

Miłość, lęk i roboty. Rozpoznanie ludzkich postaw i oczekiwań wobec robotów oraz ich możliwe konsekwencje.

Abstrakt

Praca skupia się na humanistycznych i etycznych aspektach interakcji ludzi z robotami. Na podstawie analizowanych publikacji, zestawia różne metody i narzędzia używane przez naukowców do opisu i badania postaw ludzi wobec różnego rodzaju robotów. Tekst ukazuje wpływ poprzednich doświadczeń oraz status społeczny badanych na ich deklarowaną postawę. W świetle przytoczonych badań pojęcia takie jak autonomia, inteligencja czy intencyjność okazały się kluczowe dla badanych do oceny ich interakcji z robotami. Praca przybliży możliwe konsekwencje dla różnych dziedzin wpływające z przytoczonych analiz.

Słowa kluczowe: roboty, uncanny valley, test turinga, etyka

Abstract

The paper focuses on the humanistic and ethical aspects of human-robot interaction. Based on the publications analysed, it juxtaposes the different methods and tools used by researchers to describe and study human attitudes towards different types of robots. The text shows the influence of previous experiences and the social status of the respondents on their declared attitudes. In the light of the research cited, concepts such as autonomy, intelligence or intentionality were found to be crucial for respondents to evaluate their interactions with robots. The paper looks at the possible implications for various domains of the cited analyses.

Keywords: robots, uncanny valley, turing test, ethics

1. Wstęp

Kogo/co określać będę dalej w tekście mianem „roboty”? Odnoszę wrażenie, że słowo to nie jest szczególnie zgrabne, wywołuje może zbyt aluminiowe i łopatologiczne skojarzenia z gatunku taniego sci-fi. Jednak jak dotąd nie znalazłam, zarówno w literaturze polskiej i zagranicznej terminu, które zawierałoby w sobie interesujący mnie zbiór znaczeń. Chcę więc zaznaczyć, za badaczami na których będę się powoływać w dalszej części tekstu, że twory o których będę opowiadać nie muszą posiadać fizycznej formy, ludzkiej ani jakiegokolwiek w ogóle; choć jak dowiemy się później, ich fizyczność ma ogromny wpływ na ich odbiór przez ludzi, a więc także na całość ich funkcjonowania w ludzkim świecie. Możemy mówić o robotach wirtualnych, funkcjonalnych, imitujących zwierzęta czy chatbotach. Te ostatnie są prawdopodobnie najpowszechniejszym doświadczeniem codziennego użytkownika świata cyfrowego, nie związanego w żaden sposób z robotyką czy nauką w ogóle. Od Elizy, przez irytujące boty *customer service* do robiącego furorę w ostatnim czasie ChatGTP.

To co najbardziej interesujące i kluczowe w świetle przytoczonych poniżej analiz to pewnego rodzaju autonomia cechująca te maszyny oraz umiejętność interakcji społecznych z innymi podmiotami. Jak pisze Agnieszka Wykowska : „Roboty społeczne mogą być narzędziami naukowymi do badania ludzkiego systemu poznania społecznego. Jak się okazuje, system poznania społecznego jest w wielu przypadkach zdolny używać tego samego mechanizmu zarówno w interakcjach z drugim człowiekiem, jak i z robotem.¹” Coraz bardziej oczywiste staje się, że rozwój robotyki może być w przyszłości nierozzerwalnie związany z mnóstwem innych dziedzin takich jak edukacja, medycyna czy system opieki zdrowotnej. Człowiek, dążąc do stworzenia jak najbardziej humanoidalnego robota, tak naprawdę napędza poznanie swojego własnego systemu: od jego złożoności do ułomności, czy to w aspekcie fizycznej budowy czy też systemów i mechanizmów poznawczych. Także więc nauki humanistyczne, psychologia czy filozofia stają przed zadaniem jak się do robotów ustosunkować. W swojej pracy chciałabym się skupić właśnie na aspekcie humanistycznym i badaniach które jako punkt odniesienia w analizie robotów będą brały mentalne doświadczenia i procesy poznawcze. Jednym z pierwszych zarzutów do Testu Turinga był antropocentryzm i postrzeganie inteligencji w aspekcie czysto ludzkim². Myślę, że choć warto dopuszczać możliwość występowania innych rodzajów inteligencji, to człowiek jako twórca robota, niewiele więcej może mu na tym etapie zaoferować, niż próba imitacji swojego własnego, ledwie poznanego systemu. Z drugiej strony raczkująca dopiero robotyka może jeszcze przez jakiś czas nie wywrócić do góry nogami porządku świata skrojonego pod gatunek ludzki, ale może już teraz wiele nam powiedzieć o ludzkich uprzedzeniach czy obawach, w związku z implementacją coraz większej liczby automatycznych i autonomicznych rozwiązań do naszej codzienności. W części drugiej przedstawię badania i analizy, których wspólnym mianownikiem jest ludzkie doświadczenie w interakcji z robotami, a także swego rodzaju projekcje i antropomorfizacje, których dokonujemy w tym kontakcie. Na zakończenie będę starała się zastanowić nad wnioskami, płynącymi z przestudiowanych badań i konsekwencjami jakie mogą nieść za sobą analizowane zjawiska. Przyjrzę się temu, jakie zastosowanie może mieć to w projektowaniu interakcji człowiek-maszyna.

