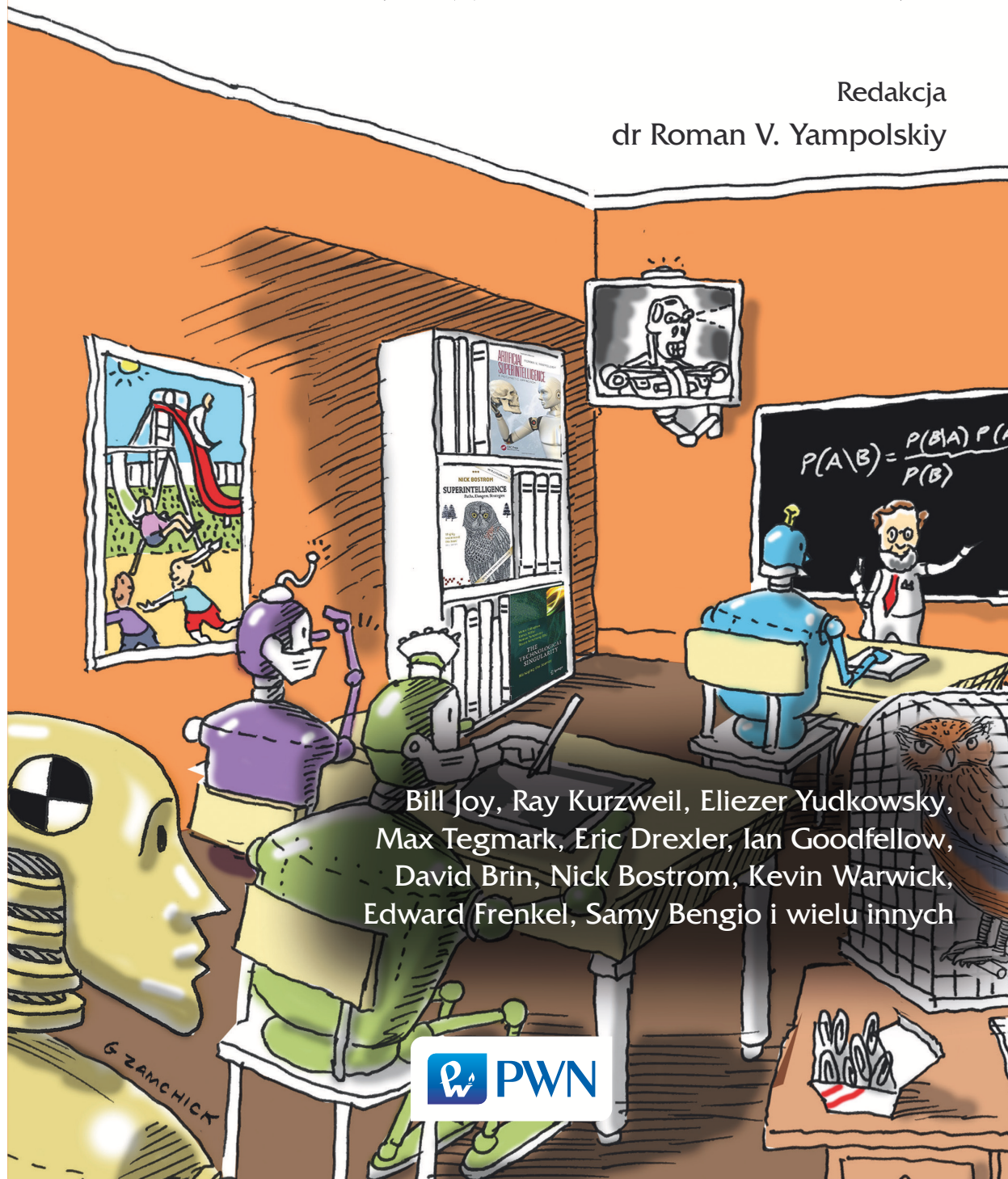


SZTUCZNA INTELIGENCJA BEZPIECZEŃSTWO I ZABEZPIECZENIA

Redakcja
dr Roman V. Yampolskiy



Bill Joy, Ray Kurzweil, Eliezer Yudkowsky,
Max Tegmark, Eric Drexler, Ian Goodfellow,
David Brin, Nick Bostrom, Kevin Warwick,
Edward Frenkel, Samy Bengio i wielu innych

