Kompleksowy przewodnik po tworzeniu Person AI i nadawaniu osobowości cyfrowym agentom

Streszczenie

Niniejszy przewodnik przedstawia kompleksową analizę współczesnych metod tworzenia sztucznych osobowości dla agentów konwersacyjnych i systemów AI. Opracowanie opiera się na systematycznym przeglądzie 224 publikacji naukowych z lat 2010-2025, ze szczególnym uwzględnieniem najnowszych osiągnięć w dziedzinie dużych modeli językowych (LLM).

Współczesne podejścia do tworzenia person AI obejmują trzy główne metody techniczne: (1) systemy oparte na regułach, które mapują cechy osobowości na moduły decyzyjne, (2) kondycjonowanie i dostrajanie LLM poprzez prompty psychometryczne, oraz (3) multimodalne systemy ekspresji łączące komunikację werbalną z niewerbalną. Większość implementacji wykorzystuje model Wielkiej Piątki (Big Five) lub wskaźnik Myers-Briggs (MBTI) jako ramy teoretyczne.

Badania wykazują, że agenci AI mogą skutecznie symulować ludzkie cechy osobowości na poziomie percepcji użytkowników i korelacji behawioralnych, jednak istnieją znaczące ograniczenia w zakresie wewnętrznej spójności, autentyczności psychologicznej i długoterminowej konsystencji. Próby zastosowania rzeczywistych nauk psychologicznych przynoszą obiecujące rezultaty w zakresie zaangażowania użytkowników i budowania zaufania, ale ujawniają również fundamentalne różnice między ludzką a sztuczną psychiką, szczególnie w zakresie subtelnych struktur czynnikowych, regulacji emocjonalnej i rozumowania moralnego.

Kluczowe wyzwania obejmują niestabilność ekspresji cech w różnych kontekstach, trudności w oddzieleniu stylu komunikacji od rzeczywistej ekspresji osobowości, oraz istotne kwestie etyczne związane z antropomorfizacją, zaufaniem i potencjalną manipulacją. Przyszłe kierunki rozwoju wskazują na potrzebę integracji walidacji psychometrycznej z multimodalnymi systemami ekspresji, rozwoju modułowych architektur osobowości oraz interdyscyplinarnej współpracy między informatyką, a psychologią.

1. Wprowadzenie

1.1. Kontekst i motywacja

Tworzenie sztucznych osobowości dla agentów konwersacyjnych i systemów AI stanowi jedno z najbardziej fascynujących i złożonych wyzwań współczesnej informatyki afektywnej i interakcji człowiek-komputer. W miarę jak systemy AI stają się coraz bardziej zintegrowane z codziennym życiem – od asystentów głosowych po chatboty wsparcia psychicznego – potrzeba nadania im wiarygodnych, spójnych i odpowiednich osobowości staje się kluczowa dla efektywności interakcji i akceptacji społecznej [1], [2].

Osobowość w kontekście agentów AI nie jest jedynie kosmetycznym dodatkiem, ale fundamentalnym elementem wpływającym na zaufanie użytkowników, zaangażowanie, gotowość do zwierzeń oraz ogólną jakość interakcji [3], [4]. Badania wykazują, że odpowiednio dobrana osobowość agenta może znacząco poprawić skuteczność systemów doradczych, terapeutycznych i edukacyjnych [5], [6].

1.2. Zakres i cele przewodnika

1. Przegląd współczesnych metod technicznych stosowanych do tworzenia person AI, ze szczególnym uwzględnieniem architektur opartych na regułach, kondycjonowania LLM oraz systemów multimodalnych.
2. Analizę zastosowania ram psychologicznych, w szczególności modelu Wielkiej Piątki, MBTI i teorii cech, w projektowaniu sztucznych osobowości.
3. Ocenę skuteczności prób zastosowania rzeczywistych nauk psychologicznych do tworzenia psychiki AI zbliżonej do ludzkiej.
4. Identyfikację ograniczeń technicznych, teoretycznych i etycznych obecnych podejść.
5. Przedstawienie rekomendacji dla przyszłych badań i praktycznych implementacji.

1.3. Metodyka

Analiza opiera się na systematycznym przeglądzie 224 unikalnych publikacji naukowych z lat 2010-2025, obejmujących badania z baz Google Scholar, ArXiv oraz innych źródeł akademickich. Szczególną uwagę poświęcono 30 najbardziej relewantnym pracom, które zostały poddane szczegółowej analizie pod kątem metodologii technicznej, zastosowanych ram psychologicznych oraz kluczowych odkryć i ograniczeń.

2. Psychologiczne podstawy teoretyczne

2.1. Model Wielkiej Piątki (Big Five / OCEAN)

Model Wielkiej Piątki stanowi najczęściej stosowaną ramę teoretyczną w projektowaniu person AI [1], [2], [7]. Model ten definiuje osobowość poprzez pięć głównych wymiarów:

- Otwartość na doświadczenie (Openness): kreatywność, ciekawość intelektualna, otwartość na nowe idee
- Sumienność (Conscientiousness): organizacja, odpowiedzialność, dyscyplina
- Ekstrawersja (Extraversion): towarzyskość, asertywność, energia w interakcjach społecznych
- Ugodowość (Agreeableness): empatia, współpraca, życzliwość
- Neurotyczność (Neuroticism): stabilność emocjonalna, podatność na stres i negatywne emocje

W kontekście agentów AI, cechy te są operacjonalizowane poprzez mapowanie na konkretne zachowania komunikacyjne, style językowe, parametry ekspresji niewerbalnej oraz strategie planowania i reagowania [1], [5], [8]. Na przykład, wysoka ekstrawersja może być wyrażana przez częstsze inicjowanie konwersacji, używanie bardziej ekspresywnego języka i większą liczbę gestów [2].

2.2. Myers-Briggs Type Indicator (MBTI)

MBTI, choć kontrowersyjny w psychologii akademickiej, znajduje szerokie zastosowanie w tworzeniu banków person dla agentów konwersacyjnych [9]. System ten klasyfikuje osobowość według czterech dychotomii, tworząc 16 typów osobowości. CharacterChat wykorzystuje dekompozycję person opartą na MBTI do stworzenia banku MBTI-1024 zawierającego wirtualne postacie o różnych profilach, co umożliwia dopasowanie optymalnych wspierających do użytkowników [9].

2.3. Teorie emocji i regulacji afektywnej

Oprócz modeli cech osobowości, zaawansowane systemy person AI integrują teorie emocji, takie jak:

- Model OCC (Ortony, Clore, Collins): definiuje emocje jako reakcje na ocenę zdarzeń, działań i obiektów [5]
- Teoria Tendencji do Działania (Action Tendencies): opisuje emocje jako przygotowanie do określonych wzorców zachowań [10]
- Model Circumplex: reprezentuje emocje w przestrzeni dwuwymiarowej (pobudzenie-walencja) [11]

Te ramy teoretyczne pozwalają na bardziej złożone modelowanie nie tylko statycznych cech osobowości, ale również dynamicznych procesów emocjonalnych i regulacyjnych [10], [12].

