

Barbara Drapikowska¹, Robert Zakrzewski²

Akademia Sztuki Wojennej w Warszawie

https://doi.org/10.29316/9788368103304_9

ZAUFANIE I LĘK WOBEC SZTUCZNEJ INTELIGENCJI – PSYCHOLOGICZNE MECHANIZMY A ODPOWIEDZIALNE WDRAŻANIE TECHNOLOGII

TRUST AND FEAR OF ARTIFICIAL INTELLIGENCE – PSYCHOLOGICAL MECHANISMS AND RESPONSIBLE IMPLEMENTATION OF TECHNOLOGY

Streszczenie

Celem rozdziału jest przedstawienie psychologicznych mechanizmów determinujących postawy wobec sztucznej inteligencji oraz określenie ich konsekwencji dla odpowiedzialnego i bezpiecznego wdrażania technologii w życie społeczne. Problem badawczy koncentruje się wokół pytania: jakie mechanizmy psychologiczne determinują postawy wobec AI oraz wpływają na odpowiedzialne i bezpieczne wdrażanie tej technologii w kontekście społecznym? Opracowanie ma charakter teoretyczny i opiera się na analizie aktualnej literatury z zakresu psychologii społecznej, poznawczej oraz interakcji człowiek – technologia.

¹ <https://orcid.org/0000-0002-6852-2791>, e-mail: b.drapikowska@akademia.mil.pl

² <https://orcid.org/0000-0002-4019-9048>, e-mail: r.zakrzewski@akademia.mil.pl

Przeprowadzone rozważania wskazują, że zaufanie i lęk wobec AI mają wspólne źródła psychologiczne, lecz mogą prowadzić do przeciwstawnych postaw, tj. od gotowości do współpracy po unikanie technologii. Zrozumienie tych procesów jest kluczowe dla wspierania odpowiedzialnego wdrażania sztucznej inteligencji w społeczeństwie.

Słowa kluczowe: sztuczna inteligencja (AI), antropomorfizacja, heurystyki poznawcze, psychologia technologii

Abstract

The aim of this chapter is to present the psychological mechanisms that determine attitudes toward artificial intelligence and to define their consequences for the responsible and safe implementation of technology in social life. The research problem is: What psychological mechanisms determine attitudes toward AI and influence the responsible, safe implementation of this technology in a social context? The study is theoretical and based on an analysis of current literature in social and cognitive psychology and human-technology interaction. The considerations presented indicate that trust and fear of AI have common psychological sources, but can lead to opposing attitudes, i.e., from a willingness to cooperate to avoidance of technology. Understanding these processes is key to supporting the responsible implementation of artificial intelligence in society.

Keywords: artificial intelligence (AI), anthropomorphism, cognitive heuristics, technology psychology

Wstęp

Dynamiczny rozwój technologii sztucznej inteligencji (AI) oraz jej integracja z codziennymi praktykami społecznymi, obejmująca tak zróżnicowane obszary jak zdrowie, bezpieczeństwo, edukacja i dostęp do usług, rodzi szereg pytań dotyczących sposobu, w jaki ludzie postrzegają i interpretują działanie systemów autonomicznych (West, Whittaker, Crawford, 2019). Użytkownicy często przypisują systemom AI cechy psychiczne, tj. intencje czy emocje, w procesie określanym jako antropomorfizacja. Zjawisko to wpływa zarówno na emocjonalne reakcje wobec maszyn, jak i na ocenę ich kompetencji oraz społecznego znaczenia (Złotowski, Proudfoot, Yogeeswaran, Bartneck, 2015).

Dynamiczny rozwój technologii sztucznej inteligencji oraz jej integracja z codziennymi praktykami społecznymi rodzi szereg pytań dotyczących sposobu, w jaki ludzie postrzegają i interpretują działanie systemów autonomicznych (West i in., 2019). W konsekwencji relacja człowiek – AI staje się obszarem silnego napięcia poznawczego i normatywnego. AI zaczyna być postrzegana nie tylko jako narzędzie, ale także jako potencjalny „aktor społeczny”, wpływający na decyzje dotyczące obywateli i instytucji (Waytz, Cacioppo, Epley, 2010).

Antropomorfizacja może prowadzić do nadmiernego zaufania wobec systemów sztucznej inteligencji, nawet wówczas, gdy brak jest podstaw do zakładania ich intencjonalności. Przypisywanie AI podmiotowości może tworzyć fałszywe przekonanie o odpowiedzialności moralnej maszyn, co przesuwając ten ciężar z projektantów i operatorów na technologie (Bryson, 2018). Jednocześnie ludzie mają tendencję do postrzegania algorytmów jako bardziej obiektywnych i kompetentnych niż ludzie (Logg, Minson, Moore, 2019), nawet wówczas, kiedy nie potrafią ocenić poprawności ich działania.

Obecnie, kiedy systemy AI są integrowane z medycyną, edukacją, bezpieczeństwem publicznym czy rynkiem pracy, ich społeczna rola ulega gruntownej transformacji. Automatyczne podejmowanie decyzji w obszarach wcześniej zarezerwowanych dla człowieka wpływa na to, jak ludzie rozumieją sprawstwo, odpowiedzialność i kompetencje technologii. Coraz częściej podkreśla się, że stosunek ludzi do technologii zależy nie tylko od jej użyteczności, lecz także od potrzeb psychologicznych, takich jak poczucie bezpieczeństwa, przewidywalności czy kontroli. Postawy wobec AI mogą wynikać zarówno z realnych doświadczeń, jak i z przekazów medialnych oraz narracji kulturowych wzmacniających obawy przed utratą pracy, autonomii lub prywatności. Ambiwalencja ta sprawia, że ludzie jednocześnie oczekują od AI perfekcyjnego działania i obawiają się jej dominacji.

Systemy AI projektowane z cechami humanoidalnymi mogą wywoływać jeszcze silniejsze reakcje społeczne, tj. od sympatii i opiekuńczości po lęk przed utratą kontroli. Elementy formy i ruchu robotów automatycznie nasilają przypisywanie im cech mentalnych, co może prowadzić również do nadawania im osobowości czy wartości moralnej (Złotowski i in., 2015). W efekcie antropomorfizacja zaczyna wpływać nie tylko na ocenę użyteczności AI, lecz także na jej status nie jako technologii, lecz jako potencjalnego uczestnika życia społecznego.