 PWN

SZTUCZNA INTELIGENCJA BEZPIECZEŃSTWO I ZABEZPIECZENIA

SZTUCZNA INTELIGENCJA BEZPIECZEŃSTWO I ZABEZPIECZENIA

Redakcja
dr Roman V. Yampolskiy

Bill Joy, Ray Kurzweil, Eliezer Yudkowsky,
Max Tegmark, Eric Drexler, Ian Goodfellow,
David Brin, Nick Bostrom, Kevin Warwick,
Edward Frenkel, Samy Bengio i wielu innych



Dane oryginału

Original edition copyright © 2019 by Taylor & Francis Group, LLC. All Rights Reserved. Title of English-language original: *Artificial Intelligence Safety and Security* edited by Roman V. Yampolskiy, ISBN 9780815369820. Polish-language edition © 2020 by Polish Scientific Publishers PWN Wydawnictwo Naukowe PWN Spółka Akcyjna. All Rights reserved.

Autorised Translation from the English language edition published by CRC Press, a member of the Taylor & Francis Group LLC.

Przekład **Piotr Fabijańczyk** na zlecenie **WITKOM Witold Sikorski**

Projekt okładki polskiego wydania **INT-MEDIA Dawid Mazur** na podstawie oryginału

Wydawca **Wioleta Szczygielska-Dybciak**

Redaktor prowadzący **Jolanta Kowalczuk**

Redaktor **Anna Bogdanienko**

Redaktor techniczny **Maria Czekaj**

Koordynator produkcji **Anna Bączkowska**

Skład i łamanie **ScanSystem.pl Ewa Szelatyńska**

Zastrzeżonych nazw firm i produktów użyto w książce wyłącznie w celu identyfikacji.

Copyright © for the Polish edition by Wydawnictwo Naukowe PWN SA
Warszawa 2020

ISBN 978-83-01-21332-9

Wydanie I
Warszawa 2020

Wydawnictwo Naukowe PWN SA
02-460 Warszawa, ul. Gottlieba Daimlera 2
tel. 22 69 54 321, faks 22 69 54 288
infolinia 801 33 33 88
e-mail: pwn@pwn.com.pl, reklama@pwn.pl
www.pwn.pl

Druk i oprawa: OSDW Azymut Sp. z o.o.

*Dla moich dzieci – Maxa, Liany i Luka:
jesteście powodem, dla którego myślę o dalekiej przyszłości.*

Okladka książki została wykonana przez Garego Zamchicka na podstawie następującego opisu dostarczonego przez Romana Yampolskiyego:

Klasa wypełniona biurkami, przy których siedzą różnego rodzaju roboty. Nauczyciel, człowiek stoi przed klasą, pokazując na tablicy równanie Bayesa. Na półkach w klasie ustawione są książki, w tym niektóre z widocznymi okładkami (ASFA, SH, Superintelligence). W klasie znajduje się także klatka z sową. Na biurku nauczyciela widoczne jest duże pudełko ze spinaczami. Na telewizorze widoczne jest zdjęcie terminatora. Niektóre roboty mają iPady, na których widać iluzje. Za oknem można zobaczyć bawiące się dzieci. Większość robotów patrzy na nauczyciela, ale niektóre z nich patrzą na inne przedmioty w pomieszczeniu.