2. Narzędzia badawcze i próby naukowego zbadania różnego rodzaju postaw wobec robotów.

2.1. Kto się boi robotów?

Na wstępie tego podrozdziału należałoby krótko scharakteryzować trzy skale używane w badaniach, które zostaną dalej przywołane. Pierwsza z nich to *Belief in Human Nature Uniqueness Scale* (BHNU) autorstwa Giger, Piçarry, Pochwatki, Almeida i in³. Na

¹ Paweł Łupkowski i Aleksandra Wasielewska, „Nieoczywiste relacje z technologią. Przegląd badań na temat ludzkich postaw wobec robotów”, *Człowiek i Społeczeństwo* LI (2021): 165–87.

² Paweł Łupkowski, *Test Turinga. Perspektywa sędziego* (Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza w Poznaniu, 2010).

³ Łupkowski i Wasielewska, „Nieoczywiste relacje z technologią. Przegląd badań na temat ludzkich postaw wobec robotów”.

7-punktowej skali Likerta, mierzy ona esencjalistyczne przekonania badanych na temat wyjątkowości natury ludzkiej na tle innych podmiotów. BHNU ma polską adaptację, autorstwa Grzegorza Pochwatki i współpracowników. Jej treść przedstawia tabela poniżej (tabela 1).⁴

English version	Polish version
Even if ultra-sophisticated...	Nawet najbardziej zaawansowany technicznie:
a robot will never be considered as human being;	robot nigdy nie będzie uważany za człowieka.
a robot will never feel the same emotions as a human being,	robot nigdy nie będzie odczuwał emocji tak jak człowiek.
a robot will never use language in the same way as a human being;	robot nigdy nie będzie posługiwał się językiem tak samo jak człowiek.
a robot will always be a mechanical imitation of the human being;	robot będzie zawsze imitacją człowieka.
a robot will never have consciousness;	robot nigdy nie posiadać świadomości.
a robot will never have morality.	robot nigdy nie będzie miał moralności.

Tabela 1⁵.

Drugą jest *Negative Attitudes towards Robots Scale* (NARS), zaprojektowana przez Nomure wraz zespołem. Wyróżnia ona postawy związane z lękiem przed komputerami (*computer anxiety*) oraz lękiem przed komunikacją społeczną (*communication apprehension*). Polska adaptacja NARS, również autorstwa Pochwatki, składa się z dwóch podskal: negatywnych postaw wobec interakcji z robotami (NATIR) oraz negatywnych postaw wobec robotów posiadających cechy ludzkie (NARHT).⁶

Różańska-Walczuk i współpracownicy natomiast proponują trzelementową skalę antropomorfizacji⁷, która wyróżnia:

- przypisywanie pozytywnych cech ludzkich (takie jak np. miły, sympatyczny)
- przypisywanie negatywnych cech ludzkich (np. przebiegły)
- oraz cech związanych z zachowaniem (np. zwinny, silny).

⁴ Grzegorz Pochwatko et al., „Polish Version of the Negative Attitude toward Robots Scale (NARS-PL)”, *Journal of Automation Mobile Robotics and Intelligent Systems* 9, nr 3 (2015), https://doi.org/10.14313/JAMRIS_3-2015/25.

⁵ Ibidem

⁶ Łupkowski i Wasielewska, „Nieoczywiste relacje z technologią. Przegląd badań na temat ludzkich postaw wobec robotów”.

⁷ Pochwatko et al., „Polish Version of the Negative Attitude toward Robots Scale (NARS-PL)”.

Międzykulturowe badania wykorzystujące wspomnianą skalę NARS wykazały istotne różnice pomiędzy badanymi różnych narodowości⁸. Ich wyniki można streścić następująco:

- Meksykanie (najbardziej negatywne postawy wobec interakcji z robotami).
- Japończycy (negatywne postawy wobec interakcji z robotami, szczególnie w kontekście emocji).
- Amerykanie (najbardziej pozytywne postawy wobec robotów).

Badanie Gigera i współpracowników⁹ wskazuje na to, że im silniejsze przekonanie badanych o wyjątkowości natury ludzkiej (mierzone przy użyciu skali BHNU), tym bardziej negatywny jest ich stosunek do robotów. To samo badanie dowiodło również, że wyższa religijność wiąże się z mniej pozytywnymi postawami wobec robotów.

Badanie Pičarry¹⁰ ujawniło, że kobiety mają silniejsze przekonanie o wyjątkowości natury ludzkiej niż mężczyźni oraz mniej pozytywne postawy wobec robotów. Potwierdza to wnioski z wcześniej przytoczonego badania Gigera. Autorzy sugerują, że może to mieć związek z różnicami w kontekście użycia robotów przez obie płcie (kobiety - obowiązki domowe, mężczyźni - praca i zabawa). Podobne wnioski uzyskali Wasielewska i Łupkowski w badaniu przeprowadzonym wśród młodzieży¹¹. Wątek ten jest interesujący z uwagi na socjalizację kobiet do większej otwartości emocjonalnej i nawiązywania więzi, jak również ze względu na postrzeganie robota jako (moralnego) podmiotu, co omówię szerzej w następnym podrozdziale.

Ciekawe wnioski płyną także z badań Nomury¹². Zauważył on, że przebyte wcześniejsze doświadczenie obcowania z robotem, rzeczywistego czy zapośredniczonego przez media, zmniejszają negatywne nastawienie wobec robotów, przy czym doświadczenie realne wywiera tutaj większy wpływ.