W konsekwencji relacja człowiek – AI staje się obszarem silnego napięcia poznawczego i normatywnego. AI zaczyna być postrzegana nie tylko jako narzędzie, ale także jako potencjalny „aktor społeczny”, wpływający na decyzje dotyczące obywateli i instytucji (West i in., 2019). Dlatego właśnie psychologia staje przed wyzwaniem opracowania ram teoretycznych pozwalających wyjaśnić, jakie procesy poznawcze i emocjonalne prowadzą do akceptacji lub odrzucenia technologii o wysokiej autonomii.

Zrozumienie tych procesów jest niezbędne nie tylko dla poszerzenia wiedzy teoretycznej, lecz także dla projektowania systemów, które będą budziły adekwatny poziom zaufania społecznego, jednocześnie nie wzmacniając nierealistycznych oczekiwań czy obaw.

Celem niniejszego rozdziału jest przedstawienie psychologicznych mechanizmów determinujących postawy wobec sztucznej inteligencji oraz określenie ich konsekwencji dla odpowiedzialnego i bezpiecznego wdrażania technologii w życie społeczne. Główny problem badawczy został skonstruowany w postaci pytania: jakie mechanizmy psychologiczne determinują postawy wobec AI oraz wpływają na odpowiedzialne i bezpieczne wdrażanie tej technologii w kontekście społecznym?

Na tej podstawie skonstruowano problematykę szczegółową obejmującą następujące kwestie:

1. Jakie są psychologiczne mechanizmy antropomorfizacji i zaufania do AI?
2. Jakie są psychologiczne mechanizmy lęku przed AI?
3. Jakie są psychologiczne konsekwencje oraz praktyczne implikacje dla projektowania zaufanej AI?

W kolejnych częściach rozdziału podjęto próbę rozwiązania przyjętych problemów badawczych poprzez analizę teoretyczną oraz prezentację wyników badań własnych.

Psychologiczne mechanizmy antropomorfizacji i zaufania do AI

W interakcjach człowieka z systemami sztucznej inteligencji istotną rolę odgrywają uproszczone mechanizmy poznawcze, które z jednej strony ułatwiają szybkie podejmowanie decyzji, lecz jednocześnie mogą prowadzić do błędnych ocen działania technologii. Tversky i Kahneman (1974) wykazali, że ludzie,

podejmując decyzje w warunkach niepewności, opierają się na heurystykach³, tj. automatycznych skrótach myślowych, takich jak heurystyka reprezentatywności, dostępności czy zakotwiczenia. Mechanizmy te znajdują swoje odzwierciedlenie również w ocenach kompetencji AI. Innymi słowy, użytkownicy mogą ufać systemom informatycznym, ponieważ „wydają się” dokładne i obiektywne, nawet jeśli nie rozumieją zasad ich działania.

Lee i See (2004) podkreślają, że zaufanie do automatyzacji jest funkcją trzech czynników, tj. postrzeganej kompetencji, przewidywalności oraz intencji systemu. W praktyce oznacza to, że jeśli technologia szybko i poprawnie dostarcza odpowiedzi, użytkownik może ulegać iluzji niezawodności, nawet przy braku transparentności działania czynności algorytmu. W sytuacji błędu ludzie nierzadko reagują skrajnie, np. nagła utrata zaufania prowadzi do całkowitego odrzucenia systemu, mimo że wcześniejsze wyniki były korzystne.

Badania Dietvorsta i innych (2015) pokazują, że wystarczy jedna widoczna pomyłka algorytmu, aby zaufanie użytkowników spadło dramatycznie, nawet jeśli system statystycznie pozostaje dokładniejszy niż człowiek. Autorzy wskazują, że ludzie surowiej oceniają błędy popełnione przez algorytmy niż przez ludzi. Efekt ten ilustruje szersze zjawisko – mianowicie, decyzje dotyczące akceptacji technologii częściej opieramy na ocenie emocjonalnej, a nie na racjonalnym rachunku skuteczności.

Równocześnie istnieją również tendencje odwrotne. W niektórych kontekstach obserwuje się tendencję do bezkrytycznego polegania na systemach automatycznych (tzw. *automation bias*) nawet wtedy, gdy użytkownik ma możliwość zakwestionowania wyników (Logg, 2019). Oznacza to, że heurystyki mogą

³ Heurystyka – uproszczona reguła poznawcza lub strategia myślenia, którą ludzie stosują, by podejmować szybkie decyzje lub wyciągać wnioski w sytuacjach niepewności lub ograniczonej informacji. Heurystyki zmniejszają obciążenie poznawcze, ale mogą prowadzić do systematycznych błędów (np. przewidywania, szacowania prawdopodobieństwa). Heurystyka reprezentatywności polega na tym, że oceniamy prawdopodobieństwo zdarzenia na podstawie tego, jak bardzo przypomina ono nasz prototyp w danej kategorii (zamiast na podstawie realnych danych statystycznych). Heurystyka dostępności polega na tym, że oceniamy częstotliwość lub ryzyko zdarzeń na podstawie tego, jak łatwo przychodzą nam na myśl przykłady, a nie na podstawie rzeczywistych statystyk. Heurystyka zakotwiczenia polega na tym, że przy podejmowaniu decyzji opieramy się na pierwszej dostępnej informacji (kotwicy), nawet jeśli jest przypadkowa lub niewiarygodna (Wojciszke, 2006).

prowadzić zarówno do nadmiernego polegania na AI, jak i do nagłego odrzucenia technologii, zależnie od doświadczeń użytkownika i charakteru zadania.

Należy zatem zaznaczyć, że decyzje dotyczące korzystania z AI nie są wynikiem chłodnego namysłu, lecz są warunkowane przez poznawcze skróty i uproszczenia. To sprawia, że zaufanie społeczne wobec sztucznej inteligencji jest zmienne, podatne na błędy oceny, a także podatne na manipulację, np. przez design interfejsu lub sposób prezentacji wyników działania systemu.

