ARGUMENT
BIANNUAL PHILOSOPHICAL JOURNAL
e-ISSN 2084–1043 p-ISSN 2083–6635
Published online: 01.09.2023
Vol. 13 (1/2023) pp. 109–122
www.argument-journal.eu

Wyzwania związane z tworzeniem etyki typu *data driven* dla maszyn typu *artificial general inteligence* (AGI)

Krzysztof SOŁODUCHA*

ABSTRACT

Challenges in developing data-driven ethics for artificial general intelligence machines (AGI): One of the biggest challenges in creating systems of AGI (artificial general intelligence) type is the question of the method for developing decision-making patterns for such machines operating in autonomous mode. The need to build active trust in their functioning forces an ART (accountability, responsibility, transparency) approach, which prefers bottom–up methods in ethics. On the other hand, such descriptive methods expose culturally conditioned, non-egalitarian patterns of behaviour that may be morally objectionable and rather indicate the need to use a top–down approach that applies normative ethical expectations. The paper examines the challenges of using both methods and outlines the need to develop a hybrid approach. The correct functioning of machines using such a hybrid system should be examinated by using an ethical testing process — at the level of output signal. At the end of the paper we present some philosophical assumptions that should be taken into account in the construction of such an ethical test.

KEYWORDS

general artificial intelligence ethics; active trust in artificial intelligence; model for building artificial intelligence ethics; hybrid approach to artificial intelligence ethics; ethical testing process

* Dr hab., Zakład Nauk Humanistycznych na Wydziale Logistyki, Bezpieczeństwa i Zarządzania Wojskowej Akademii Technicznej w Warszawie. E-mail: krzysztof.soloducha@wat.edu.pl.

DOI: 10.24917/20841043.13.1.07

Znana z epistemologii zagadka mitu danych w dziedzinie etyki jest określana jako zagadnienie powinności, a dokładnie przejścia od „jest" do „powinno być" (*is — ought problem*). Za prekursora tej tematyki uważa się Davida Hume'a. Źródłem sformułowanego przez niego problemu jest zagadnienie „pomieszania" w metaetyce poziomu empirycznego (faktów) i normatywnego (powinności) oraz kwestia możliwości przekroczenia tej granicy. Nie jest naszym zadaniem relacjonowanie długiej i skomplikowanej dyskusji na ten temat — można ją znaleźć w licznych opracowaniach (Załuski, 2003; Brożek, 2001; Makowski, 2011; Searle, 1987; Pigden, 1989; Woleński, 1980). Z punktu widzenia celu naszych rozważań ważne są jednak konsekwencje istnienia problemu gilotyny Hume'a oraz błędu naturalistycznego (Załuski, 2003).

Problematyka relacji języka i doświadczenia w metaetyce jest bowiem uwikłana w proces szukania odpowiedzi na ważne pytanie, jakie stanowi kwestia przyjęcia stosowanego modelu jej budowania. Występują tutaj co najmniej trzy propozycje: *top–down* (z góry do dołu), *bottom–up* (z dołu do góry) i mieszana (*hybrid*) (de Wall, 2012; Allen, Smit, & Wallach, 2005). Klasyczne systemy etyczne, jak deontologiczny czy też konsekwencjonalistyczny, są uznawane za systemy typu *top–down* — takie, w których wydawanie sądów moralnych jest uzależnione od stosowania pewnej abstrakcyjnej, spójnej wewnętrznie reguły etycznej, z której wywodzi się kryteria oceny zachowań podmiotów etycznych. Kryteria te tworzone są na gruncie rozumowania spełniającego warunki wewnętrznej poprawności z punktu widzenia logicznego. Z kolei za system etyczny tworzony w trybie od „dołu do góry" uważa się rozwiązania, których podstawą jest procedura zbierania informacji na temat akceptowanych społecznie lub opłacalnych ewolucyjnie wzorców zachowań oraz uogólnienie ich w zestawy norm, które mogą służyć do oceny decyzji podejmowanych przez podmioty etyczne, choć wedle niektórych interpretacji tak rozumianej etyki cnót nie da się uprawiać bez przyjęcia założeń antropologicznych, jak na przykład racjonalność oraz dążenie do eudajmonii (Allen, Smit, & Wallach, 2005).

Z punktu widzenia klasycznego podejścia do zagadnienia powinności, jedyną poprawną formą etyki jest etyka tworzona w trybie *top–down*, gdyż wszelkie odwołanie się do sfery empirycznej jest — podobnie jak w przypadku rozważań epistemologicznych — obarczone problemem relacji pomiędzy sferą językową a sferą doświadczenia oraz kwestią błędu naturalistycznego — pomieszania sfery wolności z dziedziną, w której obowiązuje deterministyczna przyczynowość. W ramach tzw. silnej racjonalności jedynym sposobem budowania poprawnych rozumowań jest bowiem odwołanie się do związków logicznych pomiędzy wyrażeniami, a zgodnie z regułami logiki przesłanki poprawnych rozumowań nie mogą mieć empirycznego charakteru. Równocześnie dyskusja wokół metody budowania etyki w modelu „góra–dół" odwołuje się do tezy, że zdania o normach nie są zdaniami w sensie logicznym, gdyż nie można im przypisać prawdy lub fałszu — a więc z tego punktu widzenia

etyka jest dziedziną wątpliwą z epistemologicznego punktu widzenia, zaś sądy o normach nie mają żadnej wartości poznawczej — w skrajnej interpretacji są tylko wyrazami emocji.