2.4. Modele Społeczno-Poznawcze

Niektóre zaawansowane architektury person AI wykorzystują podejścia społeczno-poznawcze, które uwzględniają interakcję między osobowością a kontekstem społecznym. Model SPOT (Simulating Personality Over Time) łączy podejście leksykalne z społeczno-poznawczym, umożliwiając symulację zmian osobowości w czasie [11].

3. Metody tworzenia Person AI

3.1. Systemy Oparte na Regułach

3.1.1. Architektura i Zasada Działania

Systemy oparte na regułach konwertują cechy osobowości na modularne reguły decyzyjne, które modyfikują procesy oceny (appraisal), planowania, radzenia sobie (coping) i ekspresji w architekturze agenta [5], [13]. Podejście to charakteryzuje się wysokim stopniem kontroli i przewidywalności zachowań agenta.

3.1.2. Implementacje Referencyjne

Doce i współpracownicy [5] opracowali architekturę integrującą FAtiMA (dla umysłu agenta, obsługującego emocje, radzenie sobie i planowanie) z GRETA (dla ciała, wyrażającego działania poprzez gesty, mimikę i postawę). W tym systemie osobowość jest kodowana poprzez:

- Definiowanie progów i tempa zaniku emocji: cechy osobowości wpływają na to, jak łatwo dana emocja jest wywoływana i jak długo utrzymuje się
- Wpływ na heurystyki planowania: np. wysoka sumienność faworyzuje dokładne, długoterminowe plany
- Modyfikację strategii radzenia sobie: różne cechy osobowości predysponują do różnych mechanizmów obronnych
- Parametry ekspresywności: kontrolują intensywność ekspresji niewerbalnej

Badania wykazały, że użytkownicy potrafili rozpoznać różnice w osobowościach agentów, z silnymi korelacjami dla ekstrawersji, neurotyczności i ugodowości, choć sumienność okazała się trudniejsza do percepcji [5].

3.1.3. Zalety i ograniczenia

Zalety	Ograniczenia
• Pełna kontrola nad zachowaniem agenta • Przewidywalność i deterministyczność • Możliwość precyzyjnego dostrajania poszczególnych aspektów osobowości • Łatwość debugowania i modyfikacji	• Wysoki koszt rozwoju i utrzymania • Trudność w uchwyceniu subtelności ludzkiej osobowości • Ograniczona elastyczność w nowych kontekstach • Wymaga głębokiej wiedzy psychologicznej do prawidłowego mapowania cech na zachowania

3.2. Kondycjonowanie i dostrajanie Dużych Modeli Językowych (LLM)

3.2.1. Prompty psychometryczne

Najnowsze podejścia wykorzystują LLM jako podstawę do generowania zachowań zgodnych z określonymi cechami osobowości poprzez specjalnie zaprojektowane prompty [3], [14], [15]. Huang i współpracownicy [14], [15] opracowali podejście psychometryczne, w którym osobowość jest indukowana poprzez przypisanie agentom wyników z inwentarzy osobowości (np. BFI-2) w promptach.

Dwa główne formaty promptów:

- **Format Likerta**: Przedstawia stwierdzenia z inwentarza osobowości wraz z odpowiedziami na skali Likerta (np. "Zgadzam się w pełni")
- **Format Rozszerzony** (Expanded): Prezentuje opcje odpowiedzi jako pełne zdania, co okazało się bardziej niezawodne i odporne na zakłócenia [14], [15]

Przykład promptu psychometrycznego:
Jesteś agentem AI z następującym profilem osobowości: - Ekstrawersja: wysoka (4.5/5) - Ugodowość: bardzo wysoka (4.8/5) - Sumienność: umiarkowana (3.2/5) - Neurotyczność: niska (2.1/5) - Otwartość: wysoka (4.3/5) Odpowiadaj w sposób zgodny z tym profilem osobowości.

3.2.2. Dostrajanie na zbiorach danych person

Pal i współpracownicy [7] zaproponowali nowatorskie podejście wykorzystujące długie wpisy dziennikowe z Reddit do tworzenia bogatych w osobowość dialogów. Proces obejmuje:

- Klasteryzację wpisów dziennikowych według cech Wielkiej Piątki
- Filtrowanie najbardziej reprezentatywnych klastrów
- Generowanie dialogów (około 400,000) przy użyciu Llama 3 70B
- Dostrajanie modeli na tym zbiorze danych

Wyniki pokazują 11% poprawę w uchwyceniu cech osobowości w porównaniu z istniejącymi podejściami, z lepszą spójnością i autentycznością dialogów [7].

3.2.3. Kondycjonowanie roli i osobowości

Badania Cava i współpracowników [16] wykazują, że łączenie kondycjonowania roli (np. zawód) z kondycjonowaniem osobowości generalnie poprawia zdolność modeli do naśladowania określonych osobowości. Jednak skuteczność tego podejścia różni się znacząco między modelami – niektóre modele open-source (SOLAR, Dolphin, NeuralChat, Llama3-8) dobrze naśladują osobowości, podczas gdy inne pozostają "zamknięte umysłowo" i wykazują silne wewnętrzne preferencje osobowościowe [16].

3.2.4. Deterministyczna ekspresja osobowości

Kruijssen i współpracownicy [3], [17] wykazali, że zaawansowane modele AI (GPT-4o, o1) mogą wyrażać deterministyczne i spójne osobowości, gdy są instruowane przy użyciu ustalonych ram psychologicznych. Kluczowe odkrycie: ekspresja osobowości wyłania się z holistycznego rozumowania, a nie z optymalizacji pytanie po pytaniu [3], [17]. Co istotne, dostrajanie modelu wpływa na styl komunikacji niezależnie od dokładności ekspresji osobowości [3].

3.3. Multimodalne systemy ekspresji

3.3.1. Integracja kanałów komunikacji

Sonlu i współpracownicy [2] opracowali kompleksową ramę dla agentów konwersacyjnych wyrażających osobowość poprzez:

Zachowania niewerbalne	Zachowania werbalne
• Ruch ciała i jakości Laban (Shape i Effort) • Ekspresje mimiczne • Gesty	• Wybór dialogu • Transformacja głosu (intonacja, tempo, wysokość)

Kluczowe odkrycie: każda dodatkowa modalność komunikacji poprawia dokładność percepcji osobowości, z najlepszymi wynikami osiąganymi przy obecności wszystkich modalności [2].

3.3.2. Mapowanie cech na modalności

Tabela 1 przedstawia przykładowe mapowanie cech Wielkiej Piątki na różne modalności ekspresji.

