyż oba te warunki podmiotu moralnego zakładają z jednej strony teorię umysłu, z drugiej posiadanie woli i zdolności do projektowania siebie. Propozycją wyjścia poza problemy etyki zorientowanej ontologicznie, a więc takiej, która przypisuje jakość moralną bytom posiadającym określone jakości, jest propozycja postawienia etyki przed ontologią (Gunkel 2018).

W podsumowaniu zbieram kluczowe argumenty artykułu oraz jego konkluzje.

I. Narracja konsekwencjalistyczna

Przez narrację konsekwencjalistyczną rozumiem łączenie sztucznej inteligencji z badaniami nad konsekwencjami społecznymi, ekonomicznymi, politycznymi, czy szerzej etycznymi, tworzenia, wdrażania i posługiwania się AI. W języku charakterystycznym dla tej narracji pojawiają się często „implikacje", pożądane i niepożądane „skutki", zamierzone i niezamierzone „efekty", kalkulacja „zysków i strat". Posługiwanie się sztuczną inteligencją służy w tej narracji odpowiedzi na następujące pytania: Kim możemy się stać? – jakie mamy szanse autonomicznej samorealizacji dzięki technologii AI; Co możemy zrobić? – w jakim stopniu sztuczna inte-

ligencja poszerzy bądź zawęży ludzką sprawczość; Co możemy osiągnąć? – jakie technologia AI daje możliwości jednostkowe i społeczne; Jak możemy wchodzić w interakcję ze sobą nawzajem i ze światem? – czy sztuczna inteligencja sprzyja spójności społecznej (Floridi et al., 2018: 690).

Narracja konsekwencjalistyczna najbliższa jest definiowaniu tzw. sztucznej inteligencji godnej zaufania (*trustworthy AI*). Taka definicja zaproponowana została m.in. w dokumencie Komisji Europejskiej *Wytyczne w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji* (KE 2019). Zasady, które spełniać musi AI godna zaufania, są następujące: (i) poszanowanie autonomii człowieka; (ii) zapobieganie szkodom; (iii) sprawiedliwość; (iv) możliwość wyjaśnienia (*explainability*). Sztuczna inteligencja, jakiej możemy zaufać, to w tym ujęciu przede wszystkim technologia, która nie działa na szkodę człowieka i środowiska w całym cyklu swojego życia i działania. Co istotne, w dokumencie tym zakładana jest przewodnia i nadzorcza rola człowieka (KE 2019: 17–19), innymi słowy, odpowiedzialność technologii jest ograniczona na rzecz człowieka i kontroli, jaką sprawuje on nad sztuczną inteligencją.

Luciano Floridi wraz z 12 ekspertami od sztucznej inteligencji dokonał szerszej analizy zasad, jakie postulowane są w aktach prawnych i rekomendacjach dotyczących sztucznej inteligencji i zaproponował całościowe podsumowanie zysków i strat ze sztucznej inteligencji dla Dobrego Społeczeństwa AI (*Good AI Society*). Zanim przejdę do analizy tego podsumowania, chciałabym zwrócić uwagę na język, jakim Floridi rozpoczyna swój artykuł:

> W każdym przypadku sztuczna inteligencja może być użyta (*used*) do wspierania ludzkiej natury i jej potencjału, tworząc w ten sposób możliwości; niedostatecznie wykorzystywana (*underused*), tworząc w ten sposób możliwe koszty; lub nadużywana (*overused*) lub niewłaściwie wykorzystywana (*misused*), tworząc w ten sposób ryzyko. Jak wskazuje terminologia, zakłada się, że wykorzystanie sztucznej inteligencji jest synonimem dobrych innowacji i pozytywnych zastosowań tej technologii. Jednak strach, niewiedza, nieuzasadnione obawy lub nadmierna reakcja mogą doprowadzić społeczeństwo do niedostatecznego wykorzystania technologii AI poniżej ich pełnego potencjału, z powodów, które można ogólnie określić jako niewłaściwe (Floridi et al. 2018: 690–691).

Floridi zdaje sobie sprawę z tego, że kluczowy dla korzyści i strat związanych ze sztuczną inteligencją jest sposób jej użycia. Jeśli używana będzie w celach

sprzyjających społecznemu rozwojowi, to oczywiście stwarzać będzie możliwości pozytywne, jeśli natomiast zostanie źle użyta albo w ogóle zignorowana, przyczynić się może do zmniejszenia szans rozwojowych społeczeństwa. Floridi tym samym sięga do argumentu znanego już od czasów Karola Marksa, który w maszynie fabrycznej widział sprzymierzeńca robotnika, o ile zostanie ona wyzwolona z kapitalistycznego celu produkcji, jakim jest zysk. Problem w tym, że podobnie jak Marksowski postulat właściwego zastosowania technologii dla korzyści ludzkości rozbił się o brak politycznie trwałych sukcesów rewolucji proletariackiej tam, gdzie rzeczywiście z kapitalizmem mieliśmy do czynienia, tak i argumentacja Floridiego zdaje sprawę z tego, jak chcielibyśmy, aby sztuczna inteligencja była używana, ale nie oznacza to w żadnym sensie, że z takim jej użyciem mamy faktycznie do czynienia. Ponadto ów instrumentalistyczny stosunek do sztucznej inteligencji, jako *de facto* środka do realizacji np. społeczeństwa bardziej sprawiedliwego, zdradza, że sama „moralna sztuczna inteligencja" nie jest tu celem. Raczej moglibyśmy powiedzieć, że sztuczna inteligencja jest traktowana jako instrument neutralny moralnie, skoro może być zastosowana do lepszych i gorszych społecznie celów, bądź jako instrument, który można „nauczyć" moralności rozumianej jako pożądane społecznie sposoby działania. To drugie podejście wiąże się z delegowaniem moralności na artefakty, sterowaniem ludzkim zachowaniem i współkształtowaniem przez artefakty decyzji moralnych, które podejmujemy. W szczególności zwrócił na to uwagę Peter-Paul Verbeek, gdy poddał analizie moralny wymiar badań ultrasonograficznych, jakim poddawane są kobiety w ciąży (Verbeek 2008). Verbeek zauważył, że np. decyzja o aborcji czy o leczeniu płodu podjęta może być tylko dlatego, że technologia ujawniła nam jakiś problem, który należy moralnie rozstrzygnąć. Verbeek zwraca uwagę, że technologie „pomagają kształtować naszą egzystencję i podejmowane przez nas decyzje moralne, co niezaprzeczalnie nadaje im wymiar moralny" (Verbeek, 2011: 2). Z fenomenologicznej perspektywy technologie współkształtują nasze codzienne środowisko, nasze doświadczenia, nie mogą zatem nie być traktowane jako współobecne w decyzjach moralnych. Z argumentem Verbeeka trudno się nie zgodzić, choć dyskusyjny jest zakres owego udziału technologii w decyzjach moralnych. Technologie mogą, ale nie muszą, być szczególnie istotnym czynnikiem w rozstrzyganiu dylematów moralnych, a nawet, z punktu widzenia choćby założenia o autonomii podmiotu moralnego, być może sama technologia nawet jeśli prowokuje problem, nie stanowi części w procesie znajdowania dla niego rozwiązania. Założenie, że technologie współkształtują nasze decyzje moralne, skutkuje bowiem przesunięciem dyskusji o moralności z podmiotu na przedmioty, czyli do „moralizowania technologii" (*moralizing technology*), co postuluje sam Verbeek. Dyskurs ten prowadzi zatem do większego zwracania uwagi na to, jak

