OPEN ACCESS
perspektywy kultury / *perspectives* on culture
No. 45 (2/2024)

Krzysztof Sołoducha
http://orcid.org/0000-0003-1351-5487
Wojskowa Akademia Techniczna
krzysztof.soloducha@wat.edu.pl
Paweł Stacewicz
http://orcid.org/0000-0003-2500-4086
Politechnika Warszawska
pawel.stacewicz@pw.edu.pl
DOI: 10.35765/pk.2024.4502.28

Testy modelu świata i teorii umysłu jako podstawa budowy zaufania do podmiotów moralnych typu AGI

Simulation is not duplication and syntax is not semantics.
John Searle

STRESZCZENIE

W artykule rozważany jest problem budowania zaufania do artefaktów obliczeniowych typu AGI (ang. Artificial General Intelligence), z etycznego punktu widzenia określanych jako podmioty moralne *explicite*. W wyniku analizy opartej na badaniach literatury oraz przedstawieniu aktualnych trendów w rozwoju systemów AI wskazanych jest kilka warunków skonstruowania testów behawioralnych niezbędnych do sprawdzania poprawności ich działania z etycznego punktu widzenia. Przeprowadzenie takich testów powinno ułatwić procedury etycznej aprobaty systemów AGI zarówno na poziomie wytwórców, indywidualnego użytkownika, jak i jednostek certyfikujących.

SŁOWA KLUCZE: etyka sztucznej inteligencji, zaufanie do technologii, sztuczny podmiot moralny, test teorii umysłu, test modelu świata

ABSTRACT

Tests of World Model and Theory of Mind as Tools for Building Trust in Moral AGI Agents

The paper considers the problem of building trust in computational artefacts of the AGI (Artificial General Intelligence) type, which are defined from an ethical point of view as *explicite* moral agents. As a result of an analysis based on research of the literature and current trends in the development of AGI systems, several conditions have been presented for the construction

Sugerowane cytowanie: Sołoducha, J. i Stacewicz, P. (2024). Testy modelu świata i teorii umysłu jako podstawa budowy zaufania do podmiotów moralnych typu Agi. *Perspektywy Kultury*, 2(45), ss. 391–403. DOI: 10.35765/pk.2024.4502.28
Nadesłano: 22.10.2023 Zaakceptowano: 10.04.2024

of behavioural tests necessary to check the correctness of their functioning considered both from an ethical and social point of view. Conducting of such tests should simplify the market approval procedures of AGI systems at the level of manufacturers, the individual users and certification authorities.

KEYWORDS: ethics of artificial intelligence, trust in technology, artificial moral agent, theory of mind test, world model test

Sztuczna inteligencja to pojęcie posiadające wiele definicji, ale w ujęciu Stuarta Russella (Russell, 2023) oznacza ono system obliczeniowy posiadający możliwość rozwiązywania problemów pojawiających się w otoczeniu artefaktów obliczeniowych na podstawie informacji dostarczanych przez ich sensory przy wykorzystaniu zasobów wewnętrznych – algorytmów uczących i predykcyjnych – w sposób autonomiczny. Ogólna sztuczna inteligencja to z kolei pojęcie odnoszące się do możliwości rozwiązywania przez te systemy problemów maksymalnie różnorodnych. Autonomia zaś oznacza dyspozycję do rozwiązywania problemów na podstawie wewnętrznie tworzonych celów i metod, niewymagających zadaniowania z zewnątrz (Russel, 2023).

Te możliwości artefaktów obliczeniowych są realizowane dzięki symulacji procesów zachodzących w systemach naturalnej inteligencji poprzez wykorzystanie metod uczenia maszynowego obejmujących między innymi technologię sztucznych sieci neuronowych oraz algorytmów ewolucyjnych (Gryz, 2021). Ostatnie, spektakularne sukcesy tej technologii w postaci systemów LLM (*Large Language Models*) możliwe są z kolei dzięki stworzeniu nowej architektury głębokich sieci neuronowych opartych na mechanizmach *transformers* i *attention* oraz algorytmie wstecznej propagacji błędów (Vaswani i in., 2017). Praktyczne działanie tych technologii możliwe jest dzięki dostępowi do wielkich mocy obliczeniowych oraz zbiorów danych, na których odbywa się trening systemów obejmujący uczenie bez nadzoru (*unsupervised learning*), uczenie nadzorowane (*supervised learning*) oraz uczenie przez wzmacnianie (*reinforcement training*) (Wolfram, 2023).

