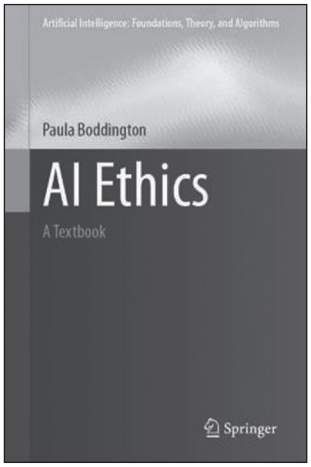
ARGUMENT
BIANNUAL PHILOSOPHICAL JOURNAL
e-ISSN 2084–1043 p-ISSN 2083–6635
Published online: 01.09.2023
Vol. 13 (1/2023) pp. 203–207
www.argument-journal.eu

Paula Boddington, *AI ethics: A textbook*

Berlin: Springer, 2023, ss. 522.
Seria: *Artificial Intelligence: Foundations, Theory, and Algorithms*.
ISSN 2365-3051, ISSN 2365-306X (electronic), ISBN 978-981-19-9381-7,
ISBN 978-981-19-9382-4 (eBook)



Sztuczna inteligencja (*artificial intelligence*, AI) jest już silnie obecna w naszym świecie, tak w domu, jak i pracy pod różnymi postaciami. Wspiera nas, wykonując różne zadania. Dużo potrafi, ale też nieustannie się rozwija. I to bardzo szybko. Najróżniejsze zastosowania sztucznej inteligencji wiążą się ze złożonymi zagadnieniami etycznymi — obecnie analizowanymi wieloaspektowo i szeroko dyskutowanymi. Dobrym świadectwem tego rodzaju analiz i dyskusji jest recenzowana książka *AI ethics: A textbook* Pauli Boddington, wydana w serii *Artificial Intelligence: Foundations, Theory and Algorithms*, której celem jest upowszechnianie aktualnej wiedzy z zakresu sztucznej inteligencji (obejmującej między innymi takie zagadnienia jak zautomatyzowane rozumowania, reprezentacja wiedzy, uczenie maszynowe, uczenie głębokie, przetwarzanie języka naturalnego, etyka AI i odpowiedzialne działanie), jak również metodologii, technologii oraz różnych zastosowań AI w życiu.

Publikacja składa się z jedenastu rozdziałów. W rozdziale I, *Introduction: Why AI ethics?*, autorka objaśnia podstawowe pojęcia z zakresu nauk o sztucznej inteligencji oraz etyki („słaba sztuczna inteligencja", „silna sztuczna inteligencja", „ogólna sztuczna inteligencja", „superinteligencja", „etyka sztucznej inteligencji") oraz uzasadnia potrzebę tego rodzaju wiedzy i studiów. W rozdziale II, *The rise of AI ethics*, Boddington zarysowuje kontekst historyczny, dokonuje przeglądu niektórych historycznych reakcji na technologię i tworzy pewne tło dla intensywnego obecnie zainteresowania kwestiami etycznymi

w AI. Ponadto wskazuje na pewne ważne i zawsze aktualne problemy: wolności i autonomii, odpowiedzialności, transparentności i zrozumienia, sprawiedliwości i uczciwości, dobroczynności i nieszkodzenia innym, prywatności, zaufania, zrównoważonego rozwoju, godności i solidarności. Rozdział III, *AI, philosophy of technology, and ethics*, zawiera analizę kilku głównych kwestii z zakresu filozofii technologii (czym ona jest i co jest jej właściwym celem), które mogą wzbogacić rozumienie kwestii etycznych, powiązanych ze sztuczną inteligencją. Rozdział IV, *Methods in applied ethics*, zawiera rozważania metodologiczne dotyczące tego, jak reagujemy na pewne sytuacje moralne i jak możemy poprawić tę pierwotną reakcję oraz czy to samo dotyczy maszyn — czy dylematy moralne oraz wartości i sądy etyczne mogą być sformalizowane i precyzyjnie wyrażone w języku sztucznej inteligencji. Rozdział V, *Humans and intelligent machines: underlying values*, pogłębia niektóre kwestie dotyczące relacji człowiek–sztuczna inteligencja, ale też powraca do fundamentalnych pytań: kim jest człowiek, jaką ma naturę, czym jest inteligencja, dlaczego ją cenimy. Rozważa niektóre etyczne implikacje odpowiedzi na te pytania. W rozdziale VI, *Normative ethical theory and AI ethics*, autorka dokonuje przeglądu głównych teorii etyki normatywnej (konsekwencjalizm, deontologia i etyka cnót) i ocenia ich przydatność w kontekście uczenia sztucznej inteligencji i robotów zachowań etycznych (Czy działające roboty wytwarzają moralne konsekwencje? Czy maszyny wykazują racjonalność praktyczną? Czy robota można nauczyć praktykowania cnót?). W rozdziale VII, *Philosophy for AI ethics: Metaethics, metaphysics, and more*, Boddington pyta między innymi o to, jakie jeszcze związane z podstawami etyki sprawy i zagadnienia musimy zbadać oraz jakie inne obszary filozofii (na przykład filozofii umysłu) uwzględnić, aby dobrze zrozumieć etykę sztucznej inteligencji. W rozdziale VIII, *Persons and AI*, autorka rozważa następujące zagadnienia: Jak rozumiano kategorię osoby? Jak przypisuje się komuś osobowość? Jakie rodzaje stworzeń mogą być osobami? Czy istnieją różne kategorie lub stopnie osobowości? W jaki sposób osoby są indywidualizowane? W jaki sposób osoby są identyfikowane w czasie? Rozważa również takie kwestie jak: ucieleśnienie osoby, sprawczość, rola poznania, jaźni i świadomości w opisach osobowości, autonomia osoby — warunki i ograniczenia. Rozdział IX, *Individuals, society, and AI: Online communication*, zawiera analizę konkretnych współczesnych problemów etycznych, wynikających między innymi z naszej obecności w sieci internetowej, na przykład problemy nadużyć i krzywdy w sieci, w tym problem mowy nienawiści (jej definicje, odmiany i stopnie), a także wykorzystanie sztucznej inteligencji (algorytmów) do ich wykrywania i monitorowania. W rozdziale X, *Towards the future with AI: Work and superintelligence*, poruszone zostały dwie grupy zagadnień: kwestie związane z zastosowaniem sztucznej inteligencji w miejscu pracy oraz problem bezpiecznej superinteligencji, kontroli superinteligencji i monitorowania, na ile będzie ona życzliwa człowiekowi i na ile będzie działała zgodnie z ludzkimi celami. Ostatni rozdział,