i dlaczego pewne rozwiązania powinny zostać zaprojektowane, aby np. maksymalizować bezpieczeństwo użytkowników czy pożądane społeczne korzyści (Verbeek & Slob 2006). Problem w tym, że jeśli owa delegacja moralności na rzeczy odbywa się w fazie projektowania, to faktyczna odpowiedzialność za to, czy artefakt działa tak, jak powinien, spada na projektantów. Nawet jeśli pierwsze smartfony z ekranem dotykowym wykluczały z pola użytkowników osoby niewidome, to nazwanie tych technologii „dyskryminującymi" nie dotyczy tak naprawdę samego urządzenia, a społecznych skutków ich ograniczonego projektu. Możliwość uzupełnienia takich artefaktów o np. asystenta głosowego poprawiło ten projekt, ale odpowiedzialność za to była w rękach producentów (zob. Goggin & Newell 2003). Dlatego, gdy Langdon Winner analizował projekty wiaduktów Roberta Mosesa prowadzących do plaży na Long Island w Nowym Jorku i stwierdził, że „artefakty są polityczne", to starał się wytłumaczyć, że artefakty technologiczne zawsze uwikłane są w jakiś porządek władzy, a z całą pewnością, określony porządek infrastrukturalny wywołują:

> Rzeczy, które nazywamy „technologiami", są sposobami budowania porządku w naszym świecie. Wiele urządzeń technicznych i systemów ważnych w codziennym życiu zawiera możliwości dla wielu różnych sposobów porządkowania ludzkiej aktywności. Świadomie lub nie, celowo lub nieumyślnie, społeczeństwa wybierają struktury dla technologii, które wpływają na to, jak ludzie będą pracować, komunikować się, podróżować, konsumować i tak dalej przez bardzo długi czas. W procesach, w których podejmowane są decyzje strukturalne, różni ludzie są różnie usytuowani i posiadają nierówne stopnie władzy, a także nierówne poziomy świadomości (Winner 1986: 28).

Dlatego właśnie narracja konsekwencjalistyczna wokół AI posiada z jednej strony dobre filozoficzne podstawy, gdyż sięga do studiów nad nauką i technologią, delegacji moralności na artefakty, fenomenologii czy częściowo teorii aktora-sieci, ale traktuje technologie jedynie jako element „infrastruktury moralnej", a nie faktyczny podmiot moralny. Dlatego sztuczna inteligencja, jakiej możemy zaufać, to w tej narracji *de facto* zaufanie do podmiotu dostarczającego nam technologię sztucznej inteligencji i przyjęcie, że jest godny zaufania co do celów i sposobów dostarczania takiej technologii użytkownikom oraz spełniania społecznych, ekonomicznych i politycznych oczekiwań co do realizacji określonych wartości.

W zasadach, jakie powinny być kluczowe w tworzeniu i wdrażaniu sztucznej inteligencji, za analizą dokumentów i deklaracji przeprowadzoną przez Floridiego, dominuje pięć kluczowych zasad: dobrobyt (*beneficience*), niekrzywdzenie (*non-maleficience*), autonomia, sprawiedliwość oraz możliwość wyjaśnienia (*explainability*).

Dobrobyt sprowadza się do troski o ludzką godność oraz o planetę. Sztuczna inteligencja ma zasadniczo służyć postępowi ludzkości, dobru wspólnemu oraz dobremu życiu. *Niekrzywdzenie* oznacza troskę o zachowanie prywatności, bezpieczeństwa oraz ocenę ryzyka kierunków rozwoju sztucznej inteligencji. Istotą tego postulatu jest zatem minimalizacja krzywd, jakie stosowanie AI może wywołać. *Autonomia* wiąże się z delegacją działania na technologię. Wraz z możliwością scedowania podejmowania wielu decyzji na sztuczną inteligencję człowiek powinien mieć wciąż prawo do kontroli tego, co AI robi. Ponadto podkreślona zostaje w tym postulacie „wartość ludzkiego wyboru (*value of human choice*) – przynajmniej co do istotnych decyzji" (Floridi et al. 2018: 698), a więc zachowanie ostrożności w delegowaniu zbyt wielu decyzji na sztuczną inteligencję i dowartościowanie decyzji podejmowanych przez człowieka. *Sprawiedliwość* sprowadza się do promocji wspólnych wartości i dobrobytu, ograniczania powielania i podtrzymywania przez sztuczną inteligencję stereotypów i uprzedzeń oraz dążenie do tego, aby ta technologia odnosiła się z szacunkiem do wszystkich ludzi bez względu na rasę, pochodzenie czy płeć. *Możliwość wyjaśnienia* oznacza natomiast, że sposób działania sztucznej inteligencji powinien być zrozumiały dla ludzi, co ma minimalizować ryzyko nieprzewidzianych konsekwencji, których człowiek nie byłby w stanie kontrolować.

Powyższe zasady – po pierwsze – sugerują, że właściwym beneficjentem działania sztucznej inteligencji jest człowiek. Innymi słowy, AI nie działa we własnym imieniu czy „dla siebie". Po drugie, zasady te mają co prawda służyć za drogowskaz w tworzeniu i wdrażaniu sztucznej inteligencji w różnych obszarach życia, ale ich faktyczna zbieżność ze sztuczną inteligencją weryfikowana jest najczęściej w procesie jej działania. Moralna AI jest zatem „moralna" dlatego, że generuje problemy wymagające rozwiązania, a skuteczność tych rozwiązań mierzona jest skutkami, jakie z nich wynikają. Gdy zatem w narracji konsekwencjalistycznej pojawia się postulat sztucznej inteligencji godnej zaufania, to wydaje się, że owo zaufanie nie dotyczy jej samej, a tego, jakie wartości zostały w nią zaimplementowane i jakie efekty to wywołuje. Powiązanie badań nad sztuczną inteligencją i etyką ma tu zatem charakter bardzo techniczny i często pozbawiony wyjaśnienia, czy pożądane „efekty" i „zyski" z wdrażania sztucznej inteligencji mierzone są ze względu na korzyści, przyjemności, szczęście, czy określają je sami badacze, opinia publiczna czy politycy.