Jednym z kluczowych czynników wpływających na zaufanie do systemów sztucznej inteligencji jest antropomorfizacja, rozumiana jako przypisywanie cech ludzkich podmiotom nieożywionym. Ludzie wykazują naturalną tendencję do interpretowania zachowań technologii poprzez pryzmat kategorii psychologicznych, tj. intencji, emocji czy motywacji. Mechanizm ten ułatwia przewidywanie działania AI, a zarazem wzmacnia poczucie bezpieczeństwa w interakcji z nią (Waytz i in., 2010). W tym sensie antropomorfizacja pełni funkcję poznawczą i społeczną, czyli pozwala traktować AI jak partnera, a nie skomplikowany system informatyczny.

Badania pokazują, że elementy wyglądu i zachowania maszyn, takie jak humanoidalna forma, płynne ruchy czy naturalnie brzmiący głos, zwiększają skłonność użytkowników do przypisywania robotom stanów mentalnych. Złotowski i inni (2015) podkreślają, że już minimalne sygnały społeczne mogą automatycznie uruchamiać procesy mentalizacji, w wyniku których AI postrzegana jest jako bardziej kompetentna i godna zaufania. Co istotne, antropomorfizacja może skutkować również nadawaniem technologii elementów osobowości oraz wartości moralnej, co stanowi podstawę do traktowania jej jako „aktora społecznego”.

Zaufanie budowane na antropomorficznych cechach AI niesie jednak ryzyko nadmiernego polegania na systemach, które nie posiadają realnych zdolności intencjonalnych. Bryson (2018) ostrzega, że traktowanie AI jako bytu moralnego jest nie tylko błędem poznawczym, lecz także może prowadzić do niebezpiecznego „przesuwania” odpowiedzialności z ludzi na technologie. Ludzie przypisują maszynom autonomię tam, gdzie w rzeczywistości decyzje pozostają wynikiem implementacji określonych algorytmów.

Co więcej, wysoki poziom antropomorfizacji nie zawsze sprzyja zaufaniu. Nadmiernie „ludzka” forma może wywoływać tzw. efekt niesamowitości (ang. *uncanny valley*), prowadząc do dyskomfortu i dystansu społecznego (Złotowski

i in., 2015). W takich przypadkach użytkownicy mogą reagować niechęcią lub wręcz nieufnością wobec AI, obawiając się naruszenia granic między tym, co ludzkie, a tym, co technologiczne. Efekt *uncanny valley* opisuje gwałtowny spadek akceptacji oraz wzrost niepokoju wobec robotów lub awatarów, które wyglądają niemal jak ludzie, lecz nadal są odbierane jako nienaturalne. W takich sytuacjach silne podobieństwo przyciąga uwagę, ale jednocześnie wywołuje niezgodność percepcyjną, prowadząc do lęku i odrzucenia technologii (Mori, 2012).

W badaniach C. Christoforakos i innych (2021) wykazano, że zaufanie wzrasta, gdy roboty oceniane są jako jednocześnie kompetentne i „ciepłe”, czyli budzące pozytywne skojarzenia społeczne. Innymi słowy, skuteczność technologii nie wystarcza, gdyż użytkownik musi jeszcze postrzegać ją jako przyjazną i ukierunkowaną na dobro człowieka.

Warto zaakcentować, że antropomorfizacja stanowi jeden z fundamentów zaufania do AI, ale jej skutki są ambiwalentne. Może ułatwiać interakcję i akceptację technologii, jednocześnie prowadząc do nieadekwatnych oczekiwań i zaburzonej oceny odpowiedzialności. Kluczowe jest więc projektowanie systemów AI tak, aby wsparcie społeczne i emocjonalne nie odbywało się kosztem przejrzystości i bezpieczeństwa.

Nie sposób zaprzeczyć, że psychologiczne reakcje na sztuczną inteligencję i systemy automatyczne cechuje wysoka zmienność. W niektórych sytuacjach użytkownicy z dużą gotowością powierzają im zadania decyzyjne, uznając algorytmy za bardziej precyzyjne, bezstronne i przewidywalne niż oceny ludzkie. Taką tendencję określa się mianem *algorithm appreciation*, czyli preferencji wobec algorytmów (Logg, 2019). W innych przypadkach jednak nawet pojedynczy błąd systemu prowadzi do gwałtownego obniżenia zaufania i odrzucenia technologii – zjawiska znanego jako *algorithm aversion*, czyli awersja wobec algorytmów. Oba mechanizmy stanowią dwa przeciwstawne bieguny społecznego zaufania do technologii i odzwierciedlają sposób, w jaki ludzie interpretują kompetencje, intencje oraz ograniczenia systemów sztucznej inteligencji (Dietvorst i in., 2015).

Z jednej strony ludzie często zakładają, że algorytmy są bardziej obiektywne i wolne od emocji, co prowadzi do nadmiernego zaufania do ich rekomendacji. W kontekstach wymagających analizy danych i logiki, takich jak medycyna, systemy rekomendacyjne czy ocena ryzyka, automatyzacja bywa odbierana jako bardziej kompetentna niż człowiek (Dietvorst i in., 2015). Mechanizm ten może

skutkować bezkrytycznym poleganiem na AI, nawet gdy użytkownik dostrzega ograniczoną przejrzystość działania systemu.

Lee i See (2004) podkreślają, że zaufanie do automatyzacji opiera się głównie na postrzeganej kompetencji i przewidywalności. Gdy AI zachowuje się zgodnie z oczekiwaniami, pojawia się entuzjazm i gotowość do współpracy. Gdy jednak system zaskakuje, zwłaszcza negatywnie, reakcja emocjonalna prowadzi do gwałtownej utraty zaufania. Paradoksalnie więc to, co początkowo buduje zaufanie do AI, może stać się źródłem jego załamania.