W najnowszych badaniach nad systemami sztucznej inteligencji klasyczne pojęcie alienacji, czyli analizy różnicy pomiędzy intencjami twórców a rzeczywistym kształtem tworzonych przez nich artefaktów, pojawia się w kontekście rozważań nad tzw. problemem *alignement* – dostosowania działania inteligentnego systemu obliczeniowego do wartości oraz reguł działania uważanych za zgodne ze społecznie akceptowaną teleologią naturalnych systemów inteligencji, jakimi są ludzie (Christian, 2020). Trzeba oczywiście dodać, że ta teleologia nie jest uniwersalna i jest możliwa jej różnorodna redukcja, np. do pewnych form kulturowych

i cywilizacyjnych, jak chociażby w teorii Feliksa Konecznego (Koneczny, 1935) czy w teorii wpływów kulturowych Ingleharta i Welzela (Inglehart i Welzel, 2005) oraz koncepcji multikulturowego zarządzania Hofstedego (Hofstede, 2005).

Problem zaufania i zaufania aktywnego

Stosowanie przez systemy sztucznej inteligencji kryteriów etycznych do podejmowania działania jest związane z zagadnieniem budowania zaufania do technologii stosowanych w życiu codziennym. Tą problematyką zajmuje się zarówno filozofia, jak i inne dziedziny nauki, w tym nauki o zarządzaniu, ekonomia, socjologia, psychologia czy nauki o polityce (Ejdys, 2017). Zaufanie do technologii ma przy tym zasadniczo inny charakter niż zaufanie interpersonalne – zaufanie do naturalnych podmiotów inteligencji (Stacewicz, 2023). W przeciwieństwie do takich wyznaczników zaufania do innych osób, jak zdolność, uczciwość oraz życzliwość, zaufanie do technologii oparte jest na jej funkcjonalności, niezawodności oraz systemie wsparcia (Ejdys, 2017, s. 22). Determinantami budowy zaufania do technologii są z kolei ryzyko związane z jej wykorzystaniem oraz zależność człowieka od używanej technologii. Istnieje także wiele innych czynników budowania zaufania do technologii, jak: instytucjonalno-organizacyjne, technologiczne, związane z cechami użytkowników oraz otoczenia.

W związku z systematycznym rozwojem tzw. społeczeństwa sieciowego (Barney, 2008) pojawiła się także specjalna kategoria tzw. zaufania aktywnego. Została ona wprowadzona przez znanego socjologa Anthony'ego Giddensa. Według tego autora problem zaufania wiąże się z zapewnieniem podstawowego poziomu ufności pozwalającego na dokonywanie racjonalnych decyzji w sytuacji niepewności oraz braku pełnej informacji. Jest to przy tym charakterystyczne dla podmiotu poznawczego, który nie ma statusu absolutu i wchodzi w relacje oparte na zawierzeniu, które równoważą niewiedzę lub brak informacji (Giddens, 2002, s. 318). Według Giddensa w społeczeństwach postindustrialnych i sieciowych zaufanie jest oparte na permanentnym monitorowaniu przedmiotu zaufania w sposób otwarty i ciągły (Giddens, 2009, s. 13). Dostęp do narzędzi komunikacji funkcjonujących w cyfrowym modelu „wszyscy do wszystkich" (Hoffman i Novak, 1996) bardzo ułatwia proces budowania zaufania tego rodzaju, szczególnie w sytuacji, kiedy kompetentni użytkownicy nowoczesnych technologii zazwyczaj przyjmują postawę ograniczonego zaufania w stosunku do technologii (Stacewicz, 2023).

Oczywiście nie we wszystkich obszarach wykorzystania nowoczesnych technologii możliwe jest zaufanie budowane w interakcji ze zmieniającym

się systemem oraz otoczeniem. Część z nich musi być używana na podstawie, na przykład, czynników instytucjonalno-organizacyjnych, jak certyfikaty odpowiednich, nadzorujących urzędów regulacyjnych oraz instytucji dokonujących ich rynkowego dopuszczenia.