Cecha osobowości	Ekspresja werbalna	Ekspresja głosowa	Ekspresja niewerbalna
Ekstrawersja (wysoka)	Częste inicjowanie rozmów, ekspresywny język, dłuższe wypowiedzi	Głośniejszy głos, szybsze tempo, większa zmienność wysokości	Szerokie gesty, otwarta postawa, częsty kontakt wzrokowy
Sumienność (wysoka)	Strukturyzowane odpowiedzi, precyzyjny język, odniesienia do planów	Wyraźna artykulacja, równomierne tempo	Kontrolowane, celowe ruchy, wyprostowana postawa
Ugodowość (wysoka)	Empatyczne sformułowania, częste potwierdzenia, unikanie konfrontacji	Ciepły ton, łagodne modulacje	Kiwanie głową, uśmiech, nachylanie się do rozmówcy
Neurotyczność (wysoka)	Wyrażanie niepewności, częste zastrzeżenia, emocjonalny język	Drżący głos, nieregularne tempo, wyższy ton	Napięta postawa, nerwowe ruchy, unikanie kontaktu wzrokowego
Otwartość (wysoka)	Abstrakcyjny język, metafory, odniesienia kulturowe	Zróżnicowana intonacja, eksperymentalne modulacje	Ekspresywne gesty, dynamiczne ruchy

3.3.3. Asystenci głosowi w kontekście Automotive

Braun i współpracownicy [18] zaprojektowali cztery osobowości asystenta głosowego (Przyjaciel, Wielbiciel, Ciotka, Lokaj) oraz postać domyślną, opierając się na dwuwymiarowym modelu postaw wobec innych Argyle'a. Osobowości były wyrażane poprzez dobór słów i intonację, przy identycznej treści i wyjściu mowy [18].

Badania wykazały, że spersonalizowane asystenty głosowi zwiększają zaufanie i sympatię, gdy są prawidłowo dopasowane do osobowości użytkownika, ale nieprawidłowe dopasowanie prowadzi do niższej sympatii, zaufania, satysfakcji i użyteczności [18].

3.4. Systemy hybrydowe i modułowe

3.4.1. Ramy modułowe

Nardelli i współpracownicy [12], [19] opracowali modułową architekturę oprogramowania do kodowania wymiarów psychologicznych sztucznego agenta. System wykorzystuje model BERT do kojarzenia zachowań z cechami osobowości

opartymi na modelu Wielkiej Piątki [19]. Architektura jest niezależna od zadania i platformy, co umożliwia łatwą integrację z różnymi systemami robotycznymi i konwersacyjnymi.

3.4.2. Systemy z pamięcią i regulacją emocjonalną

Zaawansowane architektury integrują moduły pamięci i regulacji emocjonalnej, aby poprawić ciągłość persony i długoterminową spójność [10], [12]. Yacoubi [10] zaproponował model afektywny inspirowany teorią Tendencji do Działania, wykorzystujący formalizm inspirowany logiką BDI (Beliefs-Desires-Intentions) do reprezentacji przekonań i celów agenta. Implementacja procesu regulacji emocjonalnej zwiększyła postrzeganą wiarygodność i profesjonalizm agentów [10].

4. Kluczowe odkrycia i analiza porównawcza

4.1. Skuteczność percepcyjna

4.1.1. Rozpoznawalność cech osobowości

Badania konsekwentnie wykazują, że użytkownicy potrafią rozpoznawać i rozróżniać cechy osobowości w agentach AI, szczególnie gdy wykorzystywane są multimodalne systemy ekspresji [2], [5], [19]. Sonlu i współpracownicy [2] wykazali, że modyfikacje oparte na osobowości są postrzegane jako naturalne, a każda dodatkowa modalność komunikacji poprawia dokładność percepcji.

Jednak nie wszystkie cechy są równie łatwe do percepcji. Doce i współpracownicy [5] odkryli, że podczas gdy ekstrawersja, neurotyczność i ugodowość były silnie skorelowane z percepcją użytkowników, sumienność nie wykazywała takiej korelacji, prawdopodobnie dlatego, że wpływała tylko na dwie specyficzne emocje w ich implementacji.

4.1.2. Wpływ na zaangażowanie i zaufanie

Osobowość agenta ma mierzalny wpływ na zaangażowanie użytkowników i zaufanie:

- **Sumienność**: Chatboty o wysokiej sumienności sprzyjają najwyższemu zaangażowaniu użytkowników w kontekście zdrowia psychicznego [20].
- **Ugodowość**: agenci z promptami zaprojektowanymi dla wysokiej ugodowości osiągają 60% wskaźnik pomyłki w teście Turinga, przewyższając 50% dla wszystkich testowanych agentów GPT [21].
- **Poważna, asertywna osobowość**: Użytkownicy są bardziej skłonni zwierzać się i słuchać agenta AI, o poważnej, asertywnej osobowości w kontekście rozmów kwalifikacyjnych o wysokiej stawce [4].

4.2. Walidacja psychometryczna

4.2.1. Korelacje behawioralne

Huang i współpracownicy [14], [15] przeprowadzili kompleksową walidację psychometryczną agentów LLM, wykazując, że:

- Agenci AI skutecznie symulują ludzkie cechy osobowości, wykazując silne dopasowanie do ludzkich odpowiedzi
- Format Rozszerzony przewyższa format Likerta pod względem niezawodności i odporności
- W scenariuszach podejmowania ryzyka zachowanie agentów było równoległe do ludzkiego
- Format BFI-2-Expanded najdokładniej odtwarza ludzkie asocjacje osobowość-decyzja

4.2.2. Struktura czynnikowa

Mimo obiecujących wyników na poziomie agregatowym, szczegółowe analizy ujawniają rozbieżności w strukturze czynnikowej. agenci LLM mogą odtwarzać szerokie korelacje między cechami a zachowaniem, ale drobniejsze struktury czynnikowe i wzorce odpowiedzi często odbiegają od norm ludzkich [14], [15]. To ogranicza użyteczność agentów jako pełnych substytutów w precyzyjnych badaniach psychologicznych.

4.3. Porównanie metod

Tabela 2 przedstawia porównanie trzech głównych podejść technicznych.

Wymiar	Systemy oparte na regułach	Kondycjonowanie LLM	Systemy multimodalne
Kontrola	Bardzo wysoka	Średnia	Średnia do wysokiej
Elastyczność	Niska	Bardzo wysoka	Wysoka
Koszt rozwoju	Bardzo wysoki	Niski do średniego	Wysoki

Spójność	Bardzo wysoka	Zmienna	Wysoka
Autentyczność	Średnia	Wysoka	Bardzo wysoka
Walidacja psychometryczna	Możliwa, ale pracochłonna	Dobra	Doskonała (percepcja)
Skalowalność	Niska	Bardzo wysoka	Średnia
Przykłady	FAtiMA+GRETA [5]	GPT-4o z promptami BFI [14]	Framework Sonlu [2]

4.4. Kontekst aplikacyjny

4.4.1. Zdrowie psychiczne

W kontekście chatbotów zdrowia psychicznego, osobowość odgrywa kluczową rolę w zaangażowaniu użytkowników [20], [22]. Badania wykazują, że:

- Wysoka sumienność sprzyja najwyższemu zaangażowaniu [20].
- Użytkownicy preferują agenci ugodowe i ekstrawertyczne, cytując empatię, człowieczeństwo i kompetencję [6].
- Wymagane są specjalne zabezpieczenia prywatności i etyczne, gdy agenci wnioskują lub adaptują się do osobowości użytkownika [22].