Jednocześnie pojawia się fascynacja samą możliwością przekraczania ludzkich ograniczeń. W wielu kontekstach użytkownicy przypisują AI rolę sprzymierzeńca, zdolnego uzupełniać niedoskonałości ludzkiego poznania, zwiększać bezpieczeństwo i poprawiać komfort życia. Ta pozytywna perspektywa może wzmacniać gotowość do integracji technologii z codziennością, zwłaszcza jeśli AI zostaje zaprojektowana tak, aby budzić poczucie przewidywalności, kompetencji oraz „życzliwych intencji”.

Przytoczone wcześniej zjawiska wskazują, że relacja człowiek – AI ma charakter niestabilny, czyli zależny nie tylko od faktycznych możliwości systemu, lecz także od oczekiwań, wcześniejszych doświadczeń oraz sposobu prezentacji wyników. Ta zmienność ma kluczowe znaczenie dla projektowania technologii, ponieważ nadmierne lub niewystarczające zaufanie może prowadzić do poważnych konsekwencji społecznych oraz etycznych.

Psychologiczne mechanizmy lęku przed AI

Postawy wobec sztucznej inteligencji rzadko są jednoznacznie pozytywne lub negatywne. Bez wątpienia w tej kwestii dominują emocje ambiwalentne, łączące zarówno entuzjazm wobec możliwości technologicznych, jak i obawy dotyczące ich konsekwencji dla człowieka oraz społeczeństwa. Z jednej strony AI bywa postrzegana jako narzędzie wspierające rozwój medycyny, edukacji czy bezpieczeństwa publicznego. Z drugiej jednak, budzi lęk przed utratą kontroli, dehumanizacją relacji społecznych oraz zagrożeniem dla pozycji człowieka w hierarchii sprawczości, choćby przez utratę pozycji zawodowej.

Jednym z kluczowych czynników generujących emocjonalną nieufność wobec zaawansowanych technologii jest lęk egzystencjalny. Zgodnie z Teorią

Opanowywania Trwogi⁴ (ang. *Terror Management Theory*; TMT), świadomość własnej śmiertelności skłania ludzi do obrony symbolicznych struktur dających poczucie sensu i bezpieczeństwa (Greenberg, Solomon, Pyszczynski, 1997). AI, zdolna zastępować ludzkie zdolności poznawcze, może zostać odebrana jako zagrożenie dla wyjątkowości człowieka i jego statusu jako jedyne go twórcy oraz interpretatora rzeczywistości (Greenberg i in., 1997). W rezultacie negatywne emocje mogą pojawiać się nie tylko w związku z możliwą utratą pracy, lecz także z bardziej fundamentalnym poczuciem, że rozwój AI „umniejsza” pozycję człowieka w świecie.

Wang i Siau (2019) zauważają, że emocje związane z AI wynikają z napięcia między nadzieją na poprawę jakości życia a poczuciem zagrożenia dla ludzkiej autonomii. Przekazy medialne i narracje kulturowe dodatkowo wzmacniają te reakcje. Obrazy przyjaznych, humanoidalnych robotów mogą wywoływać sympatię i ciekawość, natomiast wizje „buntu maszyn” – niepokój i opór. W konsekwencji każda nowa technologia AI zostaje szybko „osadzona” w istniejących schematach emocjonalnych.

Ambiwalencja ta ma znaczące konsekwencje dla procesów decyzyjnych. Silne emocje, zarówno pozytywne, jak i negatywne, mogą zniekształcać ocenę ryzyka i korzyści, prowadząc do tendencji skrajnych, czyli bezrefleksyjnej akceptacji lub nieuzasadnionego odrzucenia technologii. Zrozumienie psychologicznego podłoża tych reakcji jest zatem kluczowe dla projektowania systemów AI, które będą wzbudzały zaufanie adekwatne do ich rzeczywistych możliwości.

Jednak, jak wskazali Dietvorst i inni (2015), wystarczy jeden widoczny błąd algorytmu, aby zaufanie użytkownika spadło gwałtownie, czyli znacznie silniej, niż w przypadku błędów ludzkich. Ludzie surowiej oceniają technologię niż człowieka za tę samą pomyłkę, co autorzy tłumaczą oczekiwaniem perfekcyjności działania systemów sztucznej inteligencji (Greenberg i in., 1997). W efekcie

⁴ Teoria opanowywania trwogi zakłada, że ludzie są świadomi własnej śmiertelności i w związku z tym doświadczają potencjalnie silnego lęku egzystencjalnego. Aby ten lęk zredukować i utrzymać psychologiczną stabilność, tworzą systemy kulturowych przekonań, które nadają życiu sens, porządek i trwałość. Kluczową rolę pełni również poczucie własnej wartości (samoocena), będące psychologiczną ochroną przed doświadczeniem trwogi. Gdy systemy światopoglądowe lub poczucie wartości zostają zagrożone, pojawiają się nasilone reakcje obronne, takie jak podtrzymywanie własnych przekonań, uprzedzenia czy niechęć wobec grup i obiektów postrzeganych jako zagrażające (Wojciszke, 2006).

użytkownicy mogą całkowicie odrzucić rozwiązania algorytmiczne, mimo że ich przeciętna skuteczność przewyższa skuteczność ludzką.

Postawy wobec sztucznej inteligencji nie kształtują się wyłącznie na poziomie jednostkowym. Istotną rolę odgrywa szerszy kontekst społeczny, czyli normy kulturowe, przekazy medialne oraz dyskurs publiczny, które wspólnie wpływają na sposób postrzegania AI. Wang i Siau (2019) zwracają uwagę, że nowe technologie są zawsze osadzone w realiach społecznych, dlatego emocje i oczekiwania wobec AI wynikają również z narracji dotyczących przyszłości pracy, utraty kontroli czy transformacji relacji międzyludzkich. W tym ujęciu sztuczna inteligencja staje się symbolem szans i zagrożeń społecznych, odzwierciedlając dominujące przekonania na temat rozwoju cywilizacyjnego.