Our future with AI: Future projections and moral machines, jest naturalną kontynuacją poprzedniego i dotyczy przyszłości sztucznej inteligencji. Autorka stawia między innymi następujące pytania: W jaki sposób ludzie i sztuczna inteligencja odnoszą się do siebie nawzajem? Czy człowiek ma jakieś moralne zobowiązania wobec maszyn? Czy możliwe jest udoskonalenie człowieka za pomocą sztucznej inteligencji, w tym udoskonalenie moralne?

Podsumowując, książka oferuje czytelnikom niezbędny wgląd w szybko rozwijającą się dziedzinę etyki AI, a przy tym uwzględnia najlepsze dotychczasowe rozwiązania, jakie wypracowano w ramach różnych kierunków etycznych. Daje dobre wprowadzenie do wybranych zagadnień, pojęć, teorii i debat związanych z etyką, wyraźnie wskazując na silną potrzebę przemyślenia starych problemów w kontekście rozwoju sztucznej inteligencji oraz jej obecnych i przyszłych zastosowań praktycznych. Niektóre zagadnienia dotyczące etyki sztucznej inteligencji analizowane są w szerszym kontekście i uwzględniają pewne zagadnienia z zakresu filozofii, religii i kultury. Książka ma służyć jako pomoc, przewodnik lub podręcznik dla studentów, toteż wypełniona jest licznymi studiami przypadków, przykładami i ćwiczeniami. Każdy rozdział rozpoczyna się krótkim streszczeniem, a kończy „uwagami i wskazówkami nauczyciela", bibliografią oraz listą prac, z którymi czytelnik może się zapoznać w celu dodatkowego i bardziej szczegółowego opracowania poszczególnych tematów lub wątków. Na końcu publikacji znajdujemy jedenastostronicowy słowniczek terminów wykorzystanych w podręczniku oraz indeks podstawowych pojęć i nazwisk.

Książka jest imponująca i erudycyjna, co nie znaczy, że nie ma w niej pewnych drobnych mankamentów i braków. Niedosyt budzi na przykład rozdział VII. Boddington podkreśla: „Wiele pytań dotyczących etyki sztucznej inteligencji można lepiej zrozumieć, jeśli zostaną omówione kwestie metaetyczne"[1] (s. 277). I słusznie. A dwie strony dalej czytamy:

> Wiele pytań dotyczących etyki sztucznej inteligencji rodzi również pytania dotyczące metaetyki. Kwestie metaetyczne leżą u podstaw procesu formułowania jakiejkolwiek oceny na temat sztucznej inteligencji lub dowolnego kodeksu etycznego dotyczącego sztucznej inteligencji. Ponadto niektóre konkretne kwestie dotyczące potencjału sztucznej inteligencji i jej zastosowań rodzą pytania w metaetyce. Czy sztuczna inteligencja mogłaby rzeczywiście podejmować decyzje moralne, działać jako agent moralny, zasługiwać na szacunek jako moralny podmiot oraz zdobywać wiedzę i zrozumienie moralne? (s. 279)