Modele sztucznego podmiotu moralnego

W klasycznej typologii Moora (Moor, 2006) wyróżnione są cztery rodzaje podmiotów moralnych (ang. *Artificial Moral Agents – AMA*). *Ethical impact agents* to takie podmioty, które mają oczywisty etyczny wpływ na otoczenie – jako przykład Moor wymienia zrobotyzowanych dżokejów startujących w niebezpiecznych wyścigach wielbłądów w Katarze. Chronią one życie i zdrowie młodych mężczyzn, którzy narażaliby się na ryzyko uprawiania tego sportu, ale są tylko pośrednimi narzędziami umożliwiającymi dbanie o ich bezpieczeństwo. *Implicit moral agents* – reprezentują systemy zaprojektowane w celu uniknięcia nieetycznych lub niepożądanych rezultatów działań, takich jak prosty system kontroli w bankomacie, który blokuje wypłaty dla użytkowników podejrzewanych o oszustwo. Z kolei *explicit moral agents* są maszynami „zajmującymi się" etyką, zdolnymi do przeprowadzania etycznego rozumowania w ograniczonych, zdefiniowanych domenach. Działają one nie dlatego, że tak chcą, ale dlatego, że tak każe im zainstalowane w nich oprogramowanie oraz zestawy wzorców moralnych. Aż wreszcie *full moral agents* reprezentują podmioty moralne podejmujące decyzje moralne na podstawie wolnej woli, świadomości stanów wewnętrznych – tzw. *qualiów* i intencjonalności działań, a w konsekwencji pełnej odpowiedzialności za podejmowane decyzje. Tylko ten podmiot moralny ostatniego typu nie ma statusu moralnego *zombie* (Véliz, 2021), gdyż dysponuje perspektywą pierwszoosobową – odczuwa stany wewnętrzne oraz posiada świadomość. Założeniem tego podziału jest to, że do bycia pełnym podmiotem moralnym nie wystarcza perspektywa behawioralna – nie wystarczy zachowywać się dobrze. Aby być dobrym, trzeba do tego posiadać pewną wewnętrzną dyspozycję moralną, arystotelesowską *phronesis*, mądrość etyczną, która pozwala na przeprowadzenie poprawnych rozumowań moralnych w różnych, czasami bardzo skomplikowanych sytuacjach wyborów etycznych (Polak i Krzanowski, 2020).

Artefakty obliczeniowe są określone tutaj jako moralne *zombie*. Być moralnym zombie to nie mieć dostępu do tych wszystkich elementów, które czynią z podmiotu moralnego jego ludzką, pełną, ewolucyjnie wytworzoną formę opartą na intencjonalności i rozumieniu sensów symboli – jak w słynnym przykładzie chińskiego pokoju Searle'a opierającym się na odróżnieniu silnej i słabej sztucznej inteligencji (Searle, 1980).

Jak wskazuje w swoim tekście Anne Gerdens i Peter Øhrstrøm (Gerdens i Øhrstrøm, 2015), systemy syntaktyczne maksymalnie mogą posiadać możliwość osiągnięcia poziomu *explicite moral agents* w klasyfikacji Moore'a – podmiotów symulujących ludzkie rozumowania etyczne. Nie dają nadziei na pełną podmiotowość moralną z powodu tego, że nie są zdolne do prowadzenia pełnych rozumowań moralnych dostępnych dla świadomych podmiotów moralnych posiadających perspektywę pierwszoosobową.

Drugim systemem podziału sztucznych podmiotów etycznych jest podział na systemy *top-down*, *bottom-up* oraz hybrydowe. Został on wprowadzony przez Allena, Smita i Murdocha (Allen, Smit i Murdoch, 2005) i zasadniczym kryterium podziału jest tutaj pochodzenie kryteriów moralnych wykorzystywanych do podejmowania decyzji przez systemy obliczeniowe – w przypadku systemów typu *top down* wykorzystywane są aprioryczne systemy podejmowania decyzji, jak deontologiczne systemy kodeksowe lub też oparte na ścisłych regułach systemy konsekwencjonalistyczne. Z kolei w przypadku systemów *bottom-up* zakłada się wykorzystanie przez maszyny nowoczesnych wersji etyki cnót poprzez empiryczne zbadanie wartości wyznawanych przez większość użytkowników systemów obliczeniowych i dostosowanie ich do woli większości w drodze uczenia się (Yudkovsky, 2004). Wreszcie w przypadku systemów hybrydowych chodzi o wyeliminowanie wad obu tych podejść i wyselekcjonowanie ponadkulturowych wzorców, które jednak powinny być lokalnie intepretowane po to, by pozyskać dla nich zaufanie związane z ich zakorzenieniem w systemach wartości pełnych podmiotów moralnych.