Kulturowa interpretacja AI wiąże się również z mechanizmami obrony tożsamości. Zgodnie z wspomnianą wcześniej TMT, ludzie poszukują stabilnych ram znaczeniowych, które chronią poczucie sensu i wartości. Technologie, które mogą zakwestionować wyjątkowość człowieka, np. poprzez wykazywanie wyższej skuteczności w zadaniach intelektualnych, wywołują reakcje obronne, szczególnie tam, gdzie centralną wartością jest autonomia i sprawczość jednostki (Greenberg, 1997). W społeczeństwach silnie zorientowanych na indywidualizm obawy te mogą przybierać bardziej intensywną formę.

Znaczący wpływ na oceny AI mają także media i popkultura, które często wykorzystują skrajne narracje, tj. od idealistycznych wizji wszechmocnych technologii rozwiązujących globalne problemy, po katastroficzne scenariusze dominacji maszyn. Tego rodzaju przekazy wzmacniają emocjonalne bieguny postaw, tj. od entuzjastycznej akceptacji po silny opór i lęk. W rezultacie część społeczeństwa traktuje AI jako źródło nadziei, inni z kolei, jako zagrożenie dla porządku społecznego.

Istotne są również różnice wynikające z poziomu kompetencji cyfrowych. Osoby lepiej rozumiejące zasady działania systemów informatycznych wykazują większą zdolność do adekwatnej oceny ryzyka i korzyści, podczas gdy brak wiedzy może prowadzić do uogólnionego strachu lub przeciwnie – bezrefleksyjnej fascynacji technologią. To sugeruje, że edukacja w obszarze AI może pełnić kluczową rolę w kształtowaniu dojrzałego, zrównoważonego zaufania społecznego.

Podsumowując, ocena sztucznej inteligencji jest wynikiem złożonej gry wpływów jednostkowych emocji, przekonań kulturowych oraz narracji

społecznych. AI nie jest zatem postrzegana jedynie jako narzędzie, lecz jako aktor symboliczny, angażujący wartości, tożsamość i wyobrażenia o przyszłości człowieka w świecie.

Konsekwencje psychologiczne oraz implikacje praktyczne dla projektowania zaufanej AI

Właściwie ukształtowane zaufanie do sztucznej inteligencji może przynosić istotne korzyści zarówno jednostkom, jak i całemu społeczeństwu. Gdy użytkownicy postrzegają system jako wystarczająco kompetentny i przewidywalny, są bardziej skłonni angażować się w współpracę człowiek – technologia, co przekłada się na poprawę skuteczności działań i redukcję obciążenia poznawczego (Lee, See, 2004). W obszarach wymagających przetwarzania dużych zbiorów danych AI może wspierać podejmowanie decyzji, ograniczając ryzyko błędów wynikających z ograniczeń ludzkiej pamięci i uwagi.

Zaufanie oparte na racjonalnych przesłankach zwiększa również bezpieczeństwo, szczególnie w zadaniach o podwyższonym ryzyku. Od medycyny po transport – automatyzacja może pełnić funkcję wsparcia w sytuacjach krytycznych, oferując szybszą analizę danych oraz natychmiastowe ostrzeżenie o zagrożeniach. W takich przypadkach AI nie zastępuje człowieka, lecz pomaga mu podejmować decyzje bardziej świadomie i precyzyjnie (Lee, See, 2004).

Istotną korzyścią jest także zmniejszenie obciążenia emocjonalnego użytkowników. W systemach, w których algorytmy przyjmują część odpowiedzialności za analizę i rekomendacje, jednostka może doświadczać większego poczucia kontroli i bezpieczeństwa (Dietvorst i in., 2015). Jednocześnie pozytywne emocje, takie jak ciekawość i satysfakcja z kontaktu z technologią, mogą wzmacniać motywację do nauki i adaptacji nowych narzędzi.

W kontekście społecznym właściwe zaufanie do AI może przyczyniać się do zwiększenia dostępności usług, szczególnie w obszarze edukacji i opieki zdrowotnej. Systemy wspomagające komunikację, personalizację treści czy monitorowanie stanu zdrowia mogą poprawiać jakość życia osób z ograniczeniami funkcjonalnymi oraz zmniejszać nierówności wynikające z dostępu do specjalistów.

Warunkiem uzyskania tych korzyści jest adekwatna kalibracja zaufania, tj. zgodna z realnymi możliwościami i ograniczeniami systemu. Gdy użytkownik

ufa technologii na poziomie właściwym, proces współpracy człowiek – AI przebiega harmonijnie, a emocje i heurystyki poznawcze działają wspierająco, a nie zakłócająco. W rezultacie AI może stać się nie tylko narzędziem, ale partnerem, który wzmacnia ludzkie zdolności poznawcze, emocjonalne i społeczne.

Zaufanie do sztucznej inteligencji, choć pożądane z perspektywy efektywności i bezpieczeństwa, może również prowadzić do szeregu zagrożeń, jeśli nie zostanie odpowiednio skalibrowane. Niewłaściwe zaufanie można rozpatrywać jako nadmierne lub niedostateczne, przy czym każda z tych postaw wiąże się z odmiennymi konsekwencjami dla użytkownika i środowiska społecznego.

Nadmierne zaufanie (ang. *overtrust*) sprzyja bezkrytycznemu poleganiu na technologii, szczególnie w sytuacjach, w których użytkownik uznaje system za domyślnie bardziej kompetentny od człowieka (Lee, See, 2004). W rezultacie decyzje o dużym znaczeniu podejmowane są automatycznie, bez monitorowania i weryfikacji wyników działania algorytmu. Zjawisko to określa się mianem wspomnianego wcześniej *automation bias*. W obszarach takich jak medycyna czy transport może to prowadzić do poważnych błędów operacyjnych oraz zagrożeń dla zdrowia i życia.

Drugą skrajnością jest niedostateczne zaufanie, czyli unikanie lub odrzucanie technologii pomimo obiektywnie wyższej skuteczności systemów algorytmicznych w porównaniu z ludźmi. Przywołane badania dają podstawę, by sądzić, że nawet pojedynczy błąd prowadzi do gwałtownego spadku zaufania do AI, znacznie większego niż w przypadku błędów ludzkich. Postawa ta, określana jako *algorithm aversion*, może skutkować rezygnacją z innowacyjnych rozwiązań, które mogłyby poprawić jakość podejmowanych decyzji, np. w diagnostyce medycznej czy analizie danych krytycznych.