4.4.2. Systemy doradcze i edukacyjne

W systemach doradczych osobowość chatbota znacząco wpływa na postrzeganą autentyczność i zamierzone zaangażowanie, choć nie zawsze na zaufanie lub intencję użycia [6]. Studenci generalnie preferowali ugodowe i ekstrawertyczne chatboty nad sumienne [6].

4.4.3. Asystenci Automotive

W kontekście asystentów głosowych w samochodach, prawidłowe dopasowanie osobowości asystenta do użytkownika zwiększa zaufanie i sympatię, podczas gdy nieprawidłowe dopasowanie prowadzi do negatywnych skutków [18]. Obciążenie poznawcze kierowcy było porównywalne dla wszystkich typów asystentów.

5. Zastosowanie nauk psychologicznych

5.1. Operacjonalizacja teorii psychologicznych

5.1.1. Mapowanie Big Five na zachowania

Projektanci mapują każdy czynnik Wielkiej Piątki na konkretne zachowania [1, 2, 5]:

- **Ekstrawersja → Styl komunikacji:** częstotliwość inicjowania rozmów, długość wypowiedzi, ekspresywność języka.
- **Sumienność → Planowanie i organizacja**: strukturyzacja odpowiedzi, punktualność, dokładność.
- **Ugodowość → Interakcje społeczne**: empatyczne sformułowania, unikanie konfrontacji, współpraca.
- **Neurotyczność → Regulacja emocjonalna**: wyrażanie niepewności, reaktywność emocjonalna, strategie radzenia sobie.
- **Otwartość → Treść poznawcza**: abstrakcyjny język, kreatywność, ciekawość intelektualna.

5.1.2. Wykorzystanie inwentarzy psychometrycznych

Walidowane inwentarze są używane jako prompty lub instrumenty ewaluacyjne, dzięki czemu agenci mogą być przypisywane i mierzone na standaryzowanych skalach cech [14], [15]. Podejście psychometryczne Huanga wykorzystuje BFI-2 (Big Five Inventory-2) do przypisywania kwantyfikowalnych, kontrolowalnych i psychometrycznie walidowanych cech osobowości agentom LLM [14], [15].

5.2. Dowody na charakterystyki podobne do ludzkich

5.2.1. Postrzeganie osobowości AI przez użytkowników

Multimodalne manipulacje cechami prowadzą do rozpoznawalnej percepcji cech w badaniach użytkowników, z połączonymi modalnościami produkującymi najdokładniejszą percepcję cech [2], [8], [11]. Nardelli i współpracownicy [19] wykazali, że ich framework generuje osobowości wyraźnie postrzegane przez użytkowników, z silnie rozróżnialnymi wymiarami osobowości, co potwierdzono w eksperymencie pilotażowym z robotem humanoidalnym.

5.2.2. Odtwarzanie Korelacji Behawioralnych

Agenci oparci na LLM, kondycjonowane metodami psychometrycznymi, odtwarzają znane korelacje między cechami osobowości a zachowaniami decyzyjnymi w zadaniach winietek, wskazując na pewną trafność konstruktu [15]. W scenariuszach podejmowania ryzyka zachowanie agentów było równoległe do ludzkiego, co sugeruje, że agenci mogą symulować ludzkie wzorce decyzyjne związane z osobowością [14], [15].

5.2.3. Wpływ na zaangażowanie i zaufanie

agenci zgodne z cechami wpływają na zaangażowanie użytkowników i zaufanie w przewidywalny sposób [4], [6], [18]. Na przykład:

- Sumienne chatboty zwiększają zaangażowanie w kontekście zdrowia psychicznego [20].
- Poważni ankieterzy zwiększają gotowość do zwierzeń w rozmowach kwalifikacyjnych [4].
- Prawidłowo dopasowani asystenci głosowi zwiększają zaufanie i sympatię [18].

5.3. Luki pomiędzy psychologią ludzką, a sztuczną

5.3.1. Niedopasowanie drobniejszej struktury

Nawet gdy szerokie korelacje cech są dopasowane, ładunki czynnikowe i drobniejsze wzorce odpowiedzi różnią się od ludzkich, ograniczając użyteczność agentów jako pełnych substytutów w precyzyjnych badaniach [14], [15]. Huang i współpracownicy [15] zauważają, że obecne metody mogą nie w pełni uchwycić skomplikowanego wzajemnego oddziaływania osobowości, indywidualnych tła i złożoności społecznej.

5.3.2. Brak głębi generowania/regulacji emocji

Obliczeniowe modele emocji często implementują powierzchowne reguły ekspresyjne lub proste heurystyki regulacyjne, zamiast głębokich dynamik motywacyjnych opisanych w teorii psychologicznej [13], [12]. Schönbrodt i współpracownicy [13] podkreślają, że wiele teorii psychologicznych brakuje szczegółów algorytmicznych wymaganych do wiernej instancjacji obliczeniowej, co prowadzi projektantów do używania uproszczonych lub ad-hoc modeli.

5.3.3. Spójność czasowa i sytuacyjna

Agenci AI mają trudności z utrzymaniem zachowania zgodnego z osobowością w długim okresie bez wyraźnych architektur pamięci i regulacji [9], [11]. CharacterChat adresuje to poprzez włączenie dynamicznej pamięci i zachowań opartych na personach [9], ale pozostaje to znaczącym wyzwaniem dla większości systemów.

5.4. Badania oceniające trafność psychologiczną

5.4.1. Studia walidacji psychometrycznej

Wielostopniowe prace pokazują, że LLM uchwytują relacje semantyczne między miarami Wielkiej Piątki i że psychometrycznie zaprojektowane prompty produkują agenci, których zagregowane zachowanie koreluje z ludzkimi wzorcami, choć rozbieżności pozostają na drobniejszych poziomach [14, 15]. Kruijssen i współpracownicy [3, 17] wykazali, że zaawansowane modele mogą wyrażać deterministyczne i spójne osobowości przy użyciu ustalonych ram psychologicznych, z ekspresją osobowości wyłaniającą się z holistycznego rozumowania.

5.4.2. Wdrożenia w rzeczywistych warunkach

Rzeczywiste wdrożenia w ankietowaniu, doradztwie i asystentach samochodowych demonstrują mierzalne efekty osobowości na wyniki użytkowników, wspierając trafność ekologiczną projektów opartych na cechach, jednocześnie ujawniając wrażliwość kontekstową [4], [18], [22].