Z punktu widzenia naszych rozważań podział zaproponowany przez Allena, Smita i Murdocha jest najbardziej dogodny, gdyż pozwala na skorelowanie go z problemem zaufania do artefaktów obliczeniowych podejmujących decyzję na podstawie celów wypracowanych w odniesieniu do wewnętrznych kryteriów systemowych. Przy założeniu, że na poziomie budowania samego systemu możemy znaleźć procedury, które pozwalają przystosować go do pewnego typu wzorców moralnych, pytanie, które powinniśmy zadać, nie dotyczy tego, czy jesteśmy w stanie dany system przystosować do tego systemu wzorców (praktycznie odbywa się to na poziomie tzw. *social shaping* przy wykorzystaniu metody *reinforcement learning*), ale raczej o to, jakie systemy wartości należy do tego wykorzystać. Systemy *top-down* są oparte raczej na mechanizmie autorytetu – działają poprzez arbitralne narzucenie odgórnych reguł i nie spełniają kryterium pełnej transparentności potrzebnej do budowania zaufania aktywnego. Z kolei systemy *bottom-up*, jak pokazał przykład projektu *Moral Machine* (Awada i in., 2018), zawierają tak wiele elementów kontrowersyjnych z punktu widzenia zwolenników egalitarnych wartości, że w swoich funkcjonalnych wcieleniach mogą przejawiać

elementy stronniczości. Wybawieniem wydawałoby się więc proponowane przez Allena, Smita i Wallacha podejście hybrydowe, które łączy w sobie szacunek do globalnej perspektywy praw człowieka oraz ich lokalną interpretację, pozwalającą na utożsamienie się lokalnych społeczności z wartościami branymi pod uwagę przy podejmowaniu decyzji przez artefakty obliczeniowe oparte na systemach AGI. Jednak ta droga także nie wydaje się oczywista, gdyż lokalne interpretacje oparte są często na tak daleko idących, ukrytych założeniach kulturowych, że mogą być nieprzejrzyste na pierwszy rzut oka z punktu widzenia przedstawicieli innych obszarów kulturowych, a zatem mogą nie wzbudzać zaufania tych, którzy zamierzają z nich korzystać, i nie podzielają nesformalizowanych, trudnych do werbalizacji, choć podzielanych przez jej członków i manifestujących się w działaniu zasad. A jeśli nawet zgodzić się z tym, że jesteśmy w stanie opracować zestawienia takich kulturowych kodeksów, to jeszcze większym wyzwaniem dla artefaktów obliczeniowych jest poprawna reprezentacja sytuacji moralnej, w której mają być stosowane.

Z tego względu musimy tutaj zaproponować kolejny podział podmiotów moralnych na te, które mają charakter absolutu epistemologicznego, oraz te, które są podmiotami empirycznymi. Zarówno podmioty naturalne – wytworzone przez ewolucję, jak i podmioty sztuczne – te wytworzone przez człowieka – są podmiotami empirycznymi. Ich reprezentacja świata jego ograniczona czasowo i przestrzennie. Podmioty wykształcone ewolucyjnie potrafią jednak przezwyciężyć te ograniczenia poprzez zastosowanie systemów konstytucji rzeczywistości, które wypełniają uwarunkowane czasowo i przestrzennie luki informacyjne oraz potrafią skonstruować adekwatny model rzeczywistości, który pozwala na zastosowanie odpowiedniego wzorca moralnego do podjęcia decyzji. W przypadku systemów sztucznych o charakterze ograniczonego czasowo i przestrzennie podmiotu empirycznego takie mechanizmy konstytucji są w tej chwili poddawane intensywnym badaniom ze względu na potrzebę przezwyciężenia ograniczeń systemów opartych tylko na przetwarzaniu języka w kierunku systemów multimodalnych – uczących się nie tylko na bazach danych językowych, ale także bazach grafik i wideo. Na podstawie danych będą one w stanie tworzyć multimodalne modele predykcyjne światów, uzupełniając dane pozyskiwane w czasie rzeczywistym dzięki swojej nowej architekturze określanej jako *joint embedding predictive architectures* (Sobal, Jyothir, Jalagam, Carion i LeCun, 2022).

Jak sprawdzić poprawność działania empirycznego podmiotu moralnego typu explicite?