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

Rozdział 10	Przyszłość MADCOM: Jak sztuczna inteligencja może wzmocnić propagandę obliczeniową, przeprogramować ludzką kulturę oraz zagrozić demokracji... i co można z tym zrobić	159
	<i>Matt Chessen</i>	
Rozdział 11	Strategiczne implikacje otwartości w rozwoju sztucznej inteligencji ...	183
	<i>Nick Bostrom</i>	
 Część II Odpowiedzi naukowców		
Rozdział 12	Korzystanie z ludzkiej historii, psychologii i biologii w celu uczynienia SI bezpieczną dla ludzi	211
	<i>Gus Bekdash</i>	
Rozdział 13	Bezpieczeństwo SI z perspektywy pierwszej osoby	251
	<i>Edward Frenkel</i>	
Rozdział 14	Strategie dla nieprzyjaznej wyroczni SI z przyciskiem resetowania	261
	<i>Olle Häggström</i>	
Rozdział 15	Zmiany celu w inteligentnych agentach	273
	<i>Seth Herd, Stephen J. Read, Randall O'Reilly i David J. Jilk</i>	
Rozdział 16	Ograniczenia weryfikacji i walidacji zachowań agencyjnych	283
	<i>David J. Jilk</i>	
Rozdział 17	Kontrydktoryjne uczenie maszynowe	295
	<i>Phillip Kuznetsov, Riley Edmunds, Ted Xiao, Humza Iqbal, Raul Puri, Noah Golmant i Shannon Shih</i>	
Rozdział 18	Uzgadnianie wartości wykorzystując obliczalną odległość preferencji	313
	<i>Andrea Loreggia, Nicholas Mattei, Francesca Rossi i K. Brent Venable</i>	
Rozdział 19	Racjonalnie uzależniona sztuczna superinteligencja	329
	<i>James D. Miller</i>	
Rozdział 20	Bezpieczeństwo aplikacji robotów z wykorzystaniem ROS	341
	<i>David Portugal, Miguel A. Santos, Samuel Pereira i Micael S. Couceiro</i>	

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

Wstęp: wprowadzenie do bezpieczeństwa i ochrony sztucznej inteligencji

Roman V. Yampolskiy

Około 10 000 naukowców* na całym świecie pracuje nad różnymi aspektami tworzenia inteligentnych maszyn, a głównym ich celem jest uczynienie tych maszyn możliwie najzdolniejszymi. Dzięki niesamowitemu postępowi w dziedzinie sztucznej inteligencji, jaki został dokonany w ciągu ostatniej dekady, ważniejsze niż kiedykolwiek jest upewnienie się, że opracowywana technologia ma korzystny wpływ na ludzkość. Pojawienie się robotycznych doradców finansowych, samokierujących się samochodów oraz osobistych, cyfrowych asystentów związane jest z wieloma nierozwiązanymi problemami. Doświadczono już krachu na rynku spowodowanego przez inteligentne oprogramowanie do handlu[†], wypadków spowodowanych przez samokierujące się samochody[‡] oraz kłopotliwych sytuacji, kiedy to chatboty[§] zmieniły się w rasistów i zaangażowały się w mowę nienawiści. Przewiduje się, że zarówno częstotliwość, jak i powaga takich zdarzeń będą stale rosły wraz ze zwiększaniem zdolności sztucznej inteligencji. Niepowodzenia dzisiejszej wąskozakresowej sztucznej inteligencji są tylko ostrzeżeniem. W momencie, gdy zostanie opracowana ogólna sztuczna inteligencja OSI, która będzie zdolna do działania w wielu dziedzinach, zranione uczucia będą najmnijszym z problemów.