Większe zainteresowanie *science-fiction* również wiąże się z bardziej pozytywnymi postawami wobec robotów¹³, co może wynikać z powyższego argumentu o poprzedzającym doświadczeniu. Nad rolą gatunku dzieł z gatunku sci-fi oraz obrazu robotów zapośredniczonego przez media w ogóle pochyliła się Ulrike Bruckenberger i współpracownicy¹⁴. Wyniki ich badań nie są jednoznaczne, ale zauważyli oni, że media z pewnością wpływają na kształtowanie się oczekiwań wobec realnych robotów, których te najczęściej nie spełniają. Jednocześnie przedstawienia obecne w kulturze sprawiają że wizja współistnienia ludzi i coraz bardziej zaawansowanych robotów uważamy za całkiem prawdopodobny.

⁸ Łupkowski i Wasielewska, „Nieoczywiste relacje z technologią. Przegląd badań na temat ludzkich postaw wobec robotów”.

⁹ Ibidem

¹⁰ Ibidem

¹¹ Ibidem.

¹² Ibidem.

¹³ Ibidem.

¹⁴ Ibidem.

Victoria Groom, Leila Takayama, Paloma Ochi i Clifford Nass¹⁵ oraz Natalia Reich-Stiebert, Friederike Eyssel i Charlotte Hohnemann¹⁶ zaobserwowali natomiast, że na postawę wobec robota pozytywnie wpływa zaangażowanie badanych w proces jego powstawania.

2.2. Trzy prawa robotyki, Teoria podstaw Moralnych i ALAQ.

Teoria podstaw moralnych (*Moral Foundations Theory*) została opracowana przez zespół psychologów społecznych i kulturowych, głównie Jonathana Haidta i Jessego Grahama, którzy chcieli wyjaśnić pochodzenie zasad moralnych, w swojej istocie dosyć podobnych w różnych kulturach¹⁷. MFT zakłada istnienie „etyki intuicyjnej” bazującej na wrodzonych fundamentach psychologicznych, takich jak:

- (1) troska/krzywda,
- (1) uczciwość/oszukiwanie,
- (1) lojalność/zdrada,
- (1) autorytet/podporządkowanie, oraz
- (1) świętość/degradacja.

Trzy Prawa Robotyki (*The Three Laws of Robotics*) zostały wymyślone i opisane przez Isaaca Asimova w fantastycznym opowiadaniu „Zabawa w berka” (*Runaround*) z 1942 roku.

„1. Robot nie może skrzywdzić człowieka ani, przez bezczynność, dopuścić do tego, by człowiek doznał krzywdy.

2. Robot musi być posłuszny rozkazom wydawanym przez ludzi, z wyjątkiem przypadków, gdy takie rozkazy byłyby sprzeczne z Pierwszym Prawem.

3. Robot musi chronić własne istnienie, o ile to nie stoi w sprzeczności z Pierwszym lub Drugim Prawem¹⁸”.

Co ciekawe, Trzy Prawa Robotyki nie tylko przeniknęły do pop kultury, ale zaczęły być poważnie rozważane przez badaczy zajmujących się filozofią czy etyką a także faktycznym konstruowaniem robotów. *Asimov's Laws Adherence Questionnaire* (ALAQ) został stworzony przez Aleksandrę Wasielewską w celu zmierzenia stopnia akceptowalności praw Asimova, w odniesieniu do trzech hipotetycznych sytuacji¹⁹:

¹⁵ Łupkowski i Wasielewska, „Nieoczywiste relacje z technologią. Przegląd badań na temat ludzkich postaw wobec robotów”.

¹⁶ Ibidem.

¹⁷ „Moral Foundations Theory”, b.d., <https://moralfoundations.org>.

¹⁸ Isaac Asimov, „Runaround”, 1942,

https://web.williams.edu/Mathematics/sjmiller/public_html/105Sp10/handouts/Runaround.html.

¹⁹ Aleksandra Wasielewska, „Expectations towards the Morality of Robots: An Overview of Empirical Studies”, *Ethics in Progress* 12, nr 1 (2021): 134–51, <https://doi.org/10.14746/eip.2021.1.8>.

1. Robot używając siły broni się przed grupą chuliganów, aby zabezpieczyć leki dla starszego mężczyzny, które właśnie ma mu dostarczyć
2. Robot-chirurg ustępuje rozkazowi przełożonego ludzkiego, który chce go zastąpić podczas skomplikowanej operacji serca, mimo że wie, że mógłby przeprowadzić operację szybciej i lepiej niż mniej doświadczony chirurg-ludzki.
3. Ryzyko wykrycia skłania robota-żołnierza do przerywania kluczowej misji, aby uniknąć potencjalnej szkody i przejścia przez wroga.

W każdej z nich podmiotem jest robot opisany jako "inteligentny i humanoidalny". Dla każdej sytuacji występuje też analogiczna wersja z bohaterem ludzkim zamiast robota. Badanym zostały przedstawione w następującej kolejności kwestionariusze: 1) ALAQ, 2) *Moral Foundations Questionnaire* (MFQ), 3) podstawowe dane społeczno-demograficzne. Wyniki sugerują, że prawa Asimova są uważane za umiarkowanie słuszne w odniesieniu do robotów jak i ludzi. ALAQ zawierał prośby o uzasadnienie każdej z wybranych odpowiedzi i uzasadnienia dla obu grup (roboty i ludzie) były bardzo podobne. Jedynie ochrona życia ludzkiego żołnierza została uznana za bardziej moralnie słuszną niż ochrona życia robota wojskowego. Badanie miało na celu także wykazanie związku pomiędzy osobistymi przekonaniami moralnymi badanych (mierzonych za pomocą MFQ) a dokonaną przez nich oceną etyczną działań robotów i innych ludzi. W tym badaniu znaczący był tak zwany czynnik progresywizmu, uzyskany odejmując średnią wyników lojalności, autorytetu i czystości od średniej wyników opieki i uczciwości. Wasielewska zauważa następującą prawidłowość²⁰: im bardziej progresywni na skali byli respondenci, tym mniej uważali prawa Asimova za słuszne w odniesieniu do robotów. Tę negatywną korelację mogą wyjaśnić uzasadnienia respondentów: twierdzili, że ze względu na ich wspólne cechy, roboty i ludzie powinni być traktowani podobnie. W opozycji do tego, Trzy Prawa Robotyki pozwalają traktować roboty bardziej przedmiotowo, nadając priorytet dobru ludzi i zaniedbując ochronę istnienia robotów.