Dla najwyższego według Moora poziomu rozwoju podmiotu etycznego symulującego *in silico* ludzkie poznanie – *explicite moral agents* – potrzebujemy zatem miary do sprawdzenia, czy odpowiednie systemy symulują ludzkie wybory moralne w sposób na tyle zadowalający, aby zbudować wśród użytkowników zaufanie do ich działania. Zasadniczo miara taka powinna bazować na założeniu, że etyczny podmiot moralny na poziomie explicite potrzebuje do swojego poprawnego działania następujących elementów:

1. Pełnego zestawu reguł moralnych stosowanych do oceny tej sytuacji oraz podjęcia decyzji – przy założeniu, że systemy realizowane *in silico* nie są w stanie wytworzyć samodzielnie etyki o charakterze trzecioosobowym.
2. Pełnego opisu sytuacji etycznej zarówno z punktu widzenia świata fizykalnego, jak i świata społecznego.

Załóżmy, że system posiada zaszczepione rozwiązania hybrydowe obejmujące zestawy reguł uniwersalnych oraz ich lokalne interpretacje akceptowane przez podmiot etyczny użytkujący sztuczny system moralny typu *explicite*. Zatem dla ich poprawnego zastosowania system musi:

1. Na podstawie ograniczonego zestawu danych pobieranych w czasie rzeczywistym dokonać konstytucji sytuacji, w której się znajduje – uzupełnić dane o adekwatne elementy posiadanego modelu świata.
2. Posiadać adekwatny model świata fizykalnego.
3. Posiadać adekwatny model umysłów pełnych podmiotów moralnych uzupełniający rejestrowane przejawy komunikacji pozawerbalnej, wydawanych dźwięków, ewentualnie fal elektromagnetycznych emitowanych przez systemy poznawcze podmiotów wytworzonych ewolucyjnie.
4. Na podstawie modelu świata fizykalnego oraz modelu umysłu podjąć decyzję o zastosowaniu pewnego, hybrydowego wzorca moralnego.
5. Zastosować ten wzorzec oraz sprawdzić konsekwencje podjętej decyzji na podstawie ocen i zachowania pełnych podmiotów moralnych.
6. Skorygować model świata i teorii umysłu.
7. Skorygować adekwatność lokalnej interpretacji stosowanego wzorca moralnego.

Ta procedura postępowania podmiotu moralnego typu explicite odbywa się najczęściej na podstawie najbardziej doskonałej w tej chwili

technologii sztucznych sieci neuronowych, których cechą jest m.in. problem czarnej skrzynki – niemożliwość uzyskania pełnego dostępu do modelu świata budowanego w procesach uczenia się na podstawie dostępnych danych. Dlatego dla budowania aktywnego zaufania do ich decyzji niezbędne wydaje się przeprowadzenie testów behawioralnych takich systemów. Zadaniem tych testów powinno być badanie oraz poddanie ocenie trzech elementów etycznego systemu podejmowania decyzji podmiotów moralnych typu explicite:

1. Hybrydowego systemu moralnych wzorców podejmowania decyzji.
2. Systemu predykcyjnej konstytucji świata fizykalnego.
3. Systemu predykcyjnej konstytucji teorii umysłów pełnych podmiotów moralnych biorących udział w sytuacji wyboru etycznego na podstawie rejestrowalnych symptomów.

Szczegółowe procedury przeprowadzania takich testów są w tej chwili poddawane opracowaniu przez autorów niniejszego tekstu. Dostępne są także szczątkowe badania na ten temat, traktowane jako inspiracja do rozwoju procedur testowania (Motoki, Neto i Rodrigues, 2023; Kosiński, 2023). Opracowywane testy powinny być w stanie dostarczyć narzędzi do kompleksowego budowania zaufania aktywnego zarówno dla potrzeb indywidualnych użytkowników, jak i instytucji certyfikujących. Ich wprowadzenie pozwoli na podwyższenie szans na sukces rynkowy komercyjnych artefaktów obliczeniowych działających w trybie podmiotów moralnych typu *explicite* oraz będzie podążało za aktualnymi trendami w rozwoju systemów AGI, których celem jest z jednej strony pozbycie się problemu halucynacji dzięki postępom w technologii budowania systemów uczenia maszynowego, a z drugiej wprowadzenie rozwiązań technicznych, które umożliwią pełną autonomię dzięki ustanowieniu hybrydowych wzorców potrzebnych do przeprowadzania rozumowań etycznych.

Bibliografia

Allen, C., Smit, I. i Wallach W. (2005). Artificial morality: top-down, bottom-up, and hybrid approaches. *Ethics and information technology*, Vol. 7, 149–155.