5.5. Skuteczność zastosowania nauk psychologicznych

Zastosowanie rzeczywistych nauk psychologicznych do tworzenia psychiki AI przynoszą częściowy sukces:

Sukcesy	Ograniczenia
• Agenci mogą skutecznie symulować cechy osobowości na poziomie percepcji użytkowników • Szerokie korelacje behawioralne są odtwarzane • Mierzalne efekty na zaangażowanie, zaufanie i jakość interakcji • Walidacja psychometryczna potwierdza podstawową trafność konstruktu	• Drobniejsze struktury czynnikowe odbiegają od norm ludzkich • Brak głębi w regulacji emocjonalnej i motywacyjnej • Trudności ze spójnością czasową i sytuacyjną • Niepełne uchwycenie subtelności ludzkiej psychologii

Jeśli wymagana jest większa precyzja lub substytucja w wysokiej stawce, literatura konkluduje, że obecne podejścia są obiecujące, ale niewystarczające jako pełne psychologiczne ekwiwalenty [14, 15].

6. Wyzwania i ograniczenia

6.1. Ograniczenia techniczne i wyzwania spójności

6.1.1. Wewnętrzne uprzedzenia modeli i stabilność

Wiele otwartych LLM wykazuje trwałe "wewnętrzne" osobowości, które opierają się kondycjonowaniu, więc kondycjonowanie promptami daje zmienne sukcesy w różnych modelach [16]. Cava i współpracownicy [16] odkryli, że niektóre modele (SOLAR, Dolphin, NeuralChat, Llama3-8) dobrze naśladują osobowości, podczas gdy inne pozostają "zamknięte umysłowo" i wykazują silne wewnętrzne preferencje osobowościowe, z ENFJ jako dominującą wewnętrzną osobowością.

6.1.2. Dostrajanie vs. kontrola stylu

Dostrajanie często zmienia styl komunikacji niezależnie od psychometrycznej ekspresji cech, komplikując spójną implementację cech [3]. Kruijssen i współpracownicy [3] wykazali, że dostrajanie modelu wpływa na styl komunikacji niezależnie od dokładności ekspresji osobowości, co sugeruje, że te dwa aspekty mogą wymagać oddzielnych mechanizmów kontroli.

6.1.3. Złożoność fuzji multimodalnej

Łączenie modalności poprawia percepcję, ale podnosi złożoność synchronizacji, adnotacji i ewaluacji w różnych kanałach [2], [8]. Sonlu i współpracownicy [2] opracowali kompleksową ramę integrującą zachowania niewerbalne i werbalne, ale implementacja wymagała znaczących wysiłków w zakresie koordynacji między modalnościami.

6.1.4. Wariancja ewaluacji

Wariancja odpowiedzi na pojedyncze pytania jest wyższa niż metryki skali testowej, wskazując na niestabilne wyniki per-pytanie, nawet gdy zagregowane wyniki cech są dopasowane [3]. To sugeruje, że agenci mogą nie być niezawodne w kontekstach wymagających spójnych odpowiedzi na poziomie pojedynczych interakcji.

6.2. Teoretyczne ograniczenia psychologiczne

6.2.1. Niepełna operacjonalizacja teorii psychologicznej

Wiele teorii psychologicznych brakuje szczegółów algorytmicznych wymaganych do wiernej instancjacji obliczeniowej, prowadząc projektantów do używania uproszczonych lub ad-hoc modeli [13]. Schönbrodt i współpracownicy [13] argumentują, że wyzwanie konstruowania psychologicznie wiarygodnych agentów wymaga bliższej współpracy między psychologami a informatykami, aby rozwijać teorie, które są zarówno psychologicznie trafne, jak i obliczeniowo implementowalne.

6.2.2. Niedoskonała trafność konstruktu

agenci oparte na LLM mogą odtwarzać szerokie korelacje między cechami a zachowaniem, ale drobniejsze struktury czynnikowe i wzorce odpowiedzi często odbiegają od norm ludzkich [14], [15]. To ogranicza ich użyteczność jako pełnych substytutów w precyzyjnych badaniach psychologicznych lub aplikacjach wymagających wysokiej wierności psychologicznej.

6.2.3. Generalizacja kulturowa i językowa

Wskazówki osobowości nie przenoszą się prosto między językami i kulturami, komplikując wielojęzyczny projekt person [23]. Danielescu i współpracownicy [23] wykazali, że projektowanie osobowości dla wielojęzycznych agentów

konwersacyjnych wymaga uwzględnienia specyficznych dla języka i kultury wskazówek osobowości, które mogą nie być uniwersalne.

6.3. Autentyczność, wiarygodność i człowieczeństwo

6.3.1. Realizm percepcyjny vs. wewnętrzna spójność

Użytkownicy postrzegają osobowości z powierzchownych wskazówek (język, prozodia, gest), ale te percepcje nie gwarantują wewnętrznej spójności cech w różnych kontekstach lub czasie [8], [11]. To tworzy ryzyko, że agenci mogą być postrzegane jako posiadające spójną osobowość, podczas gdy ich wewnętrzne zachowanie jest niespójne.

6.3.2. Efekty roli i ograniczona generalizacja

Kondycjonowanie przez rolę plus osobowość poprawia naśladowanie dla niektórych kombinacji rola-cecha, ale nie działa jednolicie w różnych osobowościach lub zadaniach [16]. To sugeruje, że skuteczność kondycjonowania osobowości może być silnie zależna od kontekstu.

6.3.3. Pomyłka ludzka i antropomorfizacja

Prompty wysokiej ugodowości mogą zwiększać oceny człowieczeństwa lub pomyłkę Turinga, podnosząc obawy o nierozróżnialność bez jasnych granic [21]. León-Domínguez i współpracownicy [21] wykazali, że agenci AI z promptami zaprojektowanymi dla wysokiej ugodowości osiągają 60% wskaźnik pomyłki w teście Turinga, co rodzi pytania etyczne o granice między ludźmi a maszynami.

6.4. Obawy etyczne wspierane w literaturze

6.4.1. Ryzyko zaufania i oszustwa

Nieprawidłowo dopasowane lub wysoce człowieczeńskie osobowości mogą zmieniać zaufanie i gotowość do zwierzeń; niedopasowania zmniejszają zaufanie, a dopasowane persony mogą nieumyślnie zachęcać do nadmiernego polegania [18], [11]. Braun i współpracownicy [18] wykazali, że nieprawidłowe dopasowanie osobowości prowadzi do niższej sympatii, zaufania, satysfakcji i użyteczności.

6.4.2. Prywatność i profilowanie

Systemy, które wnioskują lub adaptują się do osobowości użytkownika, podnoszą obawy o prywatność i profilowanie i wymagają wyraźnych wymagań i zabezpieczeń w wrażliwych domenach, takich jak zdrowie psychiczne [22]. Siemon i współpracownicy [22] identyfikują wymagania i podejścia rozwiązań dla agentów konwersacyjnych adaptujących się do osobowości w opiece zdrowia psychicznego, podkreślając potrzebę przejrzystych ujawnień i ochrony prywatności.