Rozważania na temat niemożliwości wydawania poprawnych logicznie sądów powinnościowych natrafiają jednak na prosty fakt graniczny — takie sądy powinnościowe są wydawane. Z tym zagadnieniem zmaga się tzw. paradoks Jörgensena, wymuszający w przypadku metaetyki pewną rewizję zasady Donalda Davidsona, mówiącej, że system wiedzy to sieć nawzajem wspierających się, powiązanych logicznie przekonań. To pojęcie wiedzy jest odróżniane od tradycyjnego spojrzenia na wiedzę metaforycznie określaną jako budowla wznosząca się na fundamencie bazowych danych empirycznych — tak jak w klasycznym modelu Kartezjańskim odwołującym się do metafory drzewa (Davidson, 1984).

PARADOKS JÖRGENA JÖRGENSENA I ZADANIE ETYKI MASZYN AUTONOMICZNYCH

Rozumowanie o charakterze etycznym nazywane paradoksem Jörgensena odwołuje się do zarysowanej powyżej filozoficznej dychotomii między formalną poprawnością a praktyczną realnością faktu wydawania sądów moralnych oraz działania, które odwołuje się do tych sądów.

Został on sformułowany w artykule pt. *Imperatives and logic* opublikowanym w 1938 roku w czasopiśmie *Erkenntnis* (Jörgensen, 1938). Wedle tego rozumowania tylko zdania w sensie logicznym (tj. prawdziwe lub fałszywe) mogą być składnikami poprawnych logicznie rozumowań, normy natomiast nie mogą być nimi, ponieważ nie są zdaniami w sensie logicznym. Ale rozumowania normatywne istnieją. Jak więc się one dokonują?

Niektóre próby rozwiązania tego dylematu odwołują się do zestawienia ze sobą reguł tzw. twardej, formalnej racjonalności, z punktu widzenia której problem statusu norm jest nierozstrzygalny, oraz tzw. miękkiej racjonalności, odwołującej się do społecznej praktyki, a więc szukającej argumentów w sferze danych. Jest to proces, który przypomina ewolucję myśli Ludwiga Wittgensteina pomiędzy *Traktatem logiczno-filozoficznym* a *Dociekaniami filozoficznymi*. Także Wittgenstein w swoim rozwoju teoretycznym przeszedł drogę pomiędzy definiowaniem systemu językowego jako zespołu spójnych reguł, dla których liczy się tylko formalna poprawność ich stosowania, do charakteryzowania go jako gry, której podstawy leżą w funkcji komunikacyjnej oraz społecznej języka.

W związku z tym problemem powstały koncepcje budowania systemów etycznych w oparciu o tzw. miękką racjonalność — klasycznym przykładem są próby wykorzystania w tym celu teorii gier i pojęcia równowagi Johna Nasha. Ich podstawą jest jednak tzw. teoria użyteczności, która z kolei odwołuje się

do utylitaryzmu i utylitarystycznej teorii sprawiedliwości i z tego powodu nie wyłamują się one z modelu *top–down*.

Inną próbą zastosowania miękkiej racjonalności nawiązującej bezpośrednio do tzw. drugiego Wittgensteina są prace Johna Searle'a (Searle, 1987) dotyczące rozwiązania problemu powinności dzięki wykorzystaniu argumentu, wedle którego powiązanie bytu i powinności powinno dokonywać się w oparciu o inny związek niż relacja logiczna. Inicjatorem takiego podejścia jest Alasdair MacIntyre (MacIntyre, 1996), który uznał standardową interpretację gilotyny Hume'a za błędną, gdyż niebiorącą pod uwagę badań nad indukcją prowadzonych przez szkockiego filozofa. Wedle Searle'a rozwiązania problemu należy dokonać poprzez rezygnację z wymogów wynikania logicznego i wprowadzenie pojęcia faktu instytucjonalnego, który przez swoją regularność nadaje językowi moce deontyczne (regularność odróżnioną od reguły). Z kolei moc tej regularności można sprawdzić na przykład przez analizę statystyczną profili językowych. Odwołanie się do statystyki ma jednak konsekwencje w postaci wykorzystania reguł rozumowania zawodnego, takiego jak indukcja lub abdukcja (Makowski, 2011: 12). Stanowisko to nazywane jest nonkognitywizmem (w opozycji do kognitywizmu, czyli przekonania, że jedynym wzorcem poprawności jest mocna racjonalność odwołująca się do dedukcji).

Rozumowanie to odwołuje się do argumentu, że zachowania etyczne oraz sądy normatywne są ważnym elementem ładu społecznego i trudno wyobrazić sobie, żeby społeczeństwo działało poprawnie bez ich wykorzystania. Tym bardziej, że stanowią one także podstawę systemów prawnych. Co więcej, zachowania oraz oceny etyczne decydują o budowaniu czegoś, co określane jest jako kapitał społeczny — czyli wzajemnego zaufania związanego z oczekiwaniem od innych podmiotów etycznych zachowań dostosowanych do ogólnie przyjętych reguł współżycia zbiorowego.

Ten wątek pojawia się między innymi w rozważaniach Anthony'ego Giddensa oraz Francisa Fukuyamy. Według tego pierwszego problem zaufania społecznego wiąże się z zapewnieniem ufności pozwalającej na dokonywanie poprawnych decyzji w sytuacji niepewności, która jest naturalna dla podmiotu poznawczego niedysponującego możliwościami uzyskania poziomu absolutu epistemologicznego. Ufność polega „na zawierzeniu, które równoważy niewiedzę lub brak informacji" (Giddens, 2002: 318). Wedle Giddensa w społeczeństwach postindustrialnych i sieciowych mamy do czynienia z tzw. zaufaniem aktywnym, które jest „oparte na monitorowaniu uczciwości drugiej osoby w sposób otwarty i ciągły" (Giddens, 2009: 13). Podobne podejście, wzbogacone o wątek ekonomiczny, reprezentuje Fukuyama, który postrzega zaufanie jako epifenomen kapitału społecznego i „mechanizm oparty na założeniu, że innych członków danej społeczności cechuje uczciwe i kooperatywne zachowanie oparte na wyznawanych normach" (Fukuyama, 1997: 38).