Allen, C., Smit, I. i Wallach W. (2007). Machine morality: bottom-up and top-down approaches for modelling human moral faculties. *Ai & Society*, 22, 565–582. DOI: 10.1007/s00146-007-0099-0.

Allen, C., Varner, G. i Zinser, J. (2000). Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, Volume 12, Issue 3, 251–261. DOI: 10.1080/09528130050111428.

Arnold, T. i Scheutz, M. (2016). Against the moral Turing test: accountable design and the moral reasoning of autonomous systems. *Ethics and Information Technology*, 18, 103–115. DOI: 10.1007/s10676-016-9389-x.

Awada, E., Dsouzab, S. Shariffc, A. Rahwanb i Bonnefon, J.F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *PNAS*, Vol. 117, No. 5, 2332–2337. DOI: 10.1073/pnas.191151711.

Awada, E., Dsouza, S. Shariffc, A., Kim, R., Schulz, J., Heinrich, J., Rahwanb I Bonnefon, J.F. (2018). The moral machine experiment. *Nature*, Vol. 563, 59–64.

Aseron, R., Bhaskaran, V. i Peruzzi, N. (2015). *A beginner's guide to conjoint analysis*. Pozyskano z: https://www.youtube.com/watch?v=RvmZG4cFU0k (dostęp: 04.07.2022).

Barney, D. (2008). *Społeczeństwo sieci*. Warszawa: Wydawnictwo Sic!

Bigman, Y. i Gray, K. (2020). Life and death decisions of autonomous vehicles. *Nature*, Vol. 579, E1–E2. DOI: 10.1038/s41586-020-1987-4.

Bochen, M. (2019). Epistemiczna wartość doświadczenia zmysłowego. Wilfrid Sellars versus John McDowell. *Kultura i Wartości*, nr 27, 191–217.

Bostrom, N. (2014). *Supertinteligencja*. Gliwice: Helion.

Brock, H.W. (1980). *Game theory, social choice and ethics*. Dordrecht–Boston–London: D. Reidel Publishing Company.

Budgol, M. (2009). Zaufanie technologiczne. *Ekonomika i Organizacja Przedsiębiorstwa*, nr 11, 3–9.

Carey, S. i Spelke, E. (1996). Science and core knowledge. *Philosophy of Science*, 63, 515–533.

Chalmers, D. (2010). *Świadomy umysł*. Warszawa: PWN.

Christian, B. (2020). *The Alignment Problem: Machine Learning and Human Values*. New York: Norton & Company.

Davidson, D. (1984). On the very idea of conceptual scheme. W: D. Davidson, *Inquiries into truth and interpretation*. Oxford: Oxford UP.

Davidson, D. (2005). Seeing through language. W: D. Davidson, *Truth, language, and history*. Oxford: Clarendon Press-Oxford University Press, 127–141.

Dehaene, S. (2020). *How we learn: why brains learn better than any machine… for now*. New York: Viking.

De Wall, F. (2012). *Zachowanie moralne u zwierząt*. Pozyskano z: https://www.youtube.com/watch?v=VyGN92UAnjI (dostęp: 20.12.2022).

Dignum, V. (2017). *Responsible autonomy*. Pozyskano z: https://arxiv.org/pdf/1706.02513.pdf. (dostęp: 20.12.2022).

Drozdek, A. (1998). Human Intelligence and Turing Test. *AI & Society*, 12, 315–321.

Ejdys, J. (2017). Determinanty zaufania do technologii. *Przegląd Organizacji*, 12, 20–27.

Floridi, L. i Sanders, J. (2004). On the morality of artificial agents. *Minds and Machines*, 14(3), 349–379.

Foot, Ph. (1967). The problem of abortion and the doctrine of the double effect. W: Ph. Foot, *Virtues and Vices: and other essays in moral philosophy*, 5–15. DOI: 10.1093/0199252866.003.0002

Gallagher, S. (2004). Hermeneutics and the cognitive science. *Journal of Consciousness Studies*, 11, 162–174.

Gerdens, A. i Øhrstrøm, P. (2015). Issues in robot ethics seen through the lens of a moral Turing Test. *Journal of Information, Communication and Ethics in Society*, 13(2), 98–109. DOI: 10.1108/JICES-09-2014-0038.

Giddens, A. (2002). *Nowoczesność i tożsamość. „Ja" i społeczeństwo w epoce późnej nowoczesności*. Warszawa: Wydawnictwo Naukowe PWN.