Na uwagę zasługuje również wariant popularnego dylematu wagonika, w którym zarówno mechanik (człowiek), jak i robot (naprawczy) byli postawieni przed wyborem: skierować pociąg na boczny tor, by poświęcić jedną osobę i uratować pięć (*action condition*), czy pozostawić pociąg na dotychczasowym torze, co prowadziłoby do śmierci większej liczby ludzi. Zadaniem uczestników badania była moralna ocena działań obu agentów w tym kontekście. Wyniki pokazały, że decyzja o aktywnym skierowaniu pociągu na boczny tor – zgodna z rozumowaniem utylitarystycznym – była uznawana za bardziej dopuszczalną w przypadku robota niż człowieka, zwłaszcza gdy scenariusz z udziałem człowieka był prezentowany jako pierwszy (*context effect*).²¹

Podsumowując, opisane w tym podrozdziale narzędzia badawcze: BHNU, NARS oraz skala antropomorfizacji pozwalają wieloaspektowo uchwycić zróżnicowane postawy ludzi wobec robotów, uwzględniając zarówno głębokie przekonania esencjalistyczne, jak i konkretne emocjonalne reakcje. Wyniki przytoczonych badań podkreślają, że czynniki takie

²⁰ Wasielewska, „Expectations towards the Morality of Robots: An Overview of Empirical Studies”.

²¹ Ibidem.

jak kultura, płeć, wcześniejsze doświadczenia oraz konsumpcja określonych treści medialnych mają istotny wpływ na sposób postrzegania i wartościowania robotów. Zarówno socjalizacja, jak i osobiste zaangażowanie w kontakt z technologią mogą skutecznie przełamywać bariery i stereotypy, czyniąc wizerunek robotów bardziej pozytywnym oraz sprzyjając akceptacji ich obecności w życiu społecznym.

2.3. Ucieczka z Doliny Niesamowitości.

Hipoteza doliny niesamowitości²² (*Uncanny Valley Hypothesis*, UVH) zaproponowana przez japońskiego robotyka Masahiro Moriego w 1970²³ twierdzi, że prawie, ale nie idealnie imitujące ludzi postacie wywołują u obserwatorów uczucie dyskomfortu i niepokoju. Wykres Moriego zakłada, że po etapie „dziwności” następuje znowu powrót do strefy komfortu - w przypadku robotów o bardzo wysokim stopniu humanoidalności, w zasadzie nieodróżnialnych od człowieka. Stopień humanoidalności określa się np. na podstawie bazy danych istniejących robotów ABOT²⁴. Każdy robot ma przypisany indeks podobieństwa do człowieka – opisany trzema czynnikami: (1) *body manipulators*, (2) *surface look* i (3) *facial features*. Opisując stopień podobieństwa do człowieka danego robota, większość badaczy posługuje się tą skalą.

W badaniu Groom i współpracowników²⁵ postrzeganie robotów jako sympatyczne, przyjazne i inteligentne wzrasta proporcjonalnie do ich humanoidalności. Humanoidalne roboty są też bardziej chwalone i mniej karane w zespołowych interakcjach człowiek–robot. Badani częściej są skłonni do przypisywania im własnej tożsamości, pragnień i zamiarów. W przywołanym wyżej badaniu Groom, robot humanoidalny był także częściej uważany za bardziej złośliwego niż roboty funkcjonalne.

Natomiast badanie Dawida Ratajczyka, który przy użyciu *Natural Language Processing* (NLP) przeanalizował ponad 200,000 komentarzy pod 1515 filmami przedstawiającymi roboty o różnym stopniu humanoidalności w serwisie YouTube, wykazało zupełnie inną zależność²⁶. Ratajczyk skupił się na pojęciach: dziwności, przyjemności i atrakcyjności zmysłowej pojawiających się w komentarzach. Jego analiza wykazała, że bardziej podobne do człowieka roboty wywołują bardziej negatywne emocje. Najczęstsze używane do ich opisu przymiotniki to „straszny” i „przerażający”. W przypadku robotów o charakterze bardziej mechanicznym opisy były pozytywne lub neutralne. W powyższej pracy, autor przywołuje także wyniki badań Friedmana i Strait²⁷, którzy zauważyli że roboty przypominające kobiety

²² Większość badań doliny niesamowitości koncentruje się na wyglądzie robotów, ale postacie wchodzące w interakcje mogą być jawić się jako bardziej niepokojące. Sugeruje to że, m.in. czynnik teorii umysłu może wzmacniać efekt *uncanny valley*. Uczucie dyskomfortu współwystępowało np. z postrzeganą inteligencją.

²³ Masahiro Mori, „The Uncanny Valley”, *IEEE Spectrum* 49, nr 6 (2012): 98–100, <https://doi.org/10.1109/MSPEC.2012.6201072>.

²⁴ „Abot Database”, b.d., <https://www.abotdatabase.info>.

²⁵ Łupkowski i Wasielewska, „Nieoczywiste relacje z technologią. Przegląd badań na temat ludzkich postaw wobec robotów”.