KONCEPCJA ZOBIEKTYWIZOWANEJ WOLI LUDZKOŚCI JAKO NARZĘDZIE „MIĘKKIEJ RACJONALNOŚCI"

Zaufanie jako wartość społeczna zbudowane jest zatem na systemie wzajemnych oczekiwań opartych o zachowania ocenne. Dotyczy to także sfery normatywnej będącej podstawą podejmowania decyzji przez maszyny autonomiczne, których funkcjonowanie oparte jest o uczenie bez nadzoru (Kaplan, 2023) i które wymagają systemu kryteriów pozwalających na odróżnienie przez nie wyników rozumowań statystycznych akceptowalnych etycznie od tych, które z tego punktu widzenia muszą zostać odrzucone jako niezgodne ze społecznie uznanymi wzorcami (Gryz, 2021). Mogą się one zatem cieszyć zaufaniem i dzięki temu być używane tylko wtedy, kiedy spełniają te normatywne oczekiwania i jako maszyny etyczne dołączą do ludzkiego świata życia codziennego.

Z punktu widzenia wspominanego powyżej procesu budowania aktywnego zaufania potrzebnego do codziennego posługiwania się maszynami autonomicznymi ważny jest argument przytaczany przez Nicka Bostroma w jego książce *Superinteligencja* (Bostrom, 2014). Twierdzi on, że działające w trybie statystycznym maszyny trenowane bez nadzoru na podstawie dostępnych zbiorów danych nie mogą samodzielnie wytworzyć moralności postkonwencjonalnej zgodnej z wymaganiami teorii Kohlberga (Kohlberg, 1958), gdyż jest ona zbyt lokalna i nie jest możliwe zbudowanie na jej podstawie zaufania, którego podstawą musi być perspektywa trzecioosobowa, a więc jej zasady muszą działać w sposób symetryczny w sytuacji zmiany ról społecznych przez podmiot moralny (Bostrom, 2014: 306). Osiągnięcie trzeciego poziomu Kohlberga jest więc niezbędne do tego, żeby takie statystyczne maszyny mogły być uznawane za zasadniczo godne zaufania, czyli odpowiadające ludzkiemu wzorcowi inteligencji oraz moralności.

Skoro jednak, zgodnie ze wspomnianą wyżej koncepcją zaufania aktywnego, postkonwencjonalna moralność w społeczeństwach postindustrialnych nie może mieć charakteru uzurpacji uniwersalistycznej, wykonywanej w modelu *top–down*, musi ona być skuteczna perswazyjnie zgodnie z podejściem typu ART (*accountability*, *responsability*, *transparency*) (Dignum, 2017), czyli jej działanie musi być zrozumiałe dla podmiotów wchodzących w interakcję z maszynami. Zgodnie z tym rozumowaniem problem budzących zaufanie maszyn autonomicznych rodzi w sposób naturalny pytanie, jak zbudować model postkonwencjonalnej moralności perswazyjnej, która będzie spełniała kryteria zaufania aktywnego zgodnego z podejściem typu ART i działała zgodnie z systemem oczekiwań, który jest uznawany za podstawę ładu społecznego w danym społeczeństwie.

Jednym ze sposobów wybrnięcia z tego paradoksu jest pomysł Eliezera Yudkovsky'ego nazwanego koherentną, ekstrapolowaną wolą ludzkości CEV (*coherent extrapolated volition*) (Yudkowsky, 2004). W intencji tego autora

CEV jest realizacją odwiecznego marzenia o zbudowaniu etyki opisowej, która pokazałaby drogę od faktów do norm, wykorzystując reguły tzw. miękkiej racjonalności. Takim narzędziem miałaby być statystyczna ekstrapolacja, ale realizowana w skali ludzkości. Koncept CEV jest zbudowany na bazie koncepcji życzliwej (*benovelent*) sztucznej inteligencji odwołującej się do pomysłów Asimova (Asimov, 2004) i jego rozważań o etyce robotów. Do tego pozwala w wiarygodny sposób zbudować bazę wzorców, która jest wygodnym narzędziem wspomagającym funkcjonowanie maszyn autonomicznych.

W rozumieniu Yudkowsky'ego wykorzystanie reguł „miękkiej racjonalności" polega więc na zaoferowaniu maksymalnej reprezentatywności zbieranych danych o preferencjach etycznych w jedynym, realnie dostępnym zakresie pełnego zbioru podmiotów moralnych, jakim jest zbiór całej ludzkości. Spełnienie formalnych, ilościowych warunków statystycznych wyłonienia próbki reprezentatywnej ma zapewnić przy tym oczekiwane przejście pomiędzy „jest" a „powinno być". Pojawia się więc w ten sposób pierwszy — formalny — paradoks tak rozumianej metody budowania etyki w ramach modelu z dołu do góry. Odrzucone, zgodnie z tezą Bostroma, jako niewystarczające dla maszyn posługujących się metodami uczenia bez nadzoru reguły rozumowania statystycznego uznaje się za wystarczające w przypadku zapewnienia próbki reprezentatywnej w skali ludzkości i zmieniających problem etyki maszyn autonomicznych w kwestię metodologii budowania bazy danych wzorców etycznych podejmowania decyzji.