Wiemy już, że sztuczna inteligencja doskonale rozpoznaje obiekty, trafnie interpretuje i reaguje na pewne zdarzenia, dokonuje bezbłędnie działań logiczno-matematycznych, całkiem sprawnie komunikuje się z ludźmi (asystenci głosowi, tacy jak Amazon Alexa, ucieleśnione roboty, ChatGPT), ale wciąż ma poważne problemy ze sprawami najbardziej podstawowymi, na przykład

[1] Cytaty w przekładzie autora.

z instynktem, samopoczuciem czy intuicjami. W ramach współczesnych debat metaetycznych bardzo poważnie traktuje się intuicjonizm etyczny, głoszący prymat intuicji w poznaniu i ocenie działania, które jakiś podmiot uważa za słuszne lub niesłuszne, dobre lub złe. Postawa moralna wedle intuicjonistów jest wyrazem czyjegoś odczucia i czyjejś samoświadomości. Tutaj pojawia się mnóstwo ważnych pytań, których Boddington nie stawia: Czy intuicjonizm jest wartościowym stanowiskiem? Czym dokładnie jest intuicja? Czy można zredukować intuicję do emocji? Czy możliwe są sztuczne emocje? Czym dokładnie jest świadomość fenomenalna? Czy samoświadomość jest czymś jakościowo różnym od świadomości, czy rzeczywiście jest świadomością wyższego rzędu? Jakie podstawowe warunki należy spełnić, aby stworzyć sztuczną świadomość?

Albo dwie inne kwestie, ściśle związane z powyższymi rozważaniami: kwestia implementacji sztucznej moralności oraz kwestia możliwych typów sztucznych agentów. Problem implementacji stawia pytanie o to, jak postępować podczas projektowania sztucznego agenta moralnego. Standardowo rozróżnia się podejścia góra–dół (zdolność moralna jako zdolność stosowania przyjętych zasad moralnych w danym przypadku), dół–góra (zdolności moralne jako mądrość praktyczna, czyli wrażliwość na moralne aspekty danej sytuacji) i hybrydowe. Wszystkie trzy łączą się z różnymi poglądami etycznymi, wszystkie mają swoje mocne i słabe strony. Choć na razie jesteśmy w początkowym stadium refleksji i projektowania moralnych maszyn, powstają już pierwsze klasyfikacje agentów etycznych. Godna rozważenia i dyskusji jest na przykład klasyfikacja Jamesa H. Moora, który wyróżnił ich cztery rodzaje: 1) agenci, którzy działając, wywierają moralne konsekwencje na życie człowieka, ale bez intencji; 2) implicytni agenci etyczni, których konstrukcje nieświadomie realizują pewne wartości moralne, na przykład ostrzegają przed niebezpieczeństwem (różnego rodzaju zabezpieczenia, systemy ostrzegania); 3) eksplicytni agenci etyczni, którzy zaprojektowani są w ten sposób, że mogą wprost rozpoznawać i przetwarzać moralnie istotne informacje i podejmować decyzje moralne; 4) autonomiczni agenci etyczni, którzy są etyczni na sposób ludzki, czyli — znów — jaki? Czy klasyfikacja ta jest zadowalająca? Boddington ma świadomość tych problemów, ale nie podejmuje ich w swoich rozważaniach w pełni.

Nie są to poważne zarzuty. Raczej wskazówki, co warto podjąć i przepracować w kolejnym wydaniu.

Na koniec warto podkreślić, że istnieje już na rynku wydawniczym kilka bardzo dobrych książek, które oferują przegląd wiodących zagadnień i problemów etycznych z zakresu etyki sztucznej inteligencji. Są to między innymi: Markus D. Dubber, Frank Pasquale, Sunit Das (Red.), *The Oxford handbook of ethics of AI* (Oxford: Oxford University Press, 2020); Mark Coeckelbergh, *AI ethics* (Cambridge: The MIT Press, 2020); tegoż, *Robot ethics* (Cambridge: The MIT Press, 2022); S. Matthew Liao (Red.), *Ethics of artificial intelligence* (Oxford: Oxford University Press, 2022); Wayne Holmes, Kaśka Porayska-Pomsta, *The*

ethics of artificial intelligence in education practices, challenges, and debates (London: Routledge, 2022). Autorzy tych publikacji szczegółowo opisują, rozważają, analizują i wyjaśniają różne bardziej lub mniej ogólne problemy z zakresu etyki sztucznej inteligencji (może z wyjątkiem ostatniej pozycji, która ma nieco inne ambicje). Jednak recenzowana książka charakteryzuje się tym, że uwzględnia większość ważnych dla tej dziedziny problemów, ujmując je całościowo, w sposób jasny, uporządkowany i spójny. Ale też — co ważniejsze — realizuje ważny wymiar dydaktyczny: inspiruje i aktywizuje odbiorcę do stawiania różnych pytań, podejmowania działań oraz rozwiązywania problemów teoretycznych i praktycznych.

*Andrzej DĄBROWSKI**

* Dr hab., Instytut Filozofii i Socjologii, Uniwersytet Pedagogiczny im. KEN w Krakowie. E-mail: andrzej.dabrowski@up.krakow.pl.