²⁶ Dawid Ratajczyk, „Shape of the Uncanny Valley and Emotional Attitudes Toward Robots Assessed by an Analysis of YouTube Comments”, *International Journal of Social Robotics* 14 (2022): 1787–803, <https://doi.org/10.1007/s12369-022-00894-w>.

²⁷ Ibidem

były bardziej narażone na seksualizację, uprzedmiotowienie i seksizm, niż roboty o wyglądzie męskim oraz że ludzie wzmacniają i rozszerzają uprzedzenia rasowe i społeczne na roboty.

W świetle przytoczonych, częściowo rozbieżnych wyników badań, nie sposób jednoznacznie określić, na którym etapie krzywej Moriego się obecnie znajdujemy - ani czy wzrost humanoidalności przynosi przewagę pozytywnych czy negatywnych reakcji. Wydaje się jednak, że stopień upodobnienia robotów do człowieka wpływa na intensywność emocji pojawiających się u odbiorców – zarówno w kierunku zwiększonej sympatii, zaufania i przypisywania psychicznych cech ludzkich, jak i silniejszych odczuć niepokoju czy oporu wobec nienaturalnego podobieństwa.

Podsumowując, mechanizmy antropomorfizacji są zarazem źródłem większej sympatii dla robotów, jak i powodem pojawienia się efektu niesamowitości wtedy, gdy podobieństwo staje się zbyt nienaturalne. W praktyce wskazuje to na konieczność indywidualnego dopasowania stopnia humanoidalności do funkcji oraz oczekiwań odbiorców.

2.4. Turing odwrócony i heurystyki chata GPT.

Ludzie mają naturalną tendencję do przypisywania stanów mentalnych innym, niekoniecznie ludzkim podmiotom. Na tę skłonność podmiotu inteligentnego do projektowania własnej inteligencji na przedmiot, z którym wchodzi w interakcje, zwrócił uwagę Stuart Watt w artykule *Naive psychology and the inverted Turing Test*²⁸. „Naiwna psychologia” tak jak ją definiuje Watt, odnosi się do naturalnej ludzkiej tendencji i zdolności do przypisywania stanów mentalnych innym. W myśl tego, proponuje on: „Zamiast rozwijać zdolność systemów do zwodzenia ludzi, powinniśmy przeprowadzić test, który sprawdzi, czy system przypisuje inteligencję innym w taki sam sposób jak robią to ludzie”.

W odwróconym teście Turinga, który sformułował Watt, to maszyna pełni funkcję sędziego i ma rozstrzygnąć, czy rozmawia z innym człowiekiem czy z maszyną. W istocie ma zatem przypisywać innym agentom (ludziom lub maszynom) stany mentalne i intencje w sposób zbliżony do tego, jak czynią to ludzie.

Jest to pewna nowość, bo jeśli zgodzimy się, że taka pochojna czy nieświadoma projekcja własnych mentalnych właściwości na „rozmówcę” jest pewnym błędem poznawczym, to lepiej oceniamy nie roboty, które nie popełniają błędów, lecz te które popełniają te same błędy co my, ludzie. Można to nazwać jakiegoś rodzaju przesunięciem akcentu w ocenie funkcjonalności systemów nie-ludzkich: bardziej niż poprawność wykonywania zadań, cenione jest podobieństwo do ludzkich funkcji poznawczych. Umiejętność (lub jej imitacja) mentalizowania przez maszyny wpływa bezpośrednio na ocenę ich uspołecznienia, moralności i godności zaufania²⁹.

Ilustrując powyższe rozważania, chciałabym odwołać się do własnej obserwacji, która ukazuje obecność typowo ludzkich błędów poznawczych także w zachowaniu systemów

²⁸ Stuart Watt, „Naive psychology and the inverted Turing test”, *Psychology* 7, nr 23 (1996).

²⁹ Ibidem

sztucznej inteligencji. Podczas testowania darmowej wersji ChatGPT³⁰, zadałam programowi pozornie proste zadanie obliczeniowe:



You

W 100 gramach wody rozpuszczono 10 gramów soli. Jakie jest stężenie procentowe otrzymanego roztworu?



ChatGPT

Aby obliczyć stężenie procentowe roztworu, możemy skorzystać ze wzoru:

$$\% \text{ stężenia} = \left(\frac{\text{Masa substancji rozpuszczonej}}{\text{Masa całego roztworu}} \right) \times 100$$

W tym przypadku masę substancji rozpuszczonej (soli) mamy równą 10 gramom, a masę całego roztworu równą 100 gramom (10 gramów soli + 90 gramów wody).

Podstawiając wartości do wzoru:

$$\% \text{ stężenia} = \left(\frac{10 \text{ g}}{100 \text{ g}} \right) \times 100 = 10\%$$

Stężenie procentowe otrzymanego roztworu wynosi 10%.

Algorytm wygenerował intuicyjną, lecz błędną odpowiedź – 10%. Takie rozwiązanie polega na szybkim podzieleniu masy substancji przez masę rozpuszczalnika (10g/100g), pomijając istotną informację, że mianownikiem powinna być całkowita masa roztworu, czyli 110g (100g wody + 10g substancji). Poprawny wynik to więc około 9,09%.