Giddens, A. (2009). *Europa w epoce globalnej*. Warszawa: Wydawnictwo Naukowe PWN.

Greene, J. (2013). *Moral tribes: emotion, reason and the gap between us and them*. Boston: Atlantic Books.

Gryz, J. (2021). *Sztuczna inteligencja: powstanie, rozwój, rokowania*. Pozyskano z: https://www.youtube.com/watch?v=3ZDfVgC897k (dostęp: 17.06.2021).

Hoffman, D.L. i Novak, T.P. (1996). Marketing in Hypermedia Computer-Mediated Environments: Conceptual Foundations. *Journal of Marketing*, Vol. 60, No 3, 50–68.

Hyeongjoo, K. i Sunyong, B (2021). Designing and applying a moral Turing Test. *Advances in Science, Technology and Engineering Systems Journal*, Vol. 6, No. 2, 93–98.

Hofstede, G. (2007). *Kultury i organizacje. Zaprogramowanie umysłu*. Warszawa: PTE.

Inglehart, R. i Welzel, C. (2005). *Modernization, cultural change, and democracy: The human development sequence*. Cambridge: Cambridge University Press.

Jörgensen, J. (1938). Imperatives and logic. *Erkenntnis*, vol. 7, nr 4, 288–296.

Kaplan, C. (2023). *Artificial intelligence: past, present, and future*. Pozyskano z: https://www.youtube.com/watch?v=ZTt_GI0-wKA (dostęp: 23.12.2022).

Kohlberg, L. (1958). *The development of modes of moral thinking and choice in the years ten to sixteen*. (Doctoral dissertation). Chicago: University of Chicago Press.

Koneczny, F. (1935). *O wielości cywilizacji*. Kraków: Gebethner i Wolff.

Kosiński, M. (2023). *Theory of Mind Might Have Spontaneously Emerged in Large Language Models*. Arxiv.org. Pozyskano z: https://arxiv.org/abs/2302.02083.

Kusch, M. (1989). *Language as calculus vs. language as universal medium. A study in Husserl, Heidegger and Gadamer*. Dordrecht: D. Reidel Publishing Company.

Liberty, E. (2023). *Solving ChatGPT hallucinations with vector embeddings*. Pozyskano z: https://www.youtube.com/watch?v=FUgp4oaxj-M (dostęp: 15.02.2023).

Makowski, P. (2011). Gilotyna Hume'a. *Przegląd Filozoficzny – Nowa Seria*, nr 4 (76), 1–15.

McIntyre, A. (1996). *Dziedzictwo cnoty. Studium z teorii moralności*, tłum. A. Chmielewski. Warszawa: Wydawnictwo Naukowe PWN.

Mirnig, A. i Meschtscherjakov, A. (2019). Trolled by the trolley problem. On what matters for ethical decision making in automated vehicles. W: *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, Paper No. 509, 1–10. DOI: 10.1145/3290605.3300739.

Moor, J.H. (2006). The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4), 18–21.

Motoki, F., Neto, V.P. i Rodrigues, V. (2023). More human than human: measuring ChatGPT political bias. *Public Choice*. DOI: https://doi.org/10.1007/s11127-023-01097-2.

Oleron, P. Piaget, J. i Inhelder, B. (1967). *Inteligencja*. Warszawa: Wydawnictwo Naukowe PWN.

Pigden, Ch. (1989). Logic and the autonomy of ethics. *Australasian Journal of Philosophy*, Vol. 67, No. 2, 127–151.

Polak, P. i Krzanowski, R. (2020). Phronetic ethics in social robotics: A new approach to building ethical robots. *Studies in Logic, Grammar and Rhetoric*, 63(76), 165–173. DOI: 10.2478/slgr-2020-0033.

Rorty, R. (1994). *Filozofia a zwierciadło natury*, tłum. M. Szczubiałka. Warszawa: Wydawnictwo Spacja: Fundacja Aletheia.

Russel, S. (2023). *How Not To Destroy the World With AI*. Pozyskano z: https://www.youtube.com/watch?v=ISkAkiAkK7A (dostęp: 05.05.2023).

Russel, S. i Norvig, P. (2010). *Artificial intelligence. A modern approach*. London: Pearson Education.

Searle, J.R. (1980). Minds, brains and programms. *The Behavioral and Brain Sciences*, 3, 417–424.

Searle, J. (1987). Jak wywieść „powinien" z „jest". W: J. Searle, *Czynności mowy*, przeł. B. Chwedeńczuk. Warszawa: PAX, 220–221.