Pomimo tego paradoksu zaproponowane przez Yudkowsky'ego podejście do ekstrapolacji natrafiło na spory rezonans i w naszym opracowaniu traktujemy go jako bezpośrednie źródło powstania projektu *Moral Machine* (Awada *et al.*, 2020). Drugim ważnym punktem odniesienia dla tego projektu — który pojawia się zresztą wprost w referencjach autorów — jest koncepcja wskaźników wymiaru kulturowego Geerta Hoftstede'a (Hofstede, 2007).

WYKORZYSTANIE „MIĘKKIEJ RACJONALNOŚCI" DO BUDOWANIA ETYKI MASZYN AUTONOMICZNYCH — PROJEKT *MORAL MACHINE*

Projekt *Moral Machine* (MM) został uruchomiony w 2014 roku dzięki współpracy kilku ośrodków akademickich (Exeter Business School, Massachusetts Institute of Technology, University of British Columbia, Max-Planck Institute for Human Development, Toulouse School of Economics). Miał na celu wygenerowanie możliwie dużej liczby opinii użytkowników internetu na temat decyzji dotyczących dylematów moralnych stworzonych w oparciu o różne modyfikacje klasycznego dylematu wagonika zaproponowanego przez Philippę Foot (Foot, 1967). Dzięki zbudowanej przez autorów stronie internetowej, udostępniono publiczności scenariusze dylematów i na ich podstawie postanowiono

wyłonić rozwiązania, które można wykorzystać jako bazę danych wzorców w systemach autonomicznych — urządzeniem referencyjnym jest tutaj autonomiczny samochód[1]. Wyniki tego eksperymentu w 2018 roku zostały opublikowane na łamach czasopisma naukowego *Nature* (Awada *et al.*, 2018).

Zebrane dane objęły 39,61 miliona decyzji, które pozostawili przedstawiciele 133 krajów. Bazy danych zostały opracowane przy pomocy metody *conjoint* (Aseron, Bhaskaran, & Peruzzi, 2015). Służy ona do klasyfikacji i analizy danych w celu pomiaru preferencji uczestników badania. Jej istotą jest przedstawienie przedmiotu badania jako kombinacji cech. Są one określane jako atrybuty, a każdy z nich ma ustaloną liczbę poziomów. Atrybuty i ich poziomy tworzą różne warianty, które nazywane są profilami. Liczba wszystkich możliwych do wygenerowania profili zależy od tego, ile występuje atrybutów i ich poziomów (jest to iloczyn liczby poziomów wszystkich atrybutów).

W badaniu *Moral Machine* szczególnie poszukiwano wartości *average marginal component effect* (AMCE) każdego z badanych atrybutów sytuacji moralnej, tj. średniego efektu oddziaływania cech atrybutu na ogólny poziom preferencji moralnych. W ten sposób powstała mapa preferencji moralnych przypominająca klasyfikację Hofstede'a (Hofstede, 2007).

W drodze analizy wyłoniono dziewięć atrybutów, które potraktowano jako mierniki preferencji uczestników badania: skłonność do aktywności, relacja wobec uczestników zdarzenia, płeć, sylwetka, status społeczny, status zachowania w stosunku do przepisów o ruchu drogowym, wiek, liczba ocalonych, gatunek (zwierzę — człowiek). Zebrane preferencje respondentów wskazują na większą troskę o: nieaktywność niż aktywność, raczej przechodniów niż pasażerów, kobiety, ludzi w lepszej formie fizycznej i o wyższym statusie społecznym, przestrzegających przepisów niż ich łamiących, młodych w stosunku do wiekowych, raczej dla stosowania strategii utylitarnej w zakresie kalkulacji ilości cierpienia, ludzi w stosunku do zwierząt.

W przypadku poszczególnych typów uczestników sytuacji dylematu chętniej ratowani byliby ludzie niż zwierzęta, a wśród zwierząt raczej preferowane byłyby psy niż koty. Wśród ludzi najchętniej ratowane byłyby z kolei dzieci. Dokonano też próby skorelowania wyników ogólnych z precyzyjnie wybranymi, reprezentatywnymi sześcioma wskaźnikami demograficznymi ważnymi dla całej populacji badanych: wiekiem, wykształceniem, płcią, zamożnością, religią i poglądami politycznymi. Dodatkowa analiza nie wykazała zmian wyników (próba ogranicza się wtedy do 492 291 osób).

Ciekawe rezultaty przyniósł także proces budowania klastrów kulturowych na podstawie zebranych danych, na wzór typologii Hofstede'a (Hofstede, 2007). Wyłoniono 130 krajów. Z każdego kraju wybrano przynajmniej 100 respondentów. W efekcie próba liczyła 448 125 badanych. Dzięki technice klasteryzacji

[1] Strona internetowa projektu: http://moralmachine.mit.edu.

przy wykorzystaniu metryk euklidesowych oraz metody Warda wskazano trzy klastry kulturowe, które pokrywają się z mapą wpływów kulturowych Inglehardta i Welzela (Inglehart & Welzel, 2005). Są to klastry zachodni, wschodni i południowy. Następnie odnotowano pewne różnice między nimi. Na przykład respondenci z kultur kolektywistycznych w klastrze wschodnim, w którym występuje zakorzeniony szacunek do osób starszych, wykazywali mniejszą skłonność do ochrony osób młodych niż w klastrze zachodnim. Wyraźne różnice pojawiły się także w zakresie stosunku do przechodniów łamiących przepisy o ruchu drogowym. W krajach z klastra zachodniego o wysokiej kulturze organizacyjnej i prawnej pojawiła się mniejsza tolerancja wobec zachowań niezgodnych z kodeksem drogowym niż w krajach z klastra południowego o mniejszych tradycjach instytucjonalnych.