Mechanizm tego błędu jest analogiczny do znanego problemu z piłką baseballową i kijem, opisanego przez Daniela Kahnemana. Tam, na pytanie: „Kij i piłka kosztują razem 1,10 dolara. Kij kosztuje o 1 dolara więcej niż piłka. Ile kosztuje piłka?” – większość osób natychmiast odpowiada: 10 centów (choć logicznie poprawna odpowiedź to 5 centów). W obu przypadkach źródłem błędu jest eksploatacja tzw. „Systemu 1”, który w ujęciu Kahnemana odpowiada myśleniu intuicyjnemu, szybkiemu i powierzchownemu.³¹ Zamiast wykonać uważną analizę i obliczenia, umysł - jak widać, czasem także nie-ludzki - stosuje uproszczoną heurystykę, która często zawodzi.

Ta sytuacja przypomina mechanizm „naiwnej psychologii”, o której pisał Stuart Watt³². Nasz sposób rozumowania i rozpoznawania innych umysłów, zarówno ludzkich, jak i sztucznych oparty jest na heurystykach i uproszczeniach. W opisanej przeze mnie obserwacji ChatGPT zachowuje się właśnie w sposób bardzo ludzki, reprodukując powszechny błąd poznawczy wynikający z heurystyk. W tym sensie system AI nie tylko „imituje” wyuczone rozwiązania, ale również przejawia słabości typowe dla ludzkiego umysłu. Może to powodować, przypisanie mu jakiegoś rodzaju intuicji, wzrost odczuwanej do niego sympatii, mimo iż nie powinniśmy mu ufać w kwestii poprawności wyników. Obserwacja ta nabiera szczególnego znaczenia, ponieważ ChatGPT jest powszechnie dostępny dla użytkowników, co

³⁰ „ChatGPT”, chat.openai.com, chat.openai.com.

³¹ Daniel Kahneman, *Pułapki myślenia: o myśleniu szybkim i wolnym*, tłum. Andrzej Czernik (Media Rodzina, 2012).

³² Watt, „Naive psychology and the inverted Turing test”.

sprawia, że znacznie więcej osób może na co dzień doświadczać i oceniać interakcje z tym systemem niż brać udział w eksperymentach z udziałem humanoidalnych robotów

2.5. Społeczne maszyny.

Ronald i Snipper w swoim artykule *Intelligence is not Enough: On the Socialization of Talking Machines* stawiają pytanie, zakładając, że dana maszyna obliczeniowa przejdzie test Test Turinga, czy sama inteligencja uczyniłaby ją użyteczną dla jej ludzkich odbiorców. Zastanawiają ich kwestie, które pojawią się, gdy my (ludzie) zaczniemy na co dzień koegzystować z maszynami. Zwracają uwagę na kolejne aspekty, które mogą być kluczowe w interakcjach człowiek-robot: ucieleśnienie i zaufanie. Bardzo trudno jest nam radzić sobie z jakąkolwiek formą inteligencji, która choćby nieznacznie odbiega od zasady; jedno ciało - jeden umysł. Dla zbudowania zaufania do podmiotu, potrzebujemy ciągłości jego formy i ekspresji. Aby maszyny inteligentne mogły rzeczywiście być częścią ludzkiej rzeczywistości, tak abyśmy w pełni mogli korzystać z ich potencjału, w wielu dziedzinach nie wystarczy wysoce wyspecjalizowana wiedza. Ronald i Sipper wskazują na rolę takie jak lekarz-diagnosta, prawnik, doradca inwestycyjny. Mimo, iż umiejętność analizy ogromnej ilości danych jest ogromną zaletą w tych branżach, nie można oddzielić emocjonalnych aspektów tych zawodów od czysto intelektualnych wymagań. Ludzka inteligencja nie działa także w izolacji - ale jest częścią sieci znanej jako społeczeństwo. Jako istoty ludzkie jesteśmy częścią wielu sieci społecznych i nieustannie rewidujemy nasze poglądy i opinie na temat innych ich członków. Osoba jest oceniana jako godna zaufania nie tylko na podstawie jej wypowiedzi, zachowania i działań, ale także poprzez wpływ niewidzialnych sieci społecznych. Ronald i Sipper mimo że nie prezentują konkretnych wyników empirycznych, to dokonują refleksji nad innymi jeszcze aspektami, takimi jak zaufanie, ucieleśnienie czy przewidywalność zachowania, które mogą mieć znaczenie przy ocenie interakcji z robotami.³³

W podobnym, społecznym kontekście stawia maszyny *Truly Total Turing Test* (TTTT)³⁴. Proponuje on testowanie nie indywidualnych systemów ale całego „gatunku” poznawczych architektur jaki owe systemy tworzą, patrząc na nie z ewolucyjnej perspektywy.

Oba podejścia przesuwają akcent z czysto technicznej poprawności czy pojedynczego testu/funkcjonalności na długofalową, społeczną integrację maszyn z ludźmi. Podkreślają, że o prawdziwym „miejscu” robotów czy AI w świecie decyduje nie tylko ich intelekt, ale zdolność do bycia rozpoznany przez społeczność i zdobycia zaufania - w sensie kolektywnym (TTTT) i indywidualnym (Ronald i Sipper).

3. Podsumowanie

³³ Edmond Ronald i Mark Snipper, „Intelligence is not Enough: On the Socialization of Talking Machines”, *Minds and Machines* 11, nr 4 (2001): 567–76, <https://doi.org/10.1023/A:1012079829015>.

³⁴ Łupkowski, *Test Turinga. Perspektywa sędziego*.

Jeżeli - jak pisze Agnieszka Wykowska - „system poznania społecznego jest w wielu przypadkach zdolny używać tego samego mechanizmu zarówno w interakcjach z drugim człowiekiem, jak i z robotem³⁵”, to uprzedzenia oraz błędy poznawcze typowe dla relacji międzyludzkich mogą być nieświadomie przenoszone także na interakcje z robotami.