II. Narracja personalistyczna

W narracji personalistycznej sztuczna inteligencja zostaje potraktowana jako aktor społeczny o pewnym stopniu własnej autonomii. Tak jak w narracji konsekwencjalistycznej AI jest aktantem, gdyż działa, to jednocześnie owo działanie wpisane jest w całą infrastrukturę wartości o charakterze społecznym, kulturowym czy politycznym. W narracji personalistycznej sztuczna inteligencja staje się niejako oddzielona od swojej infrastruktury i zyskuje „własne życie", a nawet potencjalnie prawa, przywileje i godność. W znacznie większym stopniu jest ona także obecna jako *quasi*-podmiot w interakcjach z użytkownikami, gdyż przykładami, jakie często ilustrują ten rodzaj narracji, są roboty społeczne czy czatboty.

W ramach etyki przyjmowanej w narracji personalistycznej dominują dwa podejścia: zorientowane na aktora i zorientowane na pacjenta (czy inaczej odbiorcę). Orientacja na aktora zbliża w wielu miejscach narrację personalistyczną do konsekwencjalistycznej, gdyż zakłada ona, iż sztuczna inteligencja jest źródłem określonego działania, które może być wartościowane moralnie. Skupia ona zatem uwagę na sztucznej inteligencji jako spełniającej określone kryteria traktowania jej jako sprawczego podmiotu moralnego. Orientacja na pacjenta pozwala natomiast postawić pytania o to, w jaki sposób podmiot ludzki odnosić powinien się do sztucznej inteligencji i ją traktować. Etyka może zatem być postrzegana jako budowana albo pod sztuczną inteligencję w jej stosunku do człowieka, albo pod człowieka w jego stosunku do sztucznej inteligencji.

W ramach narracji personalistycznej dostrzega się istotną rolę sztucznej inteligencji w budowaniu relacji z ludźmi na poziomie emocjonalnym i związaną z tym możliwość posiadania przez AI praw. Humanizowanie sztucznej inteligencji, postrzeganie jej jako osoby, czy jako podmiotu o „sztucznym życiu", krytykowane być może jako współczesna forma antropomorfizacji i myślenia magicznego (Musiał 2019), ale bez względu na ocenę sensowności takich praktyk społecznych niewątpliwym jest, jak wykazuje wiele badań, że ludzie z robotami empatyzują, nawet gdy nie są one wyrafinowanymi AI (Suzuki et al. 2015; Darling et al. 2015). Ta empatyzacja może wiązać się ze złudzeniem wynikającym ze skutecznej komunikacji językowej, z widzenia w sztucznej inteligencji ról społecznych, w które sami jesteśmy wrzuceni (Gertz 2018), albo z niechęcią człowieka do patrzenia na cierpienie innych istot (Juchniewicz 2024), niemniej jest rzeczywistym doświadczeniem wielu ludzi i przyczynia się do dyskusji na temat rozszerzenia podmiotowości na aktorów nie-ludzkich przede wszystkim ze względu na fakt, że są bądź mogą być istotami czującymi. Dlatego Nauhäuser (2015) stwierdza, że jeśli sztuczna inteligencja będzie zdolna do odczuwania bólu, to powinna mieć prawa. Nawet jeśli aktualnie nie jesteśmy w stanie

stworzyć sztucznej inteligencji odczuwającej ból, co najwyżej skutecznie imitującą radość czy strach, to projektowanie takich praw z wyprzedzeniem może pozwolić na ich wdrożenie wtedy, gdy taka technologia rzeczywiście się pojawi. Jak zauważa Erica L. Neely (2013):

> (...) jeśli coś zachowuje się wystarczająco podobnie do mnie w szerokim zakresie sytuacji, to powinniśmy rozszerzyć na to status moralny. (…) Widzę, że moralna wina w byciu zbyt konserwatywnym jest znacznie większa niż w ryzyku bycia zbyt hojnym.

Neely chodzi, rzecz jasna, o hojność ontologiczną i moralną. Można powiedzieć, że podmiot ludzki nie traci wiele ze swojej sprawczości, rozszerzając działanie moralne na byty inne niż ludzkie, może natomiast wiele stracić, jeśli do moralności podchodzić będzie zbyt ekskluzywnie.

Bycie istotą czującą, choćby potencjalnie, zakłada zatem, że sztuczna inteligencja powinna być chroniona prawnie w sposób podobny do zwierząt czy sztucznie powoływanych na drodze prawnej osób. W filozofii nowożytnej już Tomasz Hobbes w XVI rozdziale *Lewiatana* wskazał, w jaki sposób można zapewnić reprezentację prawną rzeczy i podmiotów nie-ludzkich, a tym samym realizować ich cele czy dobro. Pośród takich podmiotów, które mogą być reprezentowane przez kogoś innego, a zatem mieć prawa, wymienia m.in. kościół czy most (Hobbes 2009: 246). Z punktu widzenia prawa, sięgającego w swej tradycji do prawa rzymskiego, możliwe jest zatem „chronienie praw sztucznej inteligencji", nawet jeśli ona sama nie może się o owe prawa ubiegać. H. Ashrafian (2015: 325) stwierdza:

> W tym przypadku roboty prawdopodobnie (…) nie replikowałyby się samodzielnie, nie zajmowałyby stanowisk publicznych ani nie byłyby właścicielami ziemi i firm, ale byłyby chronione przez prawo i miałyby możliwość wniesienia wkładu w społeczeństwo poprzez takie przykłady, jak obrona narodów i uczestnictwo w sektorze opieki zdrowotnej.

