

---

# Spis treści

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

## **Część I    Obawy luminarzy**

<b>Rozdział 1</b>	Dlaczego przyszłość nas nie potrzebuje .....	3
	<i>Bill Joy</i>	
<b>Rozdział 2</b>	Głęboko przeplatana obietnica i niebezpieczeństwo GNR .....	25
	<i>Ray Kurzweil</i>	
<b>Rozdział 3</b>	Podstawowe pobudki SI .....	59
	<i>Stephen M. Omohundro</i>	
<b>Rozdział 4</b>	Etyka sztucznej inteligencji .....	71
	<i>Nick Bostrom i Eliezer Yudkowsky</i>	
<b>Rozdział 5</b>	Przyjazna sztuczna inteligencja: Wyzwanie fizyki .....	87
	<i>Max Tegmark</i>	
<b>Rozdział 6</b>	MDL destylacja inteligencji: Poznawanie strategii bezpiecznego dostępu do superinteligentnych możliwości rozwiązywania problemów .....	93
	<i>K. Eric Drexler</i>	
<b>Rozdział 7</b>	Problem uczenia się wartości .....	111
	<i>Nate Soares</i>	
<b>Rozdział 8</b>	Przykłady kontrydiktoryjne w świecie fizycznym .....	123
	<i>Alexey Kurakin, Ian J. Goodfellow i Samy Bengio</i>	
<b>Rozdział 9</b>	W jaki sposób może zaistnieć SI? Różne podejścia i ich implikacje dla życia we wszechświecie .....	141
	<i>David Brin</i>	

<b>Rozdział 10</b>	Przyszłość MADCOM: Jak sztuczna inteligencja może wzmocnić propagandę obliczeniową, przeprogramować ludzką kulturę oraz zagrozić demokracji... i co można z tym zrobić .....	159
--------------------	--	-----

*Matt Chessen*

<b>Rozdział 11</b>	Strategiczne implikacje otwartości w rozwoju sztucznej inteligencji ...	183
--------------------	---	-----

*Nick Bostrom*

## **Część II Odpowiedzi naukowców**

<b>Rozdział 12</b>	Korzystanie z ludzkiej historii, psychologii i biologii w celu uczynienia SI bezpieczną dla ludzi .....	211
--------------------	---	-----

*Gus Bekdash*

<b>Rozdział 13</b>	Bezpieczeństwo SI z perspektywy pierwszej osoby .....	251
--------------------	---	-----

*Edward Frenkel*

<b>Rozdział 14</b>	Strategie dla nieprzyjaznej wyroczni SI z przyciskiem resetowania ....	261
--------------------	--	-----

*Olle Häggström*

<b>Rozdział 15</b>	Zmiany celu w inteligentnych agentach .....	273
--------------------	---	-----

*Seth Herd, Stephen J. Read, Randall O'Reilly i David J. Jilk*

<b>Rozdział 16</b>	Ograniczenia weryfikacji i walidacji zachowań agencyjnych .....	283
--------------------	---	-----

*David J. Jilk*

<b>Rozdział 17</b>	Kontradyktoryjne uczenie maszynowe .....	295
--------------------	--	-----

*Phillip Kuznetsov, Riley Edmunds, Ted Xiao, Humza Iqbal, Raul Puri, Noah Golmant i Shannon Shih*

<b>Rozdział 18</b>	Uzgadnianie wartości wykorzystując obliczalną odległość preferencji	313
--------------------	---	-----

*Andrea Loreggia, Nicholas Mattei, Francesca Rossi i K. Brent Venable*

<b>Rozdział 19</b>	Racjonalnie uzależniona sztuczna superinteligencja .....	329
--------------------	--	-----

*James D. Miller*

<b>Rozdział 20</b>	Bezpieczeństwo aplikacji robotów z wykorzystaniem ROS .....	341
--------------------	---	-----

*David Portugal, Miguel A. Santos, Samuel Pereira i Micael S. Couceiro*

Spis treści	ix
<b>Rozdział 21</b> Wybór preferencji społecznej i problem wyrównania wartości ..... <i>Mahendra Prasad</i>	363
<b>Rozdział 22</b> Rozłączne scenariusze katastrofalnego ryzyka SI ..... <i>Kaj Sotala</i>	395
<b>Rozdział 23</b> Realizm ofensywny i niezabezpieczona struktura systemu międzynarodowego: Sztuczna inteligencja i globalna hegemonia ..... <i>Maurizio Tinnirello</i>	423
<b>Rozdział 24</b> Superinteligencja i przyszłość rządów: Priorytetyzacja problemu kontroli na końcu historii ..... <i>Phil Torres</i>	445
<b>Rozdział 25</b> Wojskowa SI jako zbieżny cel samodoskonalącej się SI ..... <i>Alexey Turchin i David Denkenberger</i>	467
<b>Rozdział 26</b> Wrażliwe na wartości podejście do projektowania inteligentnych agentów ..... <i>Steven Umbrello i Angelo F. De Bellis</i>	491
<b>Rozdział 27</b> Konsekwencjalizm, deontologia i bezpieczeństwo sztucznej inteligencji ..... <i>Mark Walker</i>	509
<b>Rozdział 28</b> Inteligentne maszyny są zagrożeniem dla ludzkości ..... <i>Kevin Warwick</i>	523
<b>Indeks</b> .....	533