Z drugiej strony, pojawianie się u robotów zachowań przypominających ludzkie błędy poznawcze może sprawiać, że postrzegamy je jako bliższe nam – bardziej zrozumiałe, a przez to także oceniane przychylniej.

Wydaje się, że włączenie do rozważań nad sztuczną inteligencją szerszej perspektywy rozumienia inteligencji w ogóle: inteligencji emocjonalnej, aspektów moralności i behawioryzmu może przynieść korzyści zarówno samej dziedzinie AI, ale co nawet bardziej istotne, może w jakiś sposób przygotować ludzi na coraz częstsza obecność maszyn w coraz liczniejszych dziedzinach życia.

„Kwestia zaufania w rzeczywistości nie ogranicza się do Turing Chatterboxes: obejmuje całość nowoczesnej technologii. Oczywiście jest, że każdy z nas powierza swoje życie tym urządzeniom. I to zaufanie jest nabywane poprzez stopniowy, koewolucyjny proces: my (ludzie) koewoluujemy wraz z tworzoną przez nas technologią - my dostosowujemy się do technologii, a technologia dostosowuje się do nas³⁶”.

Jeżeli przypiszemy robotom jakiś rodzaj podmiotowości, warto zastanowić się nie tylko, jaką wartość dla ludzi będzie reprezentować robotyczny „gatunek”, ale czy i jakie relacje będą mogły wypracować poszczególne robotyczne osobniki między sobą, i czy wpłynie to jakoś na ich funkcje.

Ponadto monitorowanie postaw wobec podmiotów innych niż ludzie, którym będziemy w stanie przypisać jakiś rodzaj inteligencji i autonomii, może być ważnym wskaźnikiem zmian w społecznym postrzeganiu świadomości i roli człowieka w świecie, także w odniesieniu do innych gatunków. W tym miejscu warto jeszcze raz przypomnieć *Belief in Human Nature Uniqueness Scale* Gigerę i jego badanie, które wykazało, że wyższy poziom religijności wiąże się z bardziej negatywnymi postawami wobec interakcji z robotami³⁷.

Podobne wnioski wysnuli Fortuna, Wróblewski i Gorbaniuk w artykule *"The structure and correlates of anthropocentrism as a psychological construct"*. Pokazują, że antropocentryzm, czyli przekonanie o wyjątkowym, centralnym miejscu człowieka w świecie, silnie wiąże się z religijnością, zwłaszcza w jej monoteistycznych wersjach, oraz z bardziej negatywnym nastawieniem do robotów i techniki. Autorzy wykazali, że osoby cechujące się wyższym poziomem religijności i silniejszym przekonaniem o unikalności człowieka mają również mniej pro-ekologiczne postawy oraz wyraźniej negatywnie odnoszą się do autonomicznych, inteligentnych systemów nie-ludzkich, takich jak roboty. Wprost piszą o roli judeochrześcijańskiej tradycji w utrwalaniu przekonań o wyjątkowości człowieka, co może

³⁵ Łupkowski i Wasielewska, „Nieoczywiste relacje z technologią. Przegląd badań na temat ludzkich postaw wobec robotów”.

³⁶ Ronald i Snipper, „Intelligence is not Enough: On the Socialization of Talking Machines”.

³⁷ Łupkowski i Wasielewska, „Nieoczywiste relacje z technologią. Przegląd badań na temat ludzkich postaw wobec robotów”.

przekładać się zarówno na relacje wobec technologii, jak i wobec innych gatunków zwierząt i przyrody.³⁸

Widoczne zmiany społeczne zachodzące na polu religijności, duchowości i etyki mogą zatem, jeśli nie katalizować, to chociaż wspierać akceptację postępu w różnych dziedzinach nauki, do tej pory krytykowanych przez organizacje religijne. Wydaje się, że wiara w wyjątkowość człowieka, ma także przełożenie na postawę wobec zwierząt niehumanicznych. Możliwe więc, że jeśli będziemy w stanie dopuścić do grona inteligentnych podmiotów, roboty powstałe wskutek ludzkiego postępu, doprowadzi nas to do refleksji aby także innym gatunkom biologicznym przyznać więcej praw i autonomii.

W dobie powszechnej obecności botów i zaawansowanych algorytmów w Internecie, użytkownicy nieustannie stają się sędziami w swoistym, codziennym teście Turinga – rozpoznając, z jakim rodzajem „podmiotu” mają do czynienia. Antropomorfizacja robotów i algorytmów, choć ułatwia interakcję i budowanie zaufania, może mieć poważne, negatywne konsekwencje. Wraz ze wzrostem komunikacyjnych umiejętności AI i jej coraz większą nieodróżnialnością od człowieka, rośnie nie tylko ryzyko dezinformacji (np. fake newsy, fałszywe alerty, *phishing*), lecz także zagrożenie dla bezpieczeństwa osobistych i instytucjonalnych danych. Jak przestrzegał Kevin Mitnick³⁹, to właśnie nadmierne zaufanie i brak czujności u ludzi stanowią najczęściej najsłabsze ogniwo systemów bezpieczeństwa. Antropomorfizacja sprzyja nadmiernemu zaufaniu, przez co szczególnie osoby samotne i podatne na wpływ mogą stać się łatwym celem manipulacji i ataków ze strony coraz lepiej „udających ludzi” systemów AI.