Jeśli nawet prawa robotów są uzasadnialne, a odpowiednie narzędzia prawne mogą doprowadzić do realizacji tych praw, to wciąż nie jest jasne, w jaki sposób sztuczna inteligencja ma być postrzegana jako podmiot moralny i czy w ogóle za taki podmiot postrzegana być powinna. Podstawowym argumentem jest tutaj brak świadomości – AI nie jest obdarzona świadomością, choć w niektórych postaciach świetnie taką świadomość imituje (przez np. używanie języka), dlatego nie

może być podmiotem moralnym, a i przyznawanie jej praw nie jest konieczne, skoro w zasadzie jej status porównać można do rzeczy. „Roboty są art[e]faktami i dlatego, w oczach wielu, nie mają elementu świadomości, który wydaje się być powszechnie uważany za linię podziału między tym, czy zasługują na etyczne traktowanie, czy nie" (Levy 2009: 210). Jak zauważa David Levy, choć sztuczna inteligencja nie ma świadomości, a przynajmniej nie w sensie, w jakim zakładamy, że mają ją ludzie, to postępujący rozwój badań w obszarze AI może doprowadzić do wyłonienia się tzw. sztucznej świadomości (*artificial consciousness*), czyli formy świadomości innej niż ludzka. Już dzisiaj z badań Jenny Burell wiemy, a przynajmniej możemy „zobaczyć", w jak inny sposób – w porównaniu do ludzkiego sposobu rozpoznawania obiektów – sztuczna inteligencja identyfikuje tak banalne obiekty jak cyfry. Fakt, że AI widzi różnice w stopniu zaczernienia pól pikseli, które to różnice są niewidoczne dla ludzkiego oka, i że ostatecznie nie widzi ona cyfr, a stopnie szarości obiektów, oznacza, że jej sposób identyfikacji i kategoryzacji obiektów nie daje się porównać z tym, jak widzi rzeczy i kategoryzuje je człowiek (Burrell 2016: 5–7). Być może zatem, rozszerzając funkcje kognitywne AI, rzeczywiście będziemy mieć do czynienia ze sztuczną świadomością, która wymaga od nas określenia jej praw. Wendel Wallach i Colin Allen (2009: 39) postulują tzw. moralność funkcjonalną, a więc założenie, że po pierwsze, moralność do sztucznej inteligencji można zaimplementować, a po drugie, że może ona nauczyć się właściwej oceny etycznej swoich działań (zob. Coeckelbergh 2020: 52). John Sullins (2006) idzie o krok dalej i zwraca uwagę, że algorytmy sztucznej inteligencji działają i uczą się w sposób inny niż ludzie. Nie rozumiemy w pełni, dlaczego sztuczna inteligencja rozpoznaje pewne obiekty czy identyfikuje pewne jakości równie skutecznie jak człowiek, gdyż nie robi tego „na ludzki sposób". Owo wymykanie się AI samym programistom pozwala mówić o pewnym poziomie autonomii sztucznej inteligencji. Jeśli zatem jest ona częściowo autonomiczna, jeśli jednocześnie możemy wyjaśnić jej, jakimi zasadami ma się kierować, a ona je rozumie i właśnie w określony sposób postępuje, to zdaniem Sullinsa jest ona podmiotem moralnym (Coeckelbergh 2020: 54).

Levy stawia w związku z tym ważny problem: Czy moralnym jest w ogóle tworzenie przez podmiot ludzki istot potencjalnie czujących i świadomych? Czy myśląc o możliwościach rozwoju sztucznej inteligencji nie zapominamy o tym, jak tworzenie takich podmiotów/przedmiotów wpływa na człowieka i jego stosunek do siebie samego, innych ludzi i wreszcie osób nie-ludzkich? W tym kierunku idzie krytyka etyki sztucznej inteligencji zaproponowana przez Joannę Bryson (Bryson, 2010). W kontrowersyjnym artykule „Robots should be slaves" Bryson argumentuje, że sztuczna inteligencja nie jest wolna od ludzkiej woli, nie jest autonomiczna w swoim działaniu, a wszelkie próby uczynienia jej taką,

to tylko zrzucanie przez ludzi odpowiedzialności za to, co sami tworzą. Bryson (2010: 65) stwierdza:

> Projektujemy, produkujemy, posiadamy i obsługujemy roboty. Ponosimy za nie całkowitą odpowiedzialność. Określamy ich cele i zachowanie, bezpośrednio lub pośrednio poprzez określenie ich inteligencji lub nawet bardziej pośrednio poprzez określenie sposobu, w jaki nabywają własną inteligencję. Ale na końcu każdego pośrednictwa leży fakt, że nie byłoby robotów na tej planecie, gdyby nie celowe ludzkie decyzje o ich stworzeniu.

Czy zatem w narracji personalistycznej możliwe jest mówienie o sztucznej inteligencji jako odpowiedzialnej albo godnej zaufania? Model personalistyczny, jako zbudowany na koncepcji reprezentowania uzasadnialnych praw sztucznej inteligencji przez człowieka mówiącego w jej imieniu, zakłada, że sztuczna inteligencja nie może być w pełni odpowiedzialna za swoje czyny, bo owa odpowiedzialność spoczywa na człowieku. Jeśli jednakże wierzymy w to, że uda się w przyszłości powołać do życia sztuczną inteligencję czującą albo świadomą, to problem z takimi postulatami etycznymi polega na ich czystej hipotetyczności. Spełniony musi być bowiem podstawowy warunek zamiany „jeśli będzie" na „jest".

III. Krytyka narracji o „odpowiedzialnej sztucznej inteligencji" i „AI godnej zaufania"

Narracja konsekwencjalistyczna i narracja personalistyczna, mówiąc o „odpowiedzialnej sztucznej inteligencji" czy „AI godnej zaufania", mówią *de facto* o dwóch różnych problemach. Pierwszym z nich jest „zaufanie do" i „poczucie odpowiedzialności za", które powinno obowiązywać projektantów i producentów tego typu technologii oraz osób odpowiedzialnych za proces tworzenia, wdrażania i korzystania z AI. Ponadto ufność w technologię sztucznej inteligencji rośnie wraz ze skutecznością w realizacji praktycznych celów – niemożliwe jest wycofanie się z technologii, która tak wiele rzeczy ludziom ułatwia i z kapitalistycznego punktu widzenia optymalizuje i usprawnia. Koszty etyczne można zatem minimalizować, szanse wyrównywać, algorytmy lepiej trenować, sięgać do danych wolnych od uprzedzeń, ale na końcu i tak pozostanie nam po prostu nadzieja, że technologia ta nie zostanie przeciwko nam użyta, a będzie realizować postulat „dobra wspólnego". Narracja konsekwencjalistyczna, która jest źródłem dla tak rozumianej „moralnej AI", traktuje tę technologię bardziej jako instrument ludzkiego dobrobytu i rozwoju, medium w relacjach międzyludzkich, technologię, na którą delegować możemy określone

wartości i narzędzie zachowujące przynajmniej neutralność aksjologiczną. W tym sensie to podmiot ludzki i jego potrzeby oraz często praktyczne interesy mają być w takim rozumieniu AI chronione, a sztuczna inteligencja godna zaufania i tak nie jest nigdy „puszczona wolna" i choćbyśmy jej ufali, to wymagana jest od ludzi jej kontrola i nadzór. Sztuczna inteligencja godna zaufania przesuwa tak naprawdę dyskurs etyczny w kierunku polityki i tego, kto i w jaki sposób decyduje o korzystaniu z AI.