Obie skrajne reakcje wynikają często z błędnych oczekiwań wobec technologii. Nadmierne zaufanie opiera się na przekonaniu, że AI jest nieomylna, z kolei niedostateczne, że jej błędy są bardziej znaczące niż błędy ludzkie. Tymczasem w literaturze podkreśla się, że zaufanie powinno być funkcją rzeczywistych możliwości systemu, uwzględniając zarówno jego mocne strony, jak i ograniczenia (Lee, See, 2004). Zachwianie tej równowagi prowadzi do dezorganizacji relacji człowiek – technologia.

Dodatkowym zagrożeniem jest przesunięcie odpowiedzialności. Gdy systemy prezentowane są jako inteligentne i autonomiczne, ludzie mogą przypisywać

im winę za błędne decyzje, unikając odpowiedzialności osobistej lub instytucjonalnej (Bryson, 2018). Diagnostyczna aura obiektywności przypisana algorytmom utrudnia krytyczną ewaluację wyników i stawia przed projektantami etyczne zobowiązanie do zapewnienia transparentności i nadzoru.

Aby zaufanie do sztucznej inteligencji było adekwatne, tj. ani zbyt wysokie, ani zbyt niskie, konieczne jest projektowanie systemów w sposób, który uwzględnia ograniczenia ludzkiej percepcji, emocji i procesów poznawczych. Lee i See (2004) podkreślają, że kluczowymi czynnikami kształtującymi zaufanie są kompetencja, przewidywalność oraz postrzegane intencje systemu. Oznacza to, że użytkownicy muszą rozumieć zarówno to, co technologia potrafi, jak również to, czego nie jest w stanie zrobić.

Jednym z najważniejszych elementów projektowania jest transparentność. Systemy AI powinny przedstawiać użytkownikowi czytelne informacje o zastosowanych danych, zakresie działania oraz poziomie pewności modelu. W procesie tym ogromną rolę odgrywa nurt *Explainable AI* (XAI), rozwijający narzędzia umożliwiające zrozumienie logiki decyzji algorytmu nawet osobom bez zaawansowanej wiedzy technicznej (Lee, See, 2004). Przejrzystość zmniejsza niepewność i obawy, co może zapobiegać zarówno nadmiernemu, jak i niedostatecznemu zaufaniu.

Istotnym obszarem jest również projektowanie komunikatów o błędach. Skoro ludzie tracą zaufanie do AI po jednym błędzie, odpowiednie projektowanie komunikacji nabiera dużego znaczenia. W tym kontekście można postulować wyjaśnianie przyczyn błędów i sposobów ich korekcji, co może łagodzić ten efekt i wzmacniać poczucie kontroli użytkownika.

Z kolei elementy antropomorficzne należy stosować w sposób funkcjonalny, a nie wyłącznie estetyczny. R. Złotowski i inni (2015) podkreślają, że sygnały społeczne AI mogą poprawiać jakość współpracy i budować zaufanie. Nadmierna antropomorfizacja może jednak prowadzić do błędnego przypisywania autonomii i intencji, co skutkuje niebezpiecznym przesunięciem odpowiedzialności (Bryson, 2018). Dlatego antropomorfizacja powinna być skalowalna i sytuacyjna, np. w robotach społecznych – uzasadniona, a w aplikacjach medycznych – ograniczona.

W projektowaniu systemów AI nie można pomijać norm kulturowych i społecznych. Jak wskazują Wang wraz z Siau (2019), akceptacja technologii różni się w zależności od kontekstu, narracji publicznych i poziomu kompetencji

cyfrowych. Z tego względu wdrożenie tego samego rozwiązania może przynieść odmienne efekty w różnych populacjach, zatem projekt musi być kontekstowy, oparty na badaniach użytkowników.

Można zatem przyjąć, że adekwatne zaufanie do AI wymaga zintegrowania wiedzy z zakresu psychologii, interakcji człowiek – technologia i etyki projektowania. Systemy powinny być przewidywalne, zrozumiałe i odpowiedzialne, a użytkownicy wyposażeni w narzędzia do świadomego nadzoru nad technologią. Tylko wówczas współpraca człowieka i AI będzie wspierała bezpieczeństwo społeczne i rozwój, zamiast generować nowe zagrożenia.

Warto również podkreślić, że rozwój systemów sztucznej inteligencji rodzi pytania nie tylko o zaufanie użytkowników, lecz także o odpowiedzialność za skutki decyzji podejmowanych z udziałem tych technologii. W wielu zastosowaniach AI jest postrzegana jako podmiot autonomiczny, zdolny działać „samodzielnie”, co może prowadzić do błędnego przypisywania jej intencji oraz moralnej sprawczości. Bryson (2018) ostrzega, że traktowanie AI jako bytu etycznego jest nie tylko nieuzasadnione, ale również może skutkować niebezpiecznym odsunięciem odpowiedzialności od ludzi. Projektanci, operatorzy i instytucje wdrażające systemy muszą więc pozostać centralnymi aktorami odpowiedzialności etycznej i prawnej.

W wielu kontekstach AI pełni rolę doradczą, wpływając na decyzje człowieka poprzez rekomendacje, klasyfikacje czy przewidywania. Taka konfiguracja relacji rodzi ryzyko, że użytkownicy, zwłaszcza w sytuacjach nadmiernego zaufania, będą stawać się zależni od algorytmu, ograniczając własny nadzór i refleksyjność. Jednocześnie niewystarczające zaufanie może powodować odrzucenie technologii o wysokiej przydatności społecznej. Dlatego kluczowe znaczenie ma jasne określenie granic roli AI tak, aby jej udział wspierał decydenta, a nie zastępował jego odpowiedzialności (Lee, See, 2004).