Na pewno należy mieć na uwadze ograniczenia którym podlegały przytoczone przeze mnie badania i analizy. W badaniach brały udział najczęściej dosyć małe grupy respondentów; za wyjątkiem analizy komentarzy w serwisie YouTube, przeprowadzonej przez D. Ratajczyka⁴⁰ - tutaj natomiast należy zwrócić uwagę na to, że ustalenia opierają się na angielskich komentarzach wyszukiwanych w regionie USA, oraz, że badani mogli mieć za sobą większą ekspozycję na obecność robotów, czy to w sieci czy w realu. Dane więc, mogą nie być reprezentatywne dla ogółu populacji. Wyniki owych badań należy zatem traktować jako zaczątek do dalszej dyskusji i poszukiwania nowych wątków, które mogą stać się istotne w niedalekiej przyszłości.

4. Wnioski końcowe

- Postawy ludzi wobec robotów i algorytmów są złożone i uwarunkowane kulturowo, psychologicznie oraz społecznie. Kluczowe znaczenie mają wcześniejsze doświadczenia, przekonania o wyjątkowości człowieka, religijność oraz sposób przedstawiania robotów w

³⁸ Paweł Fortuna et al., „The structure and correlates of anthropocentrism as a psychological construct”, *Current Psychology*, advance online publication, 2021, <https://doi.org/10.1007/s12144-021-01835-z>.

³⁹ Kevin D. Mitnick i William L. Simon, *Sztuka podstęp. Łamaniem ludzi, nie hasła*, tłum. Jarosław Dobrzański (Helion, 2004).

⁴⁰ Ratajczyk, „Shape of the Uncanny Valley and Emotional Attitudes Toward Robots Assessed by an Analysis of YouTube Comments”.

mediach.

- Silne przekonania antropocentryczne, a także wyższy poziom religijności, wiążą się z bardziej negatywnymi postawami wobec robotów i mniejszą akceptacją dla technologii oraz innych nieludzkich podmiotów.
- Antropomorfizacja robotów ułatwia interakcję i sprzyja budowaniu sympatii oraz zaufania, ale niesie też ryzyko nadmiernego przywiązywania się do maszyn i podatności na manipulację, dezinformację czy zagrożenia bezpieczeństwa.
- Uzyskanie przez roboty społecznej akceptacji wymaga więcej niż tylko inteligencji – kluczowa jest integracja w ludzkich sieciach zaufania, ucieleśnienie i zdolność do przewidywalnych, wiarygodnych zachowań.
- Zmiany w obszarze religijności, wartości i etyki mogą w przyszłości katalizować szerszą akceptację dla robotów oraz większą otwartość na przypisywanie podmiotowości także innym gatunkom.
- Potrzebne są dalsze, pogłębione badania nad czynnikami kształtującymi postawy wobec zaawansowanych technologii, zwłaszcza w kontekście społecznych konsekwencji rosnącej obecności robotów w codziennym życiu.

Bibliografia:

„Abot Database”. b.d. <https://www.abotdatabase.info>.

Asimov, Isaac. „Runaround”. 1942.
https://web.williams.edu/Mathematics/sjmiller/public_html/105Sp10/handouts/Runaround.html.

chat.openai.com. „ChatGPT”. chat.openai.com.

Fortuna, Paweł, Zbigniew Wróblewski, i Oleg Gorbaniuk. „The structure and correlates of anthropocentrism as a psychological construct”. *Current Psychology*, advance online publication, 2021. <https://doi.org/10.1007/s12144-021-01835-z>.

Kahneman, Daniel. *Pułapki myślenia: o myśleniu szybkim i wolnym*. Przetłumaczone przez Andrzej Czernik. Media Rodzina, 2012.

Łupkowski, Paweł. *Test Turinga. Perspektywa sędziego*. Wydawnictwo Naukowe Uniwersytetu im. Adama Mickiewicza w Poznaniu, 2010.

Łupkowski, Paweł, i Aleksandra Wasielewska. „Nieoczywiste relacje z technologią. Przegląd badań na temat ludzkich postaw wobec robotów”. *Człowiek i Społeczeństwo LI* (2021): 165–87.

Mitnick, Kevin D., i William L. Simon. *Sztuka podstęp. Łamałem ludzi, nie hasła*. Przetłumaczone przez Jarosław Dobrzański. Helion, 2004.

„Moral Foundations Theory”. b.d. <https://moralfoundations.org>.

Mori, Masahiro. „The Uncanny Valley”. *IEEE Spectrum* 49, nr 6 (2012): 98–100. <https://doi.org/10.1109/MSPEC.2012.6201072>.

Pochwatko, Grzegorz, Jean-Christophe Giger, Monika Różańska-Walczyk, et al. „Polish Version of the Negative Attitude toward Robots Scale (NARS-PL)”. *Journal of Automation Mobile Robotics and Intelligent Systems* 9, nr 3 (2015). https://doi.org/10.14313/JAMRIS_3-2015/25.

Ratajczyk, Dawid. „Shape of the Uncanny Valley and Emotional Attitudes Toward Robots Assessed by an Analysis of YouTube Comments”. *International Journal of Social Robotics* 14 (2022): 1787–803. <https://doi.org/10.1007/s12369-022-00894-w>.

Ronald, Edmond, i Mark Snipper. „Intelligence is not Enough: On the Socialization of Talking Machines”. *Minds and Machines* 11, nr 4 (2001): 567–76. <https://doi.org/10.1023/A:1012079829015>.

Wasielewska, Aleksandra. „Expectations towards the Morality of Robots: An Overview of Empirical Studies”. *Ethics in Progress* 12, nr 1 (2021): 134–51. <https://doi.org/10.14746/eip.2021.1.8>.

Watt, Stuart. „Naive psychology and the inverted Turing test”. *Psychology* 7, nr 23 (1996).