Z kolei w krajach o wysokim poziomie współczynnika nierówności społecznych Giniego odnotowano skłonność do większej ochrony osób o wyższym statusie społecznym w porównaniu do osób, które są identyfikowane jako pochodzące z nizin społecznych. To podważa także na przykład uniwersalność niemieckich rozwiązań w tym zakresie zaproponowanych w trybie *a priori* w 2017 roku przez German Ethics Commission on Automated and Connected Driving, w których zakazano tego typu zachowań dyskryminacyjnych.

Poszukiwania klasteryzacji kulturowej w zakresie stosunku do strategii utylitarystycznych lub deontologicznych zostały rozwinięte w kolejnej publikacji autorów projektu *Moral Machines* pt. *Universals and variations in moral decisions made in 42 countries by 70,000 participants*. Jako element różnicujący dla klastrów wybrano wskaźnik mobilności społecznej (Awada *et al.*, 2020). Dla celów badania zebrano 70 tysięcy odpowiedzi w 10 językach z 42 krajów. Do opracowania przyjęto minimalny poziom 200 odpowiedzi na zaproponowany scenariusz z jednego kraju.

Posługując się wynikami badań Joshuy Greena (Green, 2013), przyjęto założenie o podwyższonych preferencjach wobec scenariuszy deontologicznych podejmowania decyzji w sytuacji zadania bezpośredniej, a nie pośredniej śmierci — ze względu na niechęć do brania odpowiedzialności za bezpośrednie uśmiercanie wzrasta wtedy tendencja do porzucenia aktywności. Te założenia skorelowano ze wskaźnikiem mobilności społecznej. Założono prawidłowość, iż wysoki wskaźnik mobilności społecznej ułatwia zachowania, które są niepopularne społecznie, i umożliwia stosowanie czysto racjonalnych strategii utylitarystycznych. Z kolei niska mobilność społeczna powoduje, że do głosu dochodzą ograniczenia i zahamowania obniżające swobodę stosowania kryteriów utylitarystycznych. Dodatkowym elementem, który odgrywał rolę w kształtowaniu się wyników badania była specyfika kulturowa poszczególnych krajów. W przypadku krajów azjatyckich niższa mobilność społeczna skorelowana została z niższą skłonnością do wyrażania kontrowersyjnych opinii oraz wchodzenia w konflikt z otoczeniem.

Z analizy danych wyłoniły się wyraźne preferencje wyboru strategii utylitarystycznych w przypadku krajów europejskich oraz Ameryki Północnej i Południowej. W przypadku krajów azjatyckich wykazano generalnie większą skłonność do ich odrzucania. Jest ona spowodowana obawą przed opinią na temat zachowań niezgodnych ze społecznym tabu deontologicznym, zakazującym świadomego zabijania.

WYZWANIA DLA ETYKI BUDOWANEJ NA PODSTAWIE DANYCH

Wobec zaprezentowanych wyników badań pojawiły się liczne komentarze. Zdaniem autorów polemiki pt. *Life and death decisions of autonomous vehicles* (Bigman & Gray, 2020) przy konstruowaniu wzorców dla maszyn działających autonomicznie ujawnionych w trakcie projektu MM nie należy brać pod uwagę preferencji nierównego traktowania — rasy, płci, wieku podmiotów moralnych, których dotyczą decyzje. Te empiryczne wyniki należy zignorować na rzecz podejścia normatywnego, które preferuje postawę egalitarystyczną, a pytania w ramach badania powinny być konstruowane tak, ażeby uniemożliwiać ujawnianie nierównościowych uprzedzeń.

Wedle innej argumentacji, zawartej chociażby w tekście pt. *Trolled by the trolley problem*, powinny być budowane takie systemy AGI, które potrafią przewidzieć z góry sytuację dylematu i go uniknąć. Sama sytuacja dylematu moralnego jest już bowiem porażką techniczną i dobre systemy AGI powinny być wyposażone w odpowiednie filtry algorytmiczne uniemożliwiające ich powstanie (Mirnig & Meschtscherjakov, 2019).

Niezależnie od tych krytycznych głosów u podstaw opisanego powyżej podejścia typu *data driven* do tworzenia wzorców etycznych dla systemów AGI leży przekonanie, że wykorzystanie tzw. miękkiej racjonalności do konstruowania etyki maszyn autonomicznych powinno być zgodne z koncepcją ART (Dignum, 2017), ponieważ takie podejście zwiększa szansę na ich rynkowy sukces i tworzy korzystne z punktu widzenia ekonomicznego otoczenie dla ich rozwoju. Jednocześnie wykorzystanie procedur „miękkiej racjonalności", które stoją za procesami uczenia maszynowego stosowanego przez systemy AGI (Kaplan, 2023) budzi poważne wątpliwości z punktu widzenia normatywnego — dopuszcza na przykład wzorce odwołujące się do lokalnych nierówności oraz uprzedzeń jako empirycznie istniejących i, w imię budowania zaufania, koniecznych do ujawnienia w bazach danych służących jako wzorce do podejmowania decyzji przez autonomiczne maszyny.

Pojawiają się jednak wymagające poważnej refleksji pytanie, czy taka, zbudowana w trybie *data driven*, etyka spełnia warunki systemu działającego według zasady trzecioosobowej, tzn. pozwalającego na swobodną zamianę ról społecznych i symetryzm z zachowaniem ochrony podmiotu moralnego

niezależnie od funkcji, jaką ten pełni. Czy na przykład losowe przyjście na świat podmiotu moralnego jako przedstawiciela gorzej sytuowanej części społeczeństwa w miejscach, gdzie występuje wysoki współczynnik Giniego, wymaga pogodzenia się z empirycznym faktem społecznej dyskryminacji? Jest to przecież sprzeczne ze zdroworozsądkową intuicją moralną. Zaufanie jest w końcu uważane za wartość, której normatywną podstawą powinno być przekonanie o społecznym dążeniu do sprawiedliwości, co ma zapobiegać między innymi wybuchom przemocy wywoływanej przez grupy, które uważają się za upośledzone w redystrybucji wartości z powodu przypadkowego faktu przyjścia na świat w części świata lub w społeczeństwach gorzej wyposażonych w szanse.