Drugim problemem jest postrzeganie sztucznej inteligencji jako partnera w naszej codzienności, a więc asystenta głosowego, hologramu, czatbota, wirtualnej persony, robota społecznego, który zachowując się tak jak człowiek, albo po prostu jak podmiot moralny, stawia przed nami wyzwanie rozszerzenia prawa i moralności na AI. Narracja personalistyczna, mówiąc o odpowiedzialnej AI, zwraca uwagę na fakt, że już istniejące narzędzia prawne pozwalają zabezpieczyć prawa sztucznej inteligencji choćby przez należyte reprezentowanie jej interesów w sposób porównywalny bodaj do ochrony interesów takich sztucznych podmiotów jak spółki. Ale taki stosunek do AI, choć traktuje ją jako podmiot prawa, nie przyznaje jej odpowiedzialności za czyny. Jeśli natomiast rozszerzyć tę odpowiedzialność i uznać, że jeśli AI zachowuje się tak, jak zachowałby się moralnie postępujący człowiek, innymi słowy, jeśli zdaje się ona rozumieć moralność jako koncepcję, to zobowiązani jesteśmy przynajmniej wziąć pod uwagę jej możliwe prawa. Wraz z rozwojem samej technologii można założyć, że w przyszłości uda się wytworzyć czującą AI, a być może i uznać, że posiada ona pewną sztuczną świadomość – powinniśmy zatem rozszerzyć prawa na tego typu osoby nie-ludzkie. Narracja personalistyczna ignoruje jednakże, że sztuczna inteligencja, choć może odtwarzać etykę i określone zasady moralne, to nie jest sama w stanie ich stworzyć. Nie jest zatem zdolna do autorefleksji moralnej jako wyrazu autokrytyki swojego działania ani do stworzenia etyki jako samej koncepcji. Innymi słowy – człowiek zdolny jest do autonomicznego uzasadnienia maksymy swojego postępowania i do myślenia opartego na refleksji – zwrocie ku sobie, ale przez konieczne wyjście w swoim myśleniu i wartościowaniu poza siebie, w kierunku praw, na które mogliby wyrazić zgodę inni ludzie. Sztuczna inteligencja nie buduje intersubiektywności i poczucia siebie w procesie własnego, fenomenologicznego stawania się. Może się uczyć lepiej rozpoznawać pewne wzorce, a nawet coraz lepiej dostosowywać do nich wartościowanie moralne, ale nie staje się przez to bardziej ludzka. Ponadto człowiek może chcieć zmiany, może ukierunkować zatem swoje działanie na realizację wartości, co pozwala mu być właśnie odpowiedzialnym za własne czyny; trudno podobnej autorefleksji oczekiwać od AI.

W ramach narracji personalistycznej badacze próbują z jednej strony dokonać ontologicznego rozszerzenia podmiotowości i sprawstwa, a co za tym idzie,

odpowiedzialności poza człowieka, i w zasadzie sam ten kierunek uznać należy za szlachetny „ruch wyprzedzający" możliwy rozwój sztucznej inteligencji, jednakże z drugiej popadają w błąd, który uznać można za „przesuwanie" czy „zrzucanie odpowiedzialności" za czyny generowane przez AI. Jeśli za samo pojawienie się sztucznej inteligencji jako określonego typu technologii odpowiada człowiek, to i za możliwy stopień jej emancypacji i autonomii odpowiadać powinien także człowiek. Wątpliwości etyczne budzi zresztą samo konstruowanie technologii, która mogłaby zyskać świadomość czy czuć, i to niekoniecznie ze względu na obawy, do czego taka AI byłaby zdolna, ale ze względu na wyraźny w obydwu narracjach praktyczny stosunek ludzi do tej technologii. Sztuczna inteligencja została skonstruowana, by rozwiązywać określone problemy, by przyśpieszać procesy analityczne, by poddawać obróbce wielkie zbiory danych, by te dane zbierać, by nadzorować określone procesy, by w konkretny sposób wchodzić w interakcję z ludźmi. Ta szczególna, praktyczna specjalizacja AI i fakt, że mamy do czynienia nie z jedną sztuczną inteligencją, a wieloma ich postaciami, programami, aplikacjami i urządzeniami, ujawnia, że wartości moralne pojawiają się w ramach tworzenia tej technologii dopiero *post factum*. Refleksja etyczna w obszarze AI to także skutek rozwoju samej AI, a nie przyczyna *a priori* jako np. troska o rozwój takich technologii, które służyć będą ludzkości. Innymi słowy, etyczny wymiar AI zauważyliśmy dopiero wtedy, gdy okazało się, że Siri czy Alexa mogą zabrać pracę lektorom; gdy sztuczna inteligencja przemysłowa zaczęła zastępować pracowników fabrycznych; gdy w starzejących się społeczeństwach dotarło do nas, że kiedyś pielęgniarkę może zastąpić robot asystujący, i gdy ludzie uznali, że rozmowa z hologramem jest bardziej satysfakcjonująca niż z innym człowiekiem. Przy okazji badań nad algorytmami odkryliśmy także, że technologia nie jest neutralna, że sztuczna inteligencja powiela stereotypy, uprzedzenia (Rafanelli 2022; Waelen, Wieczorek 2022), że manipulacyjny sposób posługiwania się AI szkodzi demokracji (Coeckelbergh 2024) i ma poważne konsekwencje społeczno-polityczne, a jakość danych, na których bywa trenowana, budzi wątpliwości etyczne co do ich rzetelności czy sposobu pozyskania. Ostatecznie zatem sztuczna inteligencja mówi nam więcej o nas samych niż o sobie jako Innym. Nie przesądza to jednakże, że sztuczna inteligencja nie jest Innym, którego należałoby potraktować poważnie (Gunkel 2018). Wskazuje jedynie, że wciąż borykamy się z trudnością, jaką jest formułowanie samej etyki w obszarze AI. David J. Gunkel przypomina bowiem o dobrze znanym twierdzeniu Davida Hume'a, że ze zdań o faktach nie wyprowadza się zdań o wartościach (zob. Ślęczek-Czakon 1992). Innymi słowy, zdaniem Gunkela błędem byłoby wartościowanie sztucznej inteligencji jako „moralnej" albo „godnej zaufania" dlatego, że posiada ona określone jakości, które uznajemy za podstawę „moralności" albo „godnego zaufania", gdyż wówczas wartość moralną udowodnić

