

SPIS TREŚCI

Wstęp: wprowadzenie do bezpieczeństwa i ochrony sztucznej inteligencji xi	
Podziękowania xxvii	
Redaktor naukowy xxix	
Współpracownicy xxxi	
Część I Obawy luminarzy	1
Rozdział 1 Dlaczego przyszłość nas nie potrzebuje Bill Joy	3
Rozdział 2 Głęboko przeplatana obietnica i niebezpieczeństwo GNR Ray Kurzweil	25
Rozdział 3 Podstawowe pobudki SI Stephen M. Omohundro	59
Rozdział 4 Etyka sztucznej inteligencji Nick Bostrom i Eliezer Yudkowsky	71
Rozdział 5 Przyjazna sztuczna inteligencja: Wyzwanie fizyki Max Tegmark	87
Rozdział 6 MDL destylacja inteligencji: Poznanie strategii bezpiecznego dostępu do superinteligentnych możliwości rozwiązywania problemów K. Eric Drexler	93
Rozdział 7 Problem uczenia się wartości Nate Soares	111
Rozdział 8 Przykłady kontradiktoryjne w świecie fizycznym Alexey Kurakin, Ian J. Goodfellow i Samy Bengio	123
Rozdział 9 W jaki sposób może zaistnieć SI? Różne podejścia i ich implikacje dla życia we wszechświecie David Brin	141
Rozdział 10 Przyszłość MADCOM: Jak sztuczna inteligencja może wzmocnić propagandę obliczeniową, przeprogramować ludzką kulturę oraz zagrozić demokracji i co można z tym zrobić Matt Chesson	159
Rozdział 11 Strategiczne implikacje otwartości w rozwoju sztucznej inteligencji Nick Bostrom	183
Część II Odpowiedzi naukowców	209
Rozdział 12 Korzystanie z ludzkiej historii, psychologii i biologii w celu uczynienia SI bezpieczną dla ludzi Gus Bekdash	211
Rozdział 13 Bezpieczeństwo SI z perspektywy pierwszej osoby Edward Frenkel	251
Rozdział 14 Strategie dla nieprzyjaznej wyroczni SI z przyciskiem resetowania Olle Häggström	261
Rozdział 15 Zmiany celu w inteligentnych agentach Seth Herd, Stephen J. Read, Randall O'Reilly i David J. Jilk	273
Rozdział 16 Ograniczenia weryfikacji i walidacji zachowań agencyjnych David J. Jilk	283
Rozdział 17 Kontradiktoryjne uczenie maszynowe Phillip Kuznetsov, Riley Edmunds, Ted Xiao, Humza Iqbal, Raul Puri, Noah Golmant i Shannon Shih	295

Rozdział 18 Uzgadnianie wartości wykorzystując obliczalną odległość preferencji Andrea Loreggia, Nicholas Mattei, Francesca Rossi i K. Brent Venable	313
Rozdział 19 Racjonalnie uzależniona sztuczna superinteligencja James D. Miller	329
Rozdział 20 Bezpieczeństwo aplikacji robotów z wykorzystaniem ROS David Portugal, Miguel A. Santos, Samuel Pereira i Micael S. Couceiro	341
Rozdział 21 Wybór preferencji społecznej i problem wyrównania wartości Mahendra Prasad	363
Rozdział 22 Rozłączne scenariusze katastrofalnego ryzyka SI Kaj Sotala	395
Rozdział 23 Realizm ofensywny i niezabezpieczona struktura systemu międzynarodowego: Sztuczna inteligencja i globalna hegemonia Maurizio Tinnirello	423
Rozdział 24 Superinteligencja i przyszłość rządów: Priorytetyzacja problemu kontroli na końcu historii Phil Torres	445
Rozdział 25 Wojskowa SI jako zbieżny cel samodoskonalącej się SI Alexey Turchin i David Denkenberger	467
Rozdział 26 Wrażliwe na wartości podejście do projektowania inteligentnych agentów Steven Umbrello i Angelo F. De Bellis	491
Rozdział 27 Konsekwencjalizm, deontologia i bezpieczeństwo sztucznej inteligencji Mark Walker	509
Rozdział 28 Inteligentne maszyny są zagrożeniem dla ludzkości Kevin Warwick	523
Indeks	