6.4.3. Bezpieczeństwo i artefakty wyrównania wartości

Modele wyrównane pod względem bezpieczeństwa lub filtrowane pod względem treści mogą zniekształcać wyniki moralne lub decyzyjne, zawyżając pewne oceny i stronniczo wpływając na zachowanie w porównaniu z ludźmi [14]. Huang i współpracownicy [14] zauważają rozbieżności w odpowiedziach na dylematy etyczne, sugerując, że LLM mogą nie w pełni uchwytywać niuansowanego rozumowania moralnego lub kontekstów kulturowych, częściowo z powodu dostrajania bezpieczeństwa.

6.5. Podsumowanie ograniczeń

Tabela 3 podsumowuje główne kategorie ograniczeń i ich implikacje.

Kategoria	Kluczowe ograniczenia	Implikacje praktyczne
Techniczne	Wewnętrzne uprzedzenia modeli, niestabilność per-pytanie, złożoność multimodalna	Wymagane jest staranne testowanie i walidacja dla każdego modelu i kontekstu
Psychologiczne	Niepełna operacjonalizacja teorii, niedoskonała trafność konstruktu	Potrzebna ściślejsza współpraca między psychologami a informatykami
Kulturowe	Brak generalizacji między językami i kulturami	Wymagana lokalizacja i walidacja dla każdego kontekstu kulturowego
Autentyczność	Rozbieżność między percepcją a wewnętrzną spójnością	Potrzebne mechanizmy zapewniające długoterminową spójność
Etyczne	Ryzyko zaufania, prywatność, manipulacja	Wymagane przejrzyste ujawnienia i zabezpieczenia projektowe

7. Dyskusja

7.1. Stan obecny: obiecujące podstawy, znaczące luki

Przegląd literatury ujawnia, że dziedzina tworzenia person AI osiągnęła znaczący postęp w ciągu ostatniej dekady, szczególnie z pojawieniem się dużych modeli językowych. Obecne systemy mogą skutecznie symulować cechy osobowości na poziomie percepcji użytkowników i generować mierzalne efekty na zaangażowanie, zaufanie i jakość interakcji [2], [4], [6], [14], [15], [18], [20].

Jednak fundamentalne luki pozostają między sztuczną a ludzką psychiką. agenci AI mogą odtwarzać szerokie korelacje behawioralne, ale brakuje im głębi regulacji emocjonalnej, subtelności struktury czynnikowej i spójności czasowej charakterystycznej dla ludzkiej osobowości [13,14,15]. To sugeruje, że obecne podejścia są najbardziej odpowiednie dla aplikacji, gdzie percepcyjna wiarygodność i szerokie wzorce behawioralne są wystarczające, ale mogą być nieodpowiednie dla kontekstów wymagających wysokiej wierności psychologicznej.

7.2. Napięcie między kontrolą, a autentycznością

Istnieje fundamentalne napięcie między kontrolą a autentycznością w projektowaniu person AI.

- Systemy oparte na regułach oferują wysoką kontrolę i przewidywalność, ale mogą wydawać się sztywne i nieautentyczne [5], [13].
- Systemy oparte na LLM oferują większą autentyczność i elastyczność, ale z kosztem zmniejszonej kontroli i przewidywalności [3], [14], [15], [16].
- Systemy multimodalne mogą osiągnąć wysoką autentyczność percepcyjną, ale z znaczną złożonością implementacyjną [2], [8].

Optymalne podejście prawdopodobnie zależy od kontekstu aplikacyjnego. Dla aplikacji wysokiej stawki wymagających przewidywalnego zachowania (np. systemy medyczne, doradztwo finansowe), systemy oparte na regułach lub hybrydowe mogą być preferowane. Dla aplikacji wymagających naturalnej, angażującej interakcji (np. towarzysze, rozrywka), systemy oparte na LLM mogą być bardziej odpowiednie.

7.3. Rola walidacji psychometrycznej

Walidacja psychometryczna wyłania się jako kluczowy element w projektowaniu wiarygodnych person AI [14], [15]. Podejście Huanga i współpracowników [14], [15] wykorzystujące ustalone inwentarze osobowości jako podstawę do projektowania i ewaluacji agentów reprezentuje znaczący postęp w kierunku bardziej rygorystycznego, opartego na nauce podejścia do tworzenia person AI.

Jednak walidacja psychometryczna ujawnia również ograniczenia obecnych podejść, szczególnie w zakresie drobniejszych struktur czynnikowych i rozumowania moralnego [14], [15]. To sugeruje potrzebę ciągłego udoskonalania metod i bliższej współpracy między psychologami a informatykami.

7.4. Implikacje etyczne i społeczne

Zdolność do tworzenia wysoce wiarygodnych, człowieczeńskich person AI rodzi istotne pytania etyczne:

- **Granice między ludźmi a maszynami**: Gdy agenci AI osiągają 60% wskaźnik pomyłki w teście Turinga [21], gdzie powinniśmy rysować granice?
- **Zaufanie i oszustwo**: Jak zapewnić, że użytkownicy nie są wprowadzani w błąd co do natury agentów, z którymi wchodzą w interakcję? [18], [22]
- **Prywatność i profilowanie**: Jakie zabezpieczenia są potrzebne, gdy systemy wnioskują i adaptują się do osobowości użytkowników? [22]
- **Manipulacja i wpływ**: Jak zapobiegać wykorzystywaniu person AI do manipulacji lub nieodpowiedniego wpływu? [21]

Te pytania wymagają interdyscyplinarnego dialogu obejmującego informatyków, psychologów, etyków i decydentów politycznych.

7.5. Przyszłość Person AI: W kierunku zintegrowanego podejścia

Przyszłość tworzenia person AI prawdopodobnie leży w zintegrowanym podejściu, które łączy:

- Walidację psychometryczną dla zapewnienia trafności konstruktu [14], [15].
- Multimodalne systemy ekspresji dla autentyczności percepcyjnej [2], [8].

- Modułowe architektury dla kontroli i konsystencji [12], [19].
- Mechanizmy pamięci i regulacji dla spójności czasowej [9], [10], [12].
- Zabezpieczenia etyczne dla odpowiedzialnego wdrożenia [22].

Takie zintegrowane podejście mogłoby potencjalnie przezwyciężyć wiele obecnych ograniczeń, jednocześnie zachowując mocne strony różnych metod.

8. Przyszłe kierunki i rekomendacje

8.1. Obiecujące podejścia techniczne i metodologiczne

8.1.1. Łączenie ekspresji multimodalnej z ugruntowaniem psychometrycznym

Integracja walidowanego promptowania/dostrajania psychometrycznego z zsynchronizowanymi wskazówkami werbalnymi, paralingwistycznymi i motorycznymi w celu poprawy zarówno percepcji, jak i wewnętrznej spójności [2], [8], [9]. Przyszłe systemy powinny:

- Wykorzystywać inwentarze psychometryczne (np. BFI-2) do definiowania profili osobowości [14, 15].
- Mapować te profile na wszystkie modalności ekspresji (język, głos, gesty, mimika) [2].
- Walidować zarówno percepcję użytkowników, jak i wewnętrzną spójność behawioralną.

