Spis treści

Wstęp: wprowadzenie do bezpieczeństwa i ochrony sztucznej inteligencji xi
Podziękowania xxvii
Redaktor naukowy xxix
Współpracownicy xxxi

Część I Obawy luminarzy

Rozdział 1 Dlaczego przyszłość nas nie potrzebuje 3
Bill Joy

Rozdział 2 Głęboko przeplatana obietnica i niebezpieczeństwo GNR 25
Ray Kurzweil

Rozdział 3 Podstawowe pobudki SI 59
Stephen M. Omohundro

Rozdział 4 Etyka sztucznej inteligencji 71
Nick Bostrom i Eliezer Yudkowsky

Rozdział 5 Przyjazna sztuczna inteligencja: Wyzwanie fizyki 87
Max Tegmark

Rozdział 6 MDL destylacja inteligencji: Poznawanie strategii bezpiecznego dostępu do superinteligentnych możliwości rozwiązywania problemów 93
K. Eric Drexler

Rozdział 7 Problem uczenia się wartości 111
Nate Soares

Rozdział 8 Przykłady kontradyktoryjne w świecie fizycznym 123
Alexey Kurakin, Ian J. Goodfellow i Samy Bengio

Rozdział 9 W jaki sposób może zaistnieć SI? Różne podejścia i ich implikacje dla życia we wszechświecie 141
David Brin

Rozdział 10 Przyszłość MADCOM: Jak sztuczna inteligencja może wzmocnić propagandę obliczeniową, przeprogramować ludzką kulturę oraz zagrozić demokracji... i co można z tym zrobić 159
Matt Chessen

Rozdział 11 Strategiczne implikacje otwartości w rozwoju sztucznej inteligencji ... 183
Nick Bostrom

Część II Odpowiedzi naukowców

Rozdział 12 Korzystanie z ludzkiej historii, psychologii i biologii w celu uczynienia SI bezpieczną dla ludzi 211
Gus Bekdash

Rozdział 13 Bezpieczeństwo SI z perspektywy pierwszej osoby 251
Edward Frenkel

Rozdział 14 Strategie dla nieprzyjaznej wyroczni SI z przyciskiem resetowania 261
Olle Häggström

Rozdział 15 Zmiany celu w inteligentnych agentach 273
Seth Herd, Stephen J. Read, Randall O'Reilly i David J. Jilk

Rozdział 16 Ograniczenia weryfikacji i walidacji zachowań agencyjnych 283
David J. Jilk

Rozdział 17 Kontradyktoryjne uczenie maszynowe 295
Phillip Kuznetsov, Riley Edmunds, Ted Xiao, Humza Iqbal, Raul Puri, Noah Golmant i Shannon Shih

Rozdział 18 Uzgadnianie wartości wykorzystując obliczalną odległość preferencji 313
Andrea Loreggia, Nicholas Mattei, Francesca Rossi i K. Brent Venable

Rozdział 19 Racjonalnie uzależniona sztuczna superinteligencja 329
James D. Miller

Rozdział 20 Bezpieczeństwo aplikacji robotów z wykorzystaniem ROS 341
David Portugal, Miguel A. Santos, Samuel Pereira i Micael S. Couceiro

Rozdział 21 Wybór preferencji społecznej i problem wyrównania wartości 363
Mahendra Prasad

Rozdział 22 Rozłączne scenariusze katastrofalnego ryzyka SI 395
Kaj Sotala

Rozdział 23 Realizm ofensywny i niezabezpieczona struktura systemu międzynarodowego: Sztuczna inteligencja i globalna hegemonia 423
Maurizio Tinnirello

Rozdział 24 Superinteligencja i przyszłość rządów: Priorytetyzacja problemu kontroli na końcu historii 445
Phil Torres

Rozdział 25 Wojskowa SI jako zbieżny cel samodoskonalącej się SI 467
Alexey Turchin i David Denkenberger

Rozdział 26 Wrażliwe na wartości podejście do projektowania inteligentnych agentów 491
Steven Umbrello i Angelo F. De Bellis

Rozdział 27 Konsekwencjalizm, deontologia i bezpieczeństwo sztucznej inteligencji 509
Mark Walker

Rozdział 28 Inteligentne maszyny są zagrożeniem dla ludzkości 523
Kevin Warwick

Indeks 533