musi podmiot/przedmiot przez wpisanie się w określony normatywny horyzont oczekiwań. Logika tak rozumianej etyki to przejście od „X jest A" bądź „X posiada cechę B" do „X powinno się traktować w sposób M". Z historii wiemy, że taki normatywny horyzont, po pierwsze, nie jest stały i zmienia się w czasie, po drugie, fakt, że w różnych momentach w czasie uznawano za podstawę moralności posiadanie określonych jakości, np. bycie mężczyzną, bycie reprezentantem rasy białej, bycie rozumnym w sensie sprawnej komunikacji językowej, bycie człowiekiem dorosłym itd., wcale nie oznacza, że rzeczywiście mieliśmy do czynienia z wartościowaniem, które uznalibyśmy za moralne. Gunkel sugeruje zatem, że najpierw musimy zastanowić się, jakimi wartościami należałoby się kierować w ramach działania moralnego, a w drugim kroku zweryfikować, czy my, jako ludzie, odnosimy się do bytów nie-ludzkich w taki sposób, jaki skądinąd uważamy za moralny. Gunkel zgadza się tu częściowo z Markiem Coeckelberghiem, który zauważył, że niemoralne traktowanie technologii nie oznacza pogwałcenia samej technologii, ale stanowi moralną wadę – jest przeciwieństwem cnoty (Coeckelbergh 2018: 145; zob. Darling 2012), co wpływa na podmiot ludzki w sposób deprawujący jego charakter. Gunkel przywołuje nawet mocniejszy argument, gdyż powołuje się na Emmanuela Lévinasa i jego założenie, iż etyka poprzedza ontologię (Gunkel 2018: 269). Ostatecznie bowiem, jak argumentuje Gunkel, w etyce chodzi o to, jak człowiek odnosi się do inności i różnorodności, która dotyczy rzecz jasna ludzi, ale może dotyczyć także bytów nie-ludzkich.

Podsumowanie

W niniejszym artykule wyszczególniłam dwie, dominujące narracje, a więc pewne zbiory przekonań, wyobrażeń i argumentów wokół moralnej sztucznej inteligencji. Pierwsza sprowadza się do traktowania sztucznej inteligencji jako moralnej ze względu na konsekwencje, jakie jej działanie wywołuje. Sama AI w tej narracji traktowana jest jako narzędzie, medium albo po prostu element infrastruktury wartości, w który wpisać należy określone zasady działania, tak by minimalizować niepożądane skutki. Jak pokazałam, narracja konsekwencjalistyczna, postulując „AI godną zaufania", określa tak naprawdę spektrum wartości, które powinny być brane pod uwagę, gdy technologia ta jest projektowana, wdrażana i używana. Problem w tym, że wartości te można zdefiniować jedynie w bardzo szeroki sposób, gdyż szczegółowe aplikacje sztucznej inteligencji do życia społecznego są często bardzo punktowe, a także trudne do przewidzenia. Można zatem postulować „sprawiedliwą AI", która nie powiela uprzedzeń i stereotypów, ale w praktyce owa sprawiedliwość może być rozumiana na różne sposoby w sensie kulturowym, społecznym, a zwłaszcza politycznym, a jej realizacja pozostaje w rękach człowieka.

Druga narracja, personalistyczna, stara się spojrzeć na AI jak na podmiot moralny tak w sensie sprawczości moralnej, jak i bycia podmiotem praw i przywilejów adresowanych do niej przez człowieka. Narracja ta pozwala na zwrócenie uwagi na ontologię, a więc jakości rzeczy, którym przypisujemy status moralny. Najczęściej prowadzi to do dyskusji warunków, jakie spełnić musi przedmiot, by stać się podmiotem moralnym – a więc musi posiadać umysł lub być istotą czującą. Tak, jak sama narracja personalistyczna stanowi ważny głos w dyskusji nad tym, co w podmiocie, a nie konsekwencjach jego działania decyduje o tym, że przypisujemy mu moralność, to problematyczne jest adresowanie tej narracji do sztucznej inteligencji. Dotychczas w przypadku AI mamy do czynienia z postulowaniem pewnych jakości, a nie ich faktycznym posiadaniem przez tę technologię. Narracja personalistyczna stawia ontologię przed etyką, a to samo w sobie może być błędnym podejściem. Wymaganie od przedmiotu/podmiotu, by udowodnił nam posiadanie przez siebie cech uznawanych za niezbędne do posiadania moralności, jest pewną formą moralnego konserwatyzmu i stawianiem podmiotu ludzkiego za wzór, tak jakby w samej etyce nie trwały od wieków spory, kim jest i w jakich warunkach mówić możemy w ogóle o podmiocie moralnym. Pewną propozycją wyjścia poza etykę zorientowaną na ontologię jest inspirowana Lévinasem etyka otwarcia na Innego, a więc otwarcia na radykalną różnicę. Być może w poszanowaniu Inności jako rzeczywiście innej, a nie podobnej do nas samych (pod pewnymi warunkami), tkwi szansa na zbudowanie *moralnych relacji* ze sztuczną inteligencją, a nie tylko definiowanie moralnej AI czy AI godnej zaufania.

Bibliografia

[1] Airoldi, M. (2022). *Machine Habitus. Towards a Sociology of Algorithms*. Cambridge, UK, Medford, USA: Polity Press.

[2] Andreson M., Anderson S. L. (eds) (2011). Machine Ethics. Cambridge: Cambridge University Press.

[3] Ashrafian, H. (2015). Artificial intelligence and robot responsibilities: Innovating beyond rights. *Science and Engeenering Ethics* 21 (2): 317–236. DOI 10.1007/s11948-014-9541-0

[4] Bryson, J. J. (2010). Robots should be slaves. W Y. Wilks (Ed.), *Close Engagements with Artificial Companions: Key social, psychological, ethical and design issues* (s. 63–74). John Benjamins Publishing Company. https://doi.org/10.1075/nlp.8.11bry

[5] Burrell, J. (2016). How the machine ‘thinks’: Understanding opacity in machine learning algorithms. *Big Data & Society*, *3* (1), https://doi.org/10.1177/2053951715622512

[6] Coeckelbergh. M. (2018). Why Care About Robots? Empathy, Moral Standing, and the Language of Suffering. *Kairos. Journal of Philosophy & Science* 20, 141–158. DOI 10.2478/kjps-2018-0007