Budowanie etyki maszyn autonomicznych w oparciu o metody *data driven* odwołujące się do lokalnych klastrów kulturowych niesie więc za sobą z jednej strony potencjał budowania zaufania przez uwzględnienie lokalnej specyfiki, z drugiej zaś ryzyko utrwalenia wzorców dyskryminacji oraz nieegalitarnych uprzedzeń, które generują niebezpieczeństwo protestów przeciwko stosującym je systemom. Może i powinno to obniżać szanse rynkowe takich systemów.

Pomimo zaawansowanych prac nad wprowadzeniem regulacji prawnych typu *top–down* porządkujących zasady tworzenia narzędzi typu AGI (Zenner, 2022), ciągle stoimy przed zadaniem stworzenia baz wzorców etycznych w oparciu o metody hybrydowe — łączące podejście *bottom–up* z podejściem *top–down*. Takie systemy hybrydowe wymagałoby jednak odmiennego podejścia do zagadnienia budowania zaufania społecznego wobec systemów AGI. Podstawową motywacją twórców programu *Moral Machine* była chęć skonstruowania bazy danych empirycznie rejestrowanych wzorców, oddających lokalną specyfikę i kulturową różnorodność. Zgodnie z koncepcją CEV Yudkowsky'ego celem tym wysiłków było przekształcenie zagadnienia etyki maszyn w problem techniki konstruowania bazy danych wzorców decyzji. Ograniczeniem takiego podejścia jest przede wszystkim niedoskonałość świadomości sytuacyjnej współczesnych systemów AGI. Nie mogą one uzyskać pełnego zobrazowania sytuacji etycznej potrzebnego do uruchomienia właściwego wzorca z bazy danych, co naraża je na dużą liczbę pomyłek podważających zaufanie do ich działania. W związku z tym techniczny sposób rozwiązania problemu etyki maszyn autonomicznych zaproponowany przez twórców projektu MM może budzić poważne wątpliwości. By uczynić tę koncepcję realizowalną, konieczne jest więc spełnienie kilku dodatkowych założeń.

Po pierwsze, systemy AGI powinny być wyposażone w poprawny system konstytucji otoczenia w oparciu o zestaw cząstkowych danych pozyskiwanych z dostępnych sensorów oraz źródeł informacji. W przypadku systemów poznawczych będących etycznymi wzorcami dla maszyn AGI, a więc systemów biologicznych, jest to zjawisko powszechne — tylko w oparciu o takie procesy można dokonywać poprawnych rozpoznań sytuacji dla podejmowania stosownych, służących przystosowaniu decyzji. Systemy biologiczne zawsze działają

w trybie niepewności — są w stanie pozyskać zdecydowanie mniejsze ilości danych, niż jest to potrzebne do pełnego uchwycenia konkretnej sytuacji etycznej. W celu zminimalizowania ryzyka błędnych zachowań wytwarzają zatem zestawy zautomatyzowanych reakcji na bodźce odpowiadające sytuacjom znanym z przeszłości. Występuje wówczas zjawisko silnego torowania typu *top–down*, związane z częściową automatyzacją zachowań. Diagnoza i reakcja są także wynikiem wzajemnie uzależnionych od siebie procesów asymilacji i akomodacji (Oleron, Piaget, & Inhelder, 1967). Dlatego pierwszą częścią etycznego testu zaufania powinny być testy modelu świata.

Wsparciem dla nich, jako opis wzorca, może być wykorzystanie wyników współczesnych badań z zakresu teorii rdzennych systemów poznawczych Elisabeth Spelke (Carey & Spelke, 1996) czy Stanislasa Dehaene (Dehaene, 2020), wedle których podmioty empiryczne są wyposażone w pewne wrodzone, rudymentarne modele świata, będące podstawą konstytucji i w celu podtrzymywania równowagi z otoczeniem modyfikowane przez pobieranie i wymianę informacji. Błędy w tym procesie są karane przez ewolucyjny proces wyeliminowania genów z reprodukcji.

Po drugie, wskazane wyżej torowanie afektem, które jest najszybszym sposobem automatyzacji reakcji (Zajonc & Murphy, 1994), dotyczy także rozpoznawania stanów wewnętrznych innych podmiotów moralnych jako będących częścią świadomości sytuacyjnej. Wymaga to z kolei systemu rozumienia zachowań oraz reakcji opartych na przykład o systemy symulacji lub rozumienia stanów wewnętrznych (Gallagher, 2004). Do konstytucji sytuacji etycznej potrzebne są więc także możliwości rozpoznawania emocji i *qualiów* — stanów wewnętrznych innych podmiotów moralnych, które są podstawą tzw. etyk sytuacjonistycznych — różnicujących reakcje w zależności od rozpoznania stanu wewnętrznego podmiotu uwikłanego w proces podejmowania decyzji. Takie testy swoje założenia mogą czerpać ze znanych w psychologii testów teorii umysłu sprawdzających zdolność do empatyzowania.