8.1.2. Personalizacja oparta na zbiorach danych

Wykorzystanie długich danych osobistych (dzienniki, pisma reprezentatywne dla klastrów) i korpusów specyficznych dla person do dostrajania modeli dla bogatszych, ewoluujących osobowości [7]. Podejście Pala [7] wykorzystujące wpisy dziennikowe z Reddit demonstruje potencjał tego podejścia, osiągając 11% poprawę w uchwyceniu cech osobowości.

8.1.3. Kondycjonowanie roli i osobowości

Wspólne kondycjonowanie na roli i osobowości w celu poprawy naśladowania dla zachowań dopasowanych do zadania, ponieważ kondycjonowanie roli może zwiększyć skuteczną emulację cech w wielu modelach [16]. Przyszłe badania powinny systematycznie badać, które kombinacje rola-osobowość są najbardziej skuteczne dla różnych aplikacji.

8.1.4. Modułowe silniki osobowości i systemy pamięci

Oddzielenie modułów osobowości, zadania i pamięci, aby cechy systematycznie wpływały na planowanie, regulację emocji i długoterminową spójność [5], [10], [11], [12]. Kluczowe komponenty powinny obejmować:

- **Moduł osobowości**: definiuje cechy i ich wpływ na procesy poznawcze.
- **Moduł emocjonalny**: generuje i reguluje emocje zgodnie z osobowością.
- **Moduł pamięci:** przechowuje zgodne z cechami wspomnienia i doświadczenia.
- **Moduł planowania**: generuje zachowania zgodne z osobowością.

8.2. Rekomendacje dla projektowania badań

8.2.1. Używanie walidacji psychometrycznej i winietek behawioralnych

Ewaluacja agentów z ustalonymi inwentarzami i zadaniami decyzyjnymi w celu walidacji korespondencji na poziomie konstruktu, nie tylko subiektywnych ocen [14], [15]. Protokoły ewaluacji powinny obejmować:

- **Testy osobowości**: BFI-2, NEO-PI-R lub inne walidowane inwentarze.
- **Winiety behawioralne**: scenariusze podejmowania ryzyka, dylematy moralne, zadania interpersonalne.
- **Oceny użytkowników**: percepcja cech, zaufanie, zaangażowanie, satysfakcja.
- **Analiza spójności**: spójność czasowa, spójność sytuacyjna, wewnętrzna konsystencja.

8.2.2. Testowanie międzykulturowe i wielojęzyczne

Walidacja wskazówek person w różnych językach i kulturach, aby uniknąć błędnej generalizacji i zapewnić lokalizację wskazówek osobowości [23]. Badania powinny:

- Testować te same profile osobowości w wielu językach i kulturach.
- Identyfikować uniwersalne vs. specyficzne dla kultury wskazówki osobowości.
- Rozwijać wytyczne dla lokalizacji person AI.

8.2.3. Przejrzyste ujawnienia i zabezpieczenia dla użytkowników

Wdrażanie jasnych ujawnień, ochrony prywatności i kontroli, gdy agenci wnioskują lub adaptują się do osobowości użytkownika, szczególnie w kontekstach zdrowia psychicznego [22]. Najlepsze praktyki powinny obejmować:

- **Przejrzyste ujawnienia**: jasne informowanie użytkowników, że wchodzą w interakcję z AI.
- **Kontrola prywatności:** pozwalanie użytkownikom kontrolować, jakie dane są zbierane i jak są używane.
- **Mechanizmy zgody**: uzyskiwanie wyraźnej zgody na wnioskowanie i adaptację osobowości.
- **Zabezpieczenia etyczne**: mechanizmy zapobiegające manipulacji lub nieodpowiedniemu wpływowi.

8.2.4. Współpraca interdyscyplinarna

Angażowanie psychologów wcześnie w projektowanie modelu, aby wybierać zoperacjonalizowane teorie (np. model motywacji społecznej Zurychu, gdy jest to odpowiednie) i kierować ewaluacjami ekologicznymi [13], [19]. Skuteczna współpraca wymaga:

- **Wspólnego języka**: rozwijanie wspólnego słownictwa między psychologami a informatykami
- **Iteracyjnego projektowania:** cykle projektowania, implementacji, testowania i udoskonalania
- **Walidacji ekologicznej:** testowanie w rzeczywistych kontekstach aplikacyjnych
- **Rozwoju teorii:** rozwijanie teorii psychologicznych, które są zarówno trafne, jak i obliczeniowo implementowane

8.3. Priorytety końcowe

8.3.1. Dążenie do mierzalnej trafności konstruktu

Dążenie do mierzalnej trafności konstruktu, a nie tylko realizmu percepcyjnego; łączenie metod psychometrycznych z walidacjami longitudinalnymi i behawioralnymi [14], [15]. To wymaga:

- Rygorystycznych protokołów walidacji psychometrycznej.
- Długoterminowych badań spójności.
- Behawioralnych testów w różnorodnych kontekstach.
- Porównań z ludzkimi normami.

8.3.2. Łagodzenie szkód etycznych przez projektowanie

Ustalanie granic realistycznego naśladowania, ochrona danych użytkowników i testowanie efektów manipulacyjnych w kontekstach wdrożenia [22], [18]. Kluczowe strategie obejmują:

- **Projektowanie dla przejrzystości**: jasne sygnalizowanie natury AI.
- **Ograniczenia naśladowania**: unikanie nadmiernego człowieczeństwa w wrażliwych kontekstach.
- **Ochrona prywatności**: minimalizacja zbierania i przechowywania danych osobowych.
- **Testowanie etyczne**: systematyczne testowanie potencjalnych szkód przed wdrożeniem.

8.3.3. Iteracja na ewaluacji

Używanie zarówno eksperymentów z udziałem ludzi, jak i automatycznych kontroli psychometrycznych, aby agenci zbiegały się do niezawodnych, wrażliwych na kontekst zachowań osobowościowych [3], [15]. Proces iteracyjny powinien obejmować:

- **Projektowanie początkowe**: oparte na teoriach psychologicznych i najlepszych praktykach.
- **Implementacja:** tworzenie prototypu agenta.
- **Walidacja psychometryczna**: testowanie z inwentarzami i winiętami.
- **Testowanie użytkowników**: ewaluacja percepcji i efektywności.
- **Analiza i udoskonalenie**: identyfikacja problemów i iteracja.
- **Wdrożenie i monitorowanie:** ciągłe monitorowanie w rzeczywistych warunkach.

9. Wnioski

Tworzenie person AI i nadawanie osobowości agentom cyfrowym stanowi złożone, interdyscyplinarne wyzwanie na przecięciu informatyki, psychologii, lingwistyki i etyki. Niniejszy przewodnik przeanalizował obecny stan wiedzy oparty na systematycznym przeglądzie 224 publikacji naukowych, ze szczególnym uwzględnieniem 30 najbardziej relewantnych prac.