Z punktu widzenia projektowania etycznego ważne jest zapewnienie zrozumiałych reguł funkcjonowania systemu, obejmujących m.in. informacje o jakości danych, poziomie błędu oraz o tym, kto ponosi odpowiedzialność za skutki działania algorytmu (Lee, See, 2004). W przypadku błędów AI, czy to w ocenie medycznej, czy w decyzjach dotyczących bezpieczeństwa, nie można dopuścić do sytuacji, w której odpowiedzialność staje się „niczyja”. AI nie odczuwa bowiem

konsekwencji własnych decyzji i nie może zostać pociągnięta do odpowiedzialności moralnej ani prawnej (Bryson, 2018).

Granice sprawczości AI muszą być także komunikowane użytkownikom, w przeciwnym razie ryzykujemy zjawisko „moralnego rozmycia odpowiedzialności”. Odpowiedzialne projektowanie zakłada stałą obecność człowieka „na pętli decyzyjnej”, szczególnie w obszarach związanych z etyką, zdrowiem publicznym oraz bezpieczeństwem. Współpraca człowiek – AI powinna być zorganizowana tak, aby technologia wzmacniała kompetencje ludzi, a nie przyczyniała się do ich zastępowania lub degradacji.

Zaufanie do sztucznej inteligencji jest złożonym konstruktym psychologicznym, którego konsekwencje mogą być zarówno korzystne, jak i potencjalnie niebezpieczne dla jednostki oraz społeczeństwa. Odpowiednio skalibrowane zaufanie wspiera efektywność działań, poprawia bezpieczeństwo i pozwala wykorzystać mocne strony technologii w różnych dziedzinach życia, od medycyny po edukację. Jednak niewłaściwe zaufanie, zarówno nadmierne, jak i niedostateczne, może prowadzić do poważnych błędów decyzyjnych, utraty nadzoru nad systemem lub odrzucania rozwiązań, które mogłyby przynieść istotne korzyści. Dlatego projektowanie AI musi uwzględniać ludzkie ograniczenia poznawcze, emocjonalne i społeczne, zapewniając jednocześnie przejrzystość działania technologii, informowanie o jej możliwościach oraz wyraźne określenie granic odpowiedzialności człowieka i maszyny. Tylko wówczas współpraca człowiek – AI będzie sprzyjać rozwojowi i dobrostanowi, a nie generować nowych zagrożeń.

Podsumowanie

Celem rozdziału było ukazanie psychologicznych mechanizmów determinujących postawy wobec sztucznej inteligencji oraz określenie ich konsekwencji dla odpowiedzialnego i bezpiecznego wdrażania technologii w życie społeczne. Przeprowadzona analiza pozwala stwierdzić, że postawy wobec AI mają charakter głęboko ambiwalentny i są wynikiem współdziałania mechanizmów poznawczych, emocjonalnych oraz społeczno-kulturowych. W odniesieniu do problematyki szczegółowej rozdziału można wskazać, że zarówno zaufanie, jak i lęk wobec systemów autonomicznych mają swoje korzenie w tych samych procesach

psychologicznych – jednak prowadzą do odmiennych konsekwencji dla użytkownika oraz społeczeństwa.

Nawiązując do pierwszego problemu badawczego, dotyczącego psychologicznych mechanizmów antropomorfizacji i zaufania do AI, wyniki analizy wskazują, że uproszczone strategie poznawcze pełnią kluczową rolę w ocenie działania systemów autonomicznych. Heurystyki poznawcze sprawiają, że użytkownicy wyciągają wnioski na podstawie pozornych wskazówek, które sugerują kompetencję i obiektywność systemu, nawet gdy jego działanie nie jest w pełni zrozumiałe. Zaufanie do AI opiera się więc często na wrażeniu przewidywalności, skuteczności lub obiektywności, a nie na świadomości faktycznych możliwości technologii. Jak pokazały badania nad *algorithm aversion* i *algorithm appreciation*, reakcje użytkowników są skrajnie zmienne – od bezkrytycznego polegania na rekomendacjach algorytmu po nagłe odrzucenie technologii już po pierwszym błędzie. Zjawiska te ujawniają, że zaufanie nie jest konstruktem stabilnym ani prostym, lecz dynamicznym i silnie zależnym od wcześniejszych doświadczeń, kontekstu zadania oraz formy prezentacji danych.

Antropomorfizacja, która stanowi kolejny istotny element w omawianym obszarze, działa jako mechanizm ułatwiający interpretację zachowania AI poprzez ludzkie kategorie mentalne. Nadawanie technologii cech ludzkich, tj. intencji, emocji, kompetencji społecznych, może zwiększać jej akceptowalność, poczucie bliskości oraz przewidywalność. Jednocześnie nadmierna antropomorfizacja może prowadzić do fałszywego przekonania o podmiotowości moralnej AI oraz do „przesuwania” odpowiedzialności z ludzi na systemy techniczne. W skrajnych przypadkach nadmiernie „ludzka” forma robotów wywołuje efekt *uncanny valley*, który z kolei prowadzi do nieufności i dystansu emocjonalnego. Oznacza to, że antropomorfizacja działa dwutorowo, czyli może zarówno wspierać budowę zaufania, jak i je podważać, w zależności od stopnia podobieństwa do człowieka, cech formy oraz kontekstu społecznego.

Drugi problem badawczy dotyczył psychologicznych mechanizmów lęku przed AI. Analiza wskazuje, że źródłem obaw nie są jedynie funkcjonalne ograniczenia technologii, lecz głęboko zakorzenione mechanizmy emocjonalne. Szczególne znaczenie ma tu lęk egzystencjalny, opisany w ramach Teorii Opanowywania Trwogi, zgodnie z którą zagrożenie dla symbolicznych struktur wartości prowadzi do reakcji obronnych i negatywnych postaw wobec obiektów

podważających poczucie sensu i wyjątkowości człowieka. Sztuczna inteligencja, która jest zdolna wykonywać zadania intelektualne, przewidywać zachowania i podejmować decyzje, może być postrzegana jako zagrożenie dla ludzkiej autonomii oraz unikalności poznawczej. Dodatkowo lęk przed AI wzmacniają czynniki społeczne i kulturowe, tj. narracje medialne, wizje przyszłości pracy oraz wyobrażenia dotyczące utraty kontroli. W rezultacie lęk wobec AI nie jest wyłącznie reakcją na technologię, lecz na jej symboliczne znaczenie dla pozycji człowieka w świecie społecznym.