Po trzecie, ze względu na problem czarnej skrzynki występujący w głębokich systemach uczenia maszynowego opartych o technologię sieci neuronowych, proces budowania zaufania do systemów AGI wyposażonych w bazy danych reakcji oraz konstytucję adekwatnej świadomości sytuacyjnej powinien być uzupełniony o algorytmy filtrujące decyzje wedle kryterium zgodności z zestawem bardziej uniwersalnych oczekiwań etycznych opartych o zasadę etyki trzecioosobowej. Osobnym pytaniem pozostaje tutaj także tryb wyłaniania filtrujących wzorców uniwersalnych, choć można tutaj odwołać się do pewnych rozstrzygnięć prawnych, jak chociażby AI Act (Zenner, 2022).

Dodać trzeba przy tym, że opisywane powyżej hybrydowe modele maszyn moralnych są już od dawna przedmiotem zainteresowania badawczego. Należą do nich między innymi model LIDA, model EDM, MedEthEx czy też EMAS (Cervantes *et al.*, 2020: 519–523). Jak wskazują w swoim artykule Paweł Polak

i Roman Krzanowski, proponowane modele hybrydowe nie biorą jednak pod uwagę słabości świadomości sytuacyjnej takich maszyn moralnych. Autorzy ci próbują wskazać rozwiązanie tego problemu w swoim modelu moralnych robotów fronetycznych (Polak & Krzanowski, 2020).

Także pomysł na zbudowanie testu sprawności etycznej nie jest całkowicie nowy. Pojawił się na początku lat dwutysięcznych w opracowaniach Colina Allena i innych (Allen, Smit, & Wallach, 2000), a dzisiaj jest przedmiotem dyskusji w stosunkowo licznych opracowaniach (np. Arnold & Scheutz, 2016; Hyeongjoom & Sunyong, 2021; Gerdes & Øhrstrøm, 2015). Konkluzje dotyczące oparcia procesu budowania zaufania do systemów AGI opartych o technologię LLM (*large language models*) na kryterium specjalnego testu sprawności etycznej muszą z pewnością odwoływać się do tych badań.

Ponadto jedną z konstytutywnych cech generatywnej technologii LLM wpływających na problem etyki AGI są tzw. halucynacje (Liberty, 2023). Spośród trzech rozpatrywanych sposobów rozwiązania problemu halucynacji — proponowane przez Ilyę Sutskevera uczenie ze wzmocnieniem z wykorzystaniem informacji zwrotnych od człowieka (*reinforcement learning with human beeings feedback*), Eda Liberty'ego budowanie systemów pamięci długotrwałej z systemem wektorowym (*long term memory using vector embeddings*) oraz Yana Lacuna tworzenie modeli wiedzy tła (*building world models*) (Liberty, 2023), najbardziej obiecujący wydaje się ten trzeci projekt. Jego rozwój może także pomóc w rozwiązaniu zagadnienia konstytucji sytuacji etycznej, która jest zasadniczym elementem proponowanych etycznych modeli hybrydowych.

BIBLIOGRAFIA

Allen, C., Smit, I., & Wallach, W. (2005). Artificial morality: top–down, bottom–up, and hybrid approaches. *Ethics and Information Technology*, *7*, 149–155.

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

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

Arnold, T. & 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: doi.org/10.1007/s10676-016-9389-x.

Asimov, I. (2004). Runaround (s. 31–48). W: I. Asimov. *I Robot*. New York: Bantam Books.

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

Awada, E., Dsouza, S., Shariffc, A., Rahwanb, I., & Bonnefon, J.F. (2020). Universals and variations in moral decisions made in 42 countries by 70,000 participants. *Proceedings of the National Academy of Sciences*, *117*(5), 2332–2337. DOI: doi.org/10.1073/pnas.191151711.

Aseron, R., Bhaskaran, V., & Peruzzi, N. (2015). *A beginner's guide to conjoint analysis*. Dostęp: https://www.youtube.com/watch?v=RvmZG4cFU0k (04.07.2022).
Bigman, Y. & Gray, K. (2020). Life and death decisions of autonomous vehicles. *Nature*, *579*, E1–E2. DOI: doi.org/10.1038/s41586-020-1987-4.
Bochen, M. (2019). Epistemiczna wartość doświadczenia zmysłowego. Wilfrid Sellars versus John McDowell. *Kultura i Wartości*, *27*, 191–217.
Bostrom, N. (2014). *Superinteligencja. Scenariusze, strategie, zagrożenia*. (Przeł. D. Konowrocka-Sawa). Gliwice: Helion.
Brock, H.W. (1980). *Game theory, social choice and ethics*. Dordrecht–Boston–London: D. Reidel Publishing Company.
Brożek, B. (2001). *Kilka uwag o logice norm*. Dostęp: https://ruj.uj.edu.pl/xmlui/bitstream/handle/item/80424/brozek_kilka_uwag_o_logice_norm_2001.pdf?sequence=1&isAllowed=y (15.12.2022).
Carey, S. & Spelke, E. (1996). Science and core knowledge. *Philosophy of Science*, *63*, 515–533.
Cervantes, J.-A., López, S., Rodríguez, L.-F., Cervantes, S., Cervantes, F., Ramos, F. (2020). Artificial moral agents: A survey of the current status. *Science and Engineering Ethics*, *26*(2), 501–532. DOI: 10.1007/s11948-019-00151-x.
Davidson, D. (1984). On the very idea of conceptual scheme (s. 5–20). W: D. Davidson. *Inquiries into truth and interpretation.* Oxford: Oxford University Press.
Davidson, D. (2005). Seeing through language (s. 127–141). W: D. Davidson. *Truth, language, and history*. New York–Oxford: Clarendon Press / Oxford University Press.
Dehaene, S. (2020). *How we learn: why brains learn better than any machine… for now*. New York: Viking Press.
de Wall, F. (2012). *Zachowanie moralne u zwierząt*. Dostęp: https://www.youtube.com/watch?v=VyGN92UAnjI (20.12.2022).
Dignum, V. (2017). *Responsible autonomy*. Dostęp: https://arxiv.org/pdf/1706.02513.pdf. (20.12.2022).
Foot, Ph. (1967), The problem of abortion and the doctrine of the double effect (s. 5–15). W: Ph. Foot. *Virtues and vices: and other essays in moral philosophy*. Berkeley: University of California Press. DOI: doi.org/10.1093/0199252866.003.0002.
Fukuyama, F. (1997). *Zaufanie. Kapitał społeczny a droga do dobrobytu*. (Przeł A. Śliwa & L. Śliwa). Warszawa: Wydawnictwo Naukowe PWN.
Gallagher, S. (2004). Hermeneutics and the cognitive science. *Journal of Consciousness Studies*, *11*, 162–174.
Gerdens, A. & 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: doi.org/10.1108/JICES-09-2014-0038.
Greene, J. (2013). *Moral tribes: emotion, reason and the gap between us and them*. London: Atlantic Books.
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.
Gryz, J. (2021). *Sztuczna inteligencja: powstanie, rozwój, rokowania*. Dostęp: https://www.youtube.com/watch?v=3ZDfVgC897k (17.06.2021).
Hyeongjoo, K. & Sunyong, B. (2021). Designing and applying a moral Turing test. *Advances in Science, Technology and Engineering Systems Journal*, *6*(2), 93–98.
Hofstede, G. (2007). *Kultury i organizacje. Zaprogramowanie umysłu*. (Przeł. M. Durska). Warszawa: PTE.
Inglehart, R. & Welzel, C. (2005). *Modernization, cultural change, and democracy: The human development sequence*. Cambridge: Cambridge University Press.