[7] Coeckelbergh, M. (2020). *AI Ethics.* Cambridge, Massachusetts, London, England: The MIT Press.
[8] Coeckelbergh, M. (2024). *Why AI Undermines Democracy and What To Do About It*. Cambridge: Polity Press.
[9] Darling, K. (2012). Extending Legal Protection to Social Robots: The Effects of Anthropomorphism, Empathy, and Violent Behavior Towards Robotic Objects (April 23, 2012). Robot Law, Calo, Froomkin, Kerr eds., Edward Elgar 2016, We Robot Conference 2012, University of Miami, https://ssrn.com/abstract=2044797
[10] Darling, K., Nandy, P., Brazeal, C. (2015). Empathic concern and the effect of stories in human-robot interaction. Proceedings of the 24th IEEE International Symposium on Robot and Human Interactive Communication, ed. IEEE, 770–775. DOI: 10.1109/ROMAN.2015.7333675
[11] Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28* (4), 689–707. https://doi.org/10.1007/s11023-018-9482-5
[12] Gertz, N. (2018). Hegel, the Struggle for Recognition, and Robots. *Techné: Research in Philosophy and Technology*, *22* (2), 138–157. https://doi.org/10.5840/techne201832080
[13] Goggin, G., Newell, C. (2003), *Digital disability: The social construction of disability in new media*. Lanham, Maryland, USA: Rowman & Littlefield.
[14] Gunkel, D.J. (2018). *Robot Rights*. Cambridge, Massachusetts, London, England: The MIT Press.
[15] Hobbes, T. (2009). *Lewiatan*. przeł. C. Znamierowski, Warszawa: Fundacja Aletheia.
[16] Juchniewicz, N. (2024). Towards (Unilateral) Recognition of "the Technological Other" – Vulnerability, Resistance and Adequate Regard. *Ethics in Progress* Vol. 15 No 2. 120–136. https://doi.org/10.14746/eip.20242.8
[17] Komisja Europejska, Grupa ekspertów wysokiego szczebla ds. sztucznej inteligencji. (2019). Wytyczne w zakresie etyki dotyczące godnej zaufania sztucznej inteligencji. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai [pobrane w języku polskim: 6.02.2025]
[18] Latour, B. (2013). Technologia jako utrwalone społeczeństwo, przeł. Ł. Afeltowicz. *AVANT,* wol. IV, nr 1/2013, s. 17–48.
[19] Levy, D. (2009). The ethical treatment of artificially conscious robots. *International Journal of Social Robotics* 1 (3): 209–216. https://doi.org/10.1007/s12369-009-0022-6
[20] Musiał, M. (2009). *Enchanting Robots. Intimacy, Magic, and Technology*. Cham: Palgrave Macmillan.
[21] Neely, E.L. (2014). Machines and the moral community. Philosophy & Technology 27 (1): 97–111. https://doi.org/10.1007/s13347-013-0114-y
[22] Neuhäuser, C. (2015). Some skeptical remarks regarding robot responsibility and a way forward. In: *Collective Agency and Cooperation in Natural and Artificial Systems: Explanation, Implementation, and Simulation*, ed. Catrin Misselhorn, 131–148. New York: Springer.

[23] Rafanelli, L.M. (2022). Justice, injustice, and artificial intelligence: Lessons from political theory and philosophy. *Big Data & Society* 9 (1). https://doi.org/10.1177/20539517221080676
[24] Sullins, J. (2006). When is a Robot a Moral Agent? *International Review of Information Ethics* 6, 23–30.
[25] Suzuki, Y., Galli, L., Ikeda, A., Itakura, S., Kitazaki, M. (2015). Measuring empathy for human and robot hand pain using electroencephalography. *Scientific Reports* 5: 15924. https://doi.org/10.1038/srep15924
[26] Ślęczek-Czakon, D. (1992). Gilotyna Hume'a. *Folia Philosophica* 10. 147–159.
[27] Verbeek, P.-P. (2011). *Moralizing Technology*. Chicago: University of Chicago Press.
[28] Verbeek, P.-P. (2008). Obstetric Ultrasound and the Technological Mediation of Morality: A Postphenomenological Analysis. *Human Studies* 31, 11–26. https://doi.org/10.1007/s10746-007-9079-0
[29] Verbeek, P.-P., Slob, A. (eds) (2006). *User Behavior and Technology Development. Shaping Sustainable Relations Between Consumers and Technologies*. Dordrecht: Springer.
[30] Wallach, W., Allen, C. (2009). *Moral Machines: Teaching Robots Right from Wrong*. Oxford: Oxford University Press.
[31] Waelen, R., Wieczorek, M. (2022). The Struggle for AI's Recognition: Understanding the Normative Implications of Gender Bias in AI with Honneth's Theory of Recognition. *Philosophy & Technology* Vol. 35, No 53. https://doi.org/10.1007/s13347-022-00548-w
[32] Winner, L. (1986). Do Artifacts Have Politics? In: *The Whale and the Reactor: a search for limits in an age of hight technology*. Chicago: Chicago University Press, 19–39.

Noty o autorach

prof. dr hab. inż.
Roman Słowiński

Wiceprezes Polskiej Akademii Nauk w latach 2019–2022, Prezes Oddziału PAN w Poznaniu (2011–2018), członek rzeczywisty PAN (od 2013), członek Academia Europaea (od 2013) i zagraniczny członek korespondent w Accademia delle Scienze dell'Istituto di Bologna (od 2024).

Tytuł naukowy profesora uzyskał w 1989 roku. Jest profesorem w Politechnice Poznańskiej oraz w Instytucie Badań Systemowych PAN. W 1989 roku założył Zakład Inteligentnych Systemów Wspomagania Decyzji w Instytucie Informatyki Politechniki Poznańskiej, którym kierował do 2022 roku.

Jest twórcą szkoły naukowej „inteligentnego wspomagania decyzji", która łączy badania operacyjne, sztuczną inteligencję i nowe technologie informatyczne. Wypromował 28 doktorów. Opracował m.in. oryginalną metodykę wspomagania decyzji zwaną dominacyjną teorią zbiorów przybliżonych, za co otrzymał w 2005 roku Nagrodę Fundacji na rzecz Nauki Polskiej.

Od 1999 roku jest redaktorem naczelnym „*European Journal of Operational Research*" (Elsevier) – jednego z najważniejszych czasopism w dziedzinie badań operacyjnych na świecie. Przewodniczący Europejskiej Grupy Roboczej przy EURO nt. Wielokryterialnego Wspomagania Decyzji. Członek w randze *Fellow* wielu towarzystw naukowych, m.in.: IEEE (The Institute of Electrical and Electronics Engineers – 2017), IRSS (International Rough Set Society – 2015), INFORMS (The Institute for Operations Research and the Management Sciences – 2019), IFIP (International Federation for Information Processing – 2019) oraz IFORS (International Federation of Operational Research Societies – 2022).

Sześć uniwersytetów zagranicznych nadało mu tytuł doktora *honoris causa*: Faculté Polytechnique de Mons (Belgia) w 2000 roku, Université Paris Dauphine (Francja) w 2001 roku, Technical University of Crete w Chanii (Grecja) w 2008 roku, Nanjing University of Aeronautics and Astronautics w Nankinie (Chiny) w 2018 roku, Hellenic

Mediterranean University (Grecja) w 2022 roku oraz University of West Attica w Atenach (Grecja) w 2024 roku.