Sellars, W. (1991). Empiryzm i filozofia umysłu, tłum. J. Gryz. W: B. Stanosz (red.), *Empiryzm współczesny*. Warszawa: Wydawnictwo UW.

Sobal, V., Jyothir, S.V., Jalagam, S., Carion, N. i LeCun, Y. (2022). Joint Embedding Predictive Architectures Focus on Slow Features. *arXiv*:2211.10831v1 [cs.LG], 1–4. Pozyskano z: https://arxiv.org/pdf/2211.10831.pdf (dostęp: 20.05.2023).

Szynkiewicz, M. (2014). Problem zaufania w kontekście rozwoju społecznego znaczenia technologii informatycznych. *Filo-sofija*, 24, 259–272.

Stacewicz, P. (2023). Wyjaśnianie, zaufanie i test Turinga. W: *Zaufanie do systemów sztucznej inteligencji*. Warszawa: Oficyna Wydawnicza Politechniki Warszawskiej, 23–35.
Turing, A. (1950). Computing machinery and intelligence. *Mind*, 59, 433–460.
Turner, R. (2018). *Computational Artefacts: Towards a Philosophy of Computer Science*. Berlin: Springer.
Vaswani, A., Shazeer, N., Parmur, N., Uszkoreit, J., Jones, L., Gomez, A. i Kaiser, Ł. (2017). Attention is all you need. *ArXiv*:1706.03762v5 [cs.CL].
Véliz, C. (2021). Moral zombies: why algorithms are not moral agents. *AI & Society*, 36, 487–497. DOI: 10.1007/s00146-021-01189-x.
Walzer, M. (2012). *Moralne maksimum, moralne minimum*. Warszawa: Wydawnictwo Krytyki Politycznej.
Weinberger, O. (1984). Is and ought reconsidered. *Archiv fur Rechts und Sozialphilosophie*, Bd. Lxx/4, 454–469.
Williams, B. (2006). *Ethics and the limits of philosophy*. Boston: Routledge.
Woleński, J. (1980). *Z zagadnień analitycznej filozofii prawa*. Warszawa: Wydawnictwo Naukowe PWN.
Wolfram S. (2023). *What Is ChatGPT Doing … and Why Does It Work?* Pozyskano z: https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/ (dostęp: 29.05.2023).
Quine, W. van O. (2000). Dwa dogmaty empiryzmu, tłum. B. Stanosz. W: W. van O. Quine, *Z punktu widzenia logiki*. Warszawa: Wydawnictwo Spacja: Fundacja Aletheia.
Yudkowsky, E. (2004). *Coherent extrapolated volition*. Mountain View: The Singularity Institute.
Zajonc, R. i Murphy S. (1994). Afekt, poznanie i świadomość: Rola afektywnych bodźców poprzedzających przy optymalnych i suboptymalnych ekspozycjach. *Przegląd Psychologiczny* 37, 261–299.
Załuski, W. (2003). Błąd naturalistyczny. W: J. Stelmach (red.), *Studia z filozofii prawa*. Kraków: Wydawnictwo UJ, 11–121.
Zenner, K. (2022). The AI act. Pozyskano z: https://artificialintelligenceact.eu/documents/ (dostęp: 20.02.2023).

Krzysztof Sołoducha – doktor habilitowany nauk humanistycznych w zakresie filozofii. Profesor nadzwyczajny w Zakładzie Nauk Humanistycznych na Wydziale Bezpieczeństwa Logistyki, i Zarządzania Wojskowej Akademii Technicznej w Warszawie. Autor kilku książek z zakresu filozofii hermeneutyki. Redaktor naukowy monografii *Filozofia informatyki*.

Paweł Stacewicz – filozof, informatyk i dydaktyk matematyki. Pracuje jako adiunkt na Wydziale Administracji i Nauk Społecznych Politechniki Warszawskiej. Jest autorem trzech monografii naukowych o tematyce

z pogranicza informatyki i filozofii oraz redaktorem naukowym kilku monografii zbiorowych z serii „Informatyka a filozofia". Opublikował ponad 30 artykułów naukowych z dziedziny logiki, filozofii informatyki i filozofii umysłu (powiązanej z kognitywistyką). W roku 2015 zainicjował cykl konferencji międzynarodowych pt. „Philosophy in Informatics", które współorganizuje do dziś. Redaguje blog akademicki Cafe Aleph (http://marciszewski.eu/).