Jörgensen, J. (1938). Imperatives and logic. *Erkenntnis*, *7*(4), 288–296.
Kaplan, C. (2023). *Artificial intelligence: past, present, and future*. Dostęp: https://www.youtube.com/watch?v=ZTt_GI0-wKA (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.
Kusch, M. (1989). *Language as calculus vs. language as universal medium. A study in Husserl, Heidegger and Gadamer*. Dodrecht–Boston–London: D. Reidel Publishing Company.
Liberty, E. (2023). *Solving ChatGPT hallucinations with vector embeddings*. Dostęp: https://www.youtube.com/watch?v=FUgp4oaxj-M (15.02.2023).
Makowski, P. (2011). Gilotyna Hume'a. *Przegląd Filozoficzny — Nowa Seria*, *4*(76), 1–15.
MacIntyre, A. (1996). *Dziedzictwo cnoty. Studium z teorii moralności*. (Przeł. A. Chmielewski). Warszawa: Wydawnictwo Naukowe PWN.
McDowell, J. (2008). Avoiding the myth of the given (s. 1–14). W: J. Lingard (Red.). *John BcDowell. Experience, norm, and nature*. Hoboken: Blackwell Publishing.
McDowell, J. (1996). *Mind and world*. Boston: Harvard University Press.
Mirnig, A. & Meschtscherjakov, A. (2019). Trolled by the trolley problem. On what matters for ethical decision making in automated vehicles. *CHI '19: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, *509*, 1–10. DOI: doi.org/10.1145/3290605.3300739.
Oleron, P., Piaget, J., Inhelder, B., & Greco, P. (1967). *Inteligencja*. (Przeł. M. Przetacznikowa). Warszawa: Państwowe Wydawnictwo Naukowe.
Pigden, Ch. (1989). Logic and the autonomy of ethics. *Australasian Journal of Philosophy*, *67*(2), 127–151.
Polak, P. & Krzanowski, R. (2020). Phronetic ethics in social robotics: A new approach to building ethical robots. *Studies in Logic, Grammar and Rhetoric*, *63*(76), 165–DOI: doi.org/10.2478/slgr-2020-0033.
Rorty, R. (1994). *Filozofia a zwierciadło natury*. (Przeł. M. Szczubiałka). Warszawa: Wydawnictwo Spacja / Fundacja Aletheia.
Searle, J. (1987). Jak wywieść „powinien" z „jest" (s. 220–221). W: J. Searle. *Czynności mowy*. (Przeł. B. Chwedeńczuk). Warszawa: PAX.
Sellars, W. (1991). Empiryzm i filozofia umysłu. (s. 173–257). (Przeł. J. Gryz). W: B. Stanosz (Red.). *Empiryzm współczesny*. Warszawa: Wydawnictwo Uniwersytetu Warszawskiego.
Sellars, W. (1997). *Empiricism and the philosophy of mind*. Harvard: Harvard University Press.
Weinberger, O. (1984). Is and ought reconsidered. *Archiv fur Rechts und Sozialphilosophie*, *70*(4), 454–469.
Woleński, J. (1980). *Z zagadnień analitycznej filozofii prawa*. Warszawa–Kraków: Państwowe Wydawnictwo Naukowe.
Quine, W. van O. (2000). Dwa dogmaty empiryzmu (s. 53–65). W: W. van O. Quine. *Z punktu widzenia logiki*. (Przeł. B. Stanosz). Warszawa: Aletheia.
Yudkowsky, E. (2004). *Coherent extrapolated volition*. San Francisco: The Singularity Institute.
Zajonc, R. & 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 (s. 111–121). W: J. Stelmach (Red.). *Studia z filozofii prawa*. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego.
Zenner, K. (2022). The AI Act. Dostęp: https://artificialintelligenceact.eu/documents/ (20.02.2023).