Poza Nagrodą FNP (2005), jest laureatem Złotego Medalu EURO (Aachen 1991), Nagrody im. F. Edgewortha i V. Pareto (Cape Town 1997), Subsydium Profesorskiego „Mistrz" FNP (2001), Nagrody Naukowej Miasta Poznania (2003), Nagrody Naukowej Prezesa PAN (2016), Nagrody Naukowej Prezesa Rady Ministrów (2020) i Nagrody Naukowej Humboldta (Fundacja Humboldta, RFN, 2023).

Jest również laureatem „Złotego Hipolita" (2017) i „Zasłużonym dla Miasta Poznania" (2018). W 2022 roku Premier rządu Republiki Francuskiej nadał mu tytuł i insygnia Oficera Palm Akademickich.

prof. dr hab. inż.

Przemysław Biecek

Polski matematyk i informatyk, profesor nauk inżynieryjno-technicznych o specjalności statystyka matematyczna. Profesor Politechniki Warszawskiej i Uniwersytetu Warszawskiego.

Doktorat obronił 10 lipca 2007 roku w Instytucie Matematyki i Informatyki na Wydziale Podstawowych Problemów Techniki Politechniki Wrocławskiej na podstawie rozprawy pt. *Testowanie zbioru hipotez z zadaną relacją hierarchii wraz z przykładami zastosowań w genetyce*. Habilitację uzyskał 16 kwietnia 2013 roku w Instytucie Biocybernetyki i Inżynierii Biomedycznej im. Macieja Nałęcza Polskiej Akademii Nauk. Tytuł profesora nauk inżynieryjno-technicznych uzyskał 2 lutego 2023 roku.

W działalności naukowo-badawczej zajmuje się statystyką matematyczną, wizualizacją i percepcją danych, metodologią i zastosowaniem uczenia maszynowego m.in. w medycynie. Interesuje się również popularyzacją technik analizy danych.

W 2008 roku zatrudniony na stanowisku adiunkta (później profesora) w Zakładzie Statystyki Matematycznej na Wydziale Matematyki, Informatyki i Mechaniki Uniwersytetu Warszawskiego, a w 2014 roku na stanowisku profesora nadzwyczajnego (później profesora) w Zakładzie Projektowania Systemów CAD/CAM i Komputerowego Wspomagania Medycyny na Wydziale Matematyki i Nauk Informacyjnych Politechniki Warszawskiej. W latach 2013–2015 był adiunktem w Interdyscyplinarnym Centrum Modelowania Matematycznego i Komputerowego UW. W kadencji 2020–2024 prodziekan ds. rozwoju Wydziału MiNI PW.

Pełni funkcję *associate editor* w czasopiśmie *R Journal*.

Wielokrotnie nagradzany uczelnianymi nagrodami indywidualnymi. W 2021 roku otrzymał Nagrodę im. Profesora Zdzisława Pawlaka – Komitetu Informatyki PAN – za Wybitną Monografię z Zakresu Informatyki.

Wybrane publikacje:

- Biecek Przemysław: *Analiza danych z programem R: modele liniowe z efektami stałymi, losowymi i mieszanymi.* Warszawa: Wydawnictwo Naukowe PWN, 2013. ISBN 978-83-01-17453-8.
- Biecek Przemysław*: Odkrywać! Ujawniać! Objaśniać!: zbiór esejów o sztuce prezentowania danych.* Warszawa: Fundacja Naukowa Smarter Poland.pl, 2016. ISBN 978-83-65291-05-9.
- Biecek Przemysław*: Przewodnik po pakiecie R.* Wrocław: Oficyna Wydawnicza GiS, 2014. ISBN 978-83-62780-22-8.

dr hab.

Agnieszka Lekka-Kowalik

Polska filozof, doktor habilitowana nauk humanistycznych, profesor nadzwyczajny Instytutu Filozofii Katolickiego Uniwersytetu Lubelskiego Jana Pawła II (KUL). Studia chemiczne odbyła na Uniwersytecie Marii Curie-Skłodowskiej w Lublinie, a filozoficzne na KUL oraz w International Academy of Philosophy w Liechtensteinie. W 1995 roku obroniła na KUL pracę doktorską *Rationality. A Preliminary Philosophical Account*, a w 2009 roku habilitowała się na podstawie rozprawy zatytułowanej *Odkrywanie aksjologicznego wymiaru nauki.* W latach 1993–1996 była pracownikiem naukowym w Zakładzie Filozofii na Wydziale Nauk Humanistycznych Uniwersytetu Kantonalnego w Neuchâtel. W 1997 roku została zatrudniona w Katedrze Metodologii Nauk KUL, a od 2011 roku jest kierownikiem KMN na stanowisku profesora nadzwyczajnego. Najnowsza książka nosi tytuł *Nauka i społeczeństwo. Paradygmat personalistyczny* (Lublin 2024).

Jej zainteresowania naukowe to:

filozofia nauki, filozofia techniki, etyka sztucznej inteligencji, etyka badań naukowych, logika praktyczna.

Pełnione funkcje, to m.in:
- 2000–2010: redaktor działu obejmującego metodologię i historię filozofii współczesnej w *Powszechnej encyklopedii filozofii*,
- 1.10.2009–31.08.2012: prodziekan Wydziału Filozofii,
- 1.09.2012–14.12.2013: prorektor KUL ds. promocji i współpracy z zagranicą,
- 1.10.2014–31.08.2022: dyrektor Instytutu Jana Pawła II – Ośrodka Badań nad Myślą Jana Pawła II KUL i redaktor naczelna czasopisma „Ethos. Kwartalnik Instytutu Jana Pawła II",
- od 17.09.2021 – wiceprzewodnicząca Zarządu Polskiego Towarzystwa Oceny Technologii (PTOT).

dr

Natalia Juchniewicz

Jest doktorem filozofii (Uniwersytet Warszawski, 2015) i doktorem socjologii (Uniwersytet SWPS, 2018). Rozprawa doktorska z filozofii poświęcona była ontologii techniki w filozofii Hegla i Marksa. Rozprawę doktorską z socjologii poświęciła interakcjom w świecie technologii mobilnych. Od 2016 roku pracuje jako adiunktka na Wydziale Filozofii Uniwersytetu Warszawskiego.

Od grudnia 2016 roku kieruje Pracownią Filozofii Techniki i Komunikacji na Wydziale Filozofii UW, zaś od 2023 roku pełni obowiązki kierowniczki Zakładu Filozofii Społecznej. Publikowała m.in. w *Techné: Research in Philosophy and Technology*, *International Journal of Digital Earth* czy *Phenomenology and the Cognitive Sciences*. Od 2018 roku jest członkinią Society for Philosophy and Technology. W pracy naukowej dr Juchniewicz zajmuje się recepcją myśli Hegla w kontekście filozofii techniki i mediów; teorią uznania w obszarze sztucznej inteligencji, społecznie zorientowanym designem oraz postfenomenologią techniki.