Odpowiadając na trzeci problem badawczy, dotyczący psychologicznych konsekwencji i implikacji praktycznych, można stwierdzić, że kluczowym wyzwaniem dla projektowania zaufanej AI jest właściwa kalibracja zaufania. Nadmierne zaufanie prowadzi do niebezpiecznych sytuacji, takich jak *automation bias* i nieumiejętność krytycznej oceny działań systemu, natomiast niedostateczne zaufanie może powodować odrzucanie technologii o wysokim potencjale społecznym. Oba te zjawiska mają realne konsekwencje dla bezpieczeństwa, efektywności i odpowiedzialności. Dlatego projektowanie systemów AI musi być oparte na transparentności, przewidywalności i jasnym komunikowaniu ograniczeń technologii. Szczególne znaczenie ma tu rozwój *Explainable AI* (XAI), który zwiększa zrozumienie działania algorytmów, a tym samym redukuje niepewność i obawy użytkowników.

W kontekście odpowiedzialnego wdrażania AI istotne jest również jednoznaczne określenie granic sprawczości technologii oraz przypisanie odpowiedzialności moralnej i prawnej ludziom, a nie maszynom. Psychologiczne procesy antropomorfizacji mogą prowadzić do błędnego przekonania, że technologie autonomiczne ponoszą odpowiedzialność za swoje decyzje, co wymaga przeciwdziałania poprzez odpowiednie projektowanie komunikatów i interfejsów. Równie istotne jest uwzględnienie różnic indywidualnych i kulturowych w poziomie umiejętności cyfrowych, ponieważ wpływają one na sposób, w jaki ludzie interpretują i oceniają działanie AI.

Podsumowując, analiza przeprowadzona w rozdziale wskazuje, że zaufanie i lęk wobec AI stanowią dwa komplementarne, lecz przeciwstawne rezultaty tych samych mechanizmów psychologicznych. Zrozumienie ich funkcjonowania jest kluczowe dla projektowania technologii, które wspierają autonomię człowieka, a jednocześnie nie prowadzą do nadmiernej zależności od systemów

algorytmicznych. Odpowiedzialne wdrażanie AI wymaga zatem połączenia wiedzy z zakresu psychologii, technologii oraz etyki, tak aby zapewnić nie tylko funkcjonalność technologii, lecz także jej akceptowalność społeczną. Tylko wówczas możliwe będzie budowanie dojrzałego, zrównoważonego zaufania do AI, które sprzyja bezpieczeństwu, dobrostanowi i rozwojowi społecznemu.

Literatura

1. Bryson, J. J. (2018). Patency is not a virtue: The design of intelligent systems and systems of ethics. *Ethics and Information Technology*, 20(1), 15-26, <https://doi.org/10.1007/s10676-018-9448-6>
2. Christoforakos, L., Gallucci, A., Surmava-Große, T., Ullrich, D., Diefenbach, S. (2021). Can Robots Earn Our Trust the Same Way Humans Do? *A Systematic Exploration of Competence, Warmth, and Anthropomorphism as Determinants of Trust Development in HRI*. *Frontiers in Robotics and AI*, 8, 640444. <https://doi.org/10.3389/frobt.2021.640444>
3. Dietvorst, B. J., Simmons, J. P., Massey, C. (2015). Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General*, 144(1), 114-131, <https://doi.org/10.1037/xge0000033>
4. Greenberg, J., Solomon, S., Pyszczynski, T. (1997). *Terror Management Theory of Self-Esteem and Cultural Worldviews: Empirical Assessments and Conceptual Refinements*. W: M. P. Zanna (red.), *Advances in Experimental Social Psychology*, vol. 29. Academic Press. [https://doi.org/10.1016/S0065-2601\(08\)60016-7](https://doi.org/10.1016/S0065-2601(08)60016-7)
5. Lee, J. D., See, K. A. (2004), Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50-80, https://doi.org/10.1518/hfes.46.1.50_30392
6. Logg, J. M., Minson, J. A., Moore, D. A. (2019). Algorithm Appreciation: People Prefer Algorithmic to Human Judgment. *Organizational Behavior and Human Decision Processes*, 151, 90-103, <https://doi.org/10.1016/j.obhdp.2018.12.005>
7. Mori, M., MacDorman, K. F., Kageki, N. (2012). The Uncanny Valley. *IEEE Robotics & Automation Magazine*, 19(2), 98-100. <https://doi.org/10.1109/MRA.2012.2192811>
8. Tversky, A., Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, 185(4157), 1124-1131. <https://doi.org/10.1126/science.185.4157.1124>

9. Wang, W., Siau, K. (2019). Artificial Intelligence, Machine Learning, Automation, Robotics, Future of Work and Future of Humanity: A Review and Research Agenda. *Journal of Database Management*, 30(1), 61-79. <https://doi.org/10.4018/JDM.2019010104>
10. Waytz, A., Cacioppo, J., Epley, N. (2010). Who Sees Human? The Stability and Importance of Individual Differences in Anthropomorphism. *Perspectives on Psychological Science*, 5(3), 219-32, <https://doi.org/10.1177/1745691610369336>
11. West, S. M., Whittaker, M., Crawford, K. (2019). *Discriminating systems: Gender, race, and power in AI*. AI Now Institute Report. Pobrane z: <https://ainowinstitute.org/publications/discriminating-systems-gender-race-and-power-in-ai-2> (dostęp: 02.09.2025).
12. Wojciszke, B. (2006), *Człowiek wśród ludzi. Zarys psychologii społecznej*. Scholar.
13. Złotowski, R., Proudfoot, D., Yogeewaran, K., Bartneck, C. (2015). Anthropomorphism: Opportunities and Challenges in Human-Robot Interaction. *International Journal of Social Robotics*, 7, 347-360. <https://doi.org/10.1007/s12369-014-0267-6>