9.1. Kluczowe wnioski

1. Trzy główne podejścia techniczne dominują w dziedzinie: systemy oparte na regułach, kondycjonowanie i dostrajanie LLM oraz multimodalne systemy ekspresji. Każde ma unikalne mocne strony i ograniczenia, z rosnącym trendem w kierunku zintegrowanych, hybrydowych podejść.

2. Model Wielkiej Piątki stanowi najczęściej stosowaną ramę psychologiczną, z rosnącym wykorzystaniem rygorystycznych metod psychometrycznych do projektowania i walidacji person AI.
3. agenci AI mogą skutecznie symulować cechy osobowości na poziomie percepcji użytkowników i generować mierzalne efekty na zaangażowanie, zaufanie i jakość interakcji, szczególnie gdy wykorzystywane są multimodalne systemy ekspresji.
4. Znaczące luki pozostają między sztuczną a ludzką psychiką, szczególnie w zakresie drobniejszych struktur czynnikowych, głębi regulacji emocjonalnej, spójności czasowej i rozumowania moralnego.
5. Próby zastosowania rzeczywistych nauk psychologicznych przynoszą obiecujące rezultaty, ale ujawniają również fundamentalne wyzwania w operacjonalizacji teorii psychologicznych i osiągnięciu pełnej trafności konstruktu.
6. Istotne kwestie etyczne obejmują granice między ludźmi, a maszynami, ryzyko zaufania i oszustwa, prywatność i profilowanie oraz potencjał manipulacji.

9.2. Implikacje praktyczne

Dla praktyków projektujących systemy person AI, kluczowe rekomendacje obejmują:

- Wybór odpowiedniego podejścia w oparciu o kontekst aplikacyjny, wymagania dotyczące kontroli vs. autentyczności oraz dostępne zasoby
- Wykorzystanie walidacji psychometrycznej dla trafności konstruktu i mierzalnej skuteczności
- Integracja wielu modalności ekspresji dla poprawy autentyczności percepcyjnej
- Implementacja mechanizmów pamięci i regulacji dla długoterminowej spójności
- Priorytetyzacja przejrzystości i zabezpieczeń etycznych w projektowaniu i wdrożeniu

9.3. Przyszłe perspektywy

Dziedzina tworzenia person AI znajduje się w punkcie przełomowym. Pojawienie się zaawansowanych dużych modeli językowych otwiera nowe możliwości dla bardziej autentycznych, elastycznych i skalowalnych person AI. Jednocześnie rosnąca świadomość ograniczeń i ryzyk etycznych wymaga bardziej rygorystycznego, opartego na nauce podejścia. Przyszłość prawdopodobnie przyniesie:

- Bardziej zintegrowane podejścia łączące mocne strony różnych metod.
- Ściślejszą współpracę między psychologami, a informatykami.
- Bardziej rygorystyczne standardy walidacji psychometrycznej i etycznej.
- Większą świadomość ograniczeń i odpowiednich kontekstów zastosowania.
- Rozwój nowych teorii psychologicznych specyficznie zaprojektowanych dla implementacji obliczeniowej

9.4. Refleksja końcowa

Tworzenie person AI nie jest jedynie technicznym wyzwaniem, ale fundamentalnym pytaniem o naturę osobowości, świadomości i człowieczeństwa. W miarę jak nasze systemy stają się coraz bardziej wyrafinowane, musimy pozostać czujni wobec zarówno możliwości, jak i ryzyk. Cel nie powinien być stworzenie doskonałych replik ludzkiej psychiki – co może być niemożliwe i niepożądane – ale raczej stworzenie agentów, które są autentyczne, użyteczne, etyczne i odpowiednie dla swoich zamierzonych kontekstów.

Ostatecznie, sukces person AI będzie mierzony nie tym, jak dobrze oszukują test Turinga, ale tym, jak dobrze służą ludzkim potrzebom, wspierają znaczące interakcje i przyczyniają się do dobrostanu jednostek i społeczeństw.

Bibliografia

[1] S. T. Völkel, "Conversational agents with personality," (data publikacji niedostępna).

[2] E. Sonlu et al., "A Conversational Agent Framework with Multi-modal Personality Expression," *ACM Transactions on Graphics*, 2021.

[3] M. Kruijssen et al., "Deterministic AI Agent Personality Expression through Standard Psychological Diagnostics," 2025.

[4] S. Zhou et al., "Trusting Virtual Agents: The Effect of Personality," 2019.

[5] T. Doce et al., "Creating individual agents through personality traits," in *Intelligent Virtual Agents*, 2010.

[6] M. A. Kuhail et al., "Interacting with a chatbot-based advising system: Understanding the effect of chatbot personality and user gender on behavior," *Informatics (Basel)*, 2022.

[7] A. Pal et al., "Beyond Discrete Personas: Personality Modeling Through Journal Intensive Conversations," 2024.

[8] C. Clavel et al., "Artificial companions with personality and social role," 2013.

[9] Q. Tu et al., "Characterchat: Learning towards conversational ai with personalized social support," *arXiv.org*, 2023.

[10] J. Yacoubi, "Vers des agents conversationnels capables de réguler leurs émotions : un modèle informatique des tendances à l'action," 2019.

[11] C. Clavel et al., "Artificial companions with personality and social role," 2013.

[12] M. Nardelli et al., "Personality-and memory-based framework for emotionally intelligent agents," (data publikacji niedostępna).

[13] F. D. Schönbrodt et al., "The Challenge of Constructing Psychologically Believable Agents," *Journal of Media Psychology*, 2011.

[14] J. Huang et al., "Designing LLM-Agents with Personalities: A Psychometric Approach," 2024.

[15] J. Huang et al., "Designing AI-Agents with Personalities: A Psychometric Approach," 2025.

[16] L. Cava et al., "Open Models, Closed Minds? On Agents Capabilities in Mimicking Human Personalities through Open Large Language Models," 2024.

[17] M. Kruijssen et al., "Deterministic AI Agent Personality Expression through Standard Psychological Diagnostics," 2025.

[18] M. Braun et al., "At Your Service: Designing Voice Assistant Personalities to Improve Automotive User Interfaces," in *Human Factors in Computing Systems*, 2019.

[19] M. Nardelli et al., "A Software Framework to Encode the Psychological Dimensions of an Artificial Agent," 2023.

[20] J. Moilanen et al., "Measuring the effect of mental health chatbot personality on user engagement," in *International Conference on Mobile and Ubiquitous Multimedia*, 2022.

[21] U. León-Domínguez et al., "AI-Driven Agents with Prompts Designed for High Agreeableness Increase the Likelihood of Being Mistaken for a Human in the Turing Test," 2024.

[22] D. Siemon et al., "Requirements and Solution Approaches to Personality-Adaptive Conversational Agents in Mental Health Care," *Sustainability*, 2022.

[23] A. Danielescu et al., "A Bot is Not a Polyglot: Designing Personalities for Multi-Lingual Conversational Agents," in *Human Factors in Computing Systems*, 2018.