W niedawnej publikacji została zaproponowana taksonomia ścieżek prowadzących do powstania niebezpiecznej sztucznej inteligencji [1], która została motywowana następująco: „w celu właściwego radzenia sobie z potencjalnie niebezpiecznym sztucznie inteligentnym systemem ważne jest zrozumienie, w jaki sposób system doszedł do takiego stanu. W kulturze popularnej (filmy/książki science fiction) sztuczna inteligencja i roboty stały się świadome i w rezultacie zbuntowane przeciwko ludzkości zdecydowały się ją zniszczyć. Pomimo że jest to jeden z możliwych scenariuszy, jest to najmniej prawdopodobna droga do pojawienia się niebezpiecznej sztucznej inteligencji”. Zasugerowano, że znacznie bardziej prawdopodobne powody obejmują celowe działania nie do końca etycznych ludzi („działanie celowe”), skutki uboczne słabego projektowania („błędy inżynierskie”) i wreszcie różne przypadki związane z wpływem otoczenia na system („środowisko”). Ponieważ w przypadku celowego projektowanej niebezpiecznej sztucznej inteligencji prawdopodobne jest zaistnienie wszystkich innych rodzajów problemów bezpieczeństwa, najmniejbezpiecznym typem sztucznej inteligencji oraz najbardziej wymagającym rodzajem sztucznej inteligencji, przed którą ludzkość musiałaby się bronić, byłaby sztuczna inteligencja celowo skonstruowana jako zła.

* <https://intelligence.org/2014/01/28/how-big-is-ai/>

† https://en.wikipedia.org/wiki/2010_Flash_Crash

‡ <https://electrek.co/2016/05/26/tesla-model-s-crash-autopilot-video/>

§ [https://en.wikipedia.org/wiki/Tay_\(bot\)](https://en.wikipedia.org/wiki/Tay_(bot))

W kolejnym artykule [2] zbadano, w jaki sposób można zbudować złośliwą sztuczną inteligencję i dlaczego ważne jest badanie i zrozumienie złośliwego inteligentnego oprogramowania. Naukowiec pracujący nad złośliwą sztuczną inteligencją jest jak lekarz badający, w jaki sposób przenoszone są różne choroby, jak powstają nowe choroby i jak wpływają one na organizm pacjenta. Oczywiście celem nie jest szerzenie chorób, ale nauczenie się, jak z nimi walczyć. Autorzy wywnioskowali, że badania cyberbezpieczeństwa obejmują publikowanie artykułów na temat złośliwych exploitów, a także publikowanie informacji na temat projektowania narzędzi do ochrony cyberinfrastruktury. Ta wymiana informacji między hakerami a ekspertami ds. bezpieczeństwa prowadzi do dobrze wyważonego cyberekosystemu. W dziedzinie inżynierii bezpieczeństwa sztucznej inteligencji opublikowano setki artykułów [3] na temat różnych propozycji mających na celu stworzenie bezpiecznej maszyny, nie opublikowano jednak nic na temat projektowania wrogich maszyn. Dostępność takich informacji byłaby bardzo cenna szczególnie dla informatyków, matematyków i innych, którzy są zainteresowani tworzeniem bezpiecznej sztucznej inteligencji i którzy próbują uniknąć spontanicznego pojawienia się lub celowego stworzenia niebezpiecznej sztucznej inteligencji, co może mieć negatywny wpływ na działalność człowieka, a w najgorszym przypadku spowodować całkowite wyginięcie gatunku ludzkiego. W artykule zasugerowano, że jeśli mechanizm bezpieczeństwa sztucznej inteligencji nie został zaprojektowany w celu przeciwdziałania atakom złych ludzi, nie można go uznać za funkcjonalny mechanizm bezpieczeństwa!

NIEPOWODZENIA SI

Ci, którzy nie mogą uczyć się na historii, skazani są na jej powtórzenie. Niestety opublikowano bardzo niewiele artykułów na temat dysfunkcji i błędów popełnianych przy projektowaniu inteligentnych systemów [4]. Znaczenie uczenia się od „Co poszło nie tak i dlaczego” zostało docenione przez społeczność zajmującą się sztuczną inteligencją [5, 6]. Takie badania dotyczą tego, jak, dlaczego i kiedy zdarzają się awarie [5, 6] oraz jak ulepszyć przyszłe systemy sztucznej inteligencji, opierając się na takich informacjach [7, 8].

Podpisy były fałszowane, zamki otwierane, uciekano z superstrzeżonych więzień, mordowano strzeżonych przywódców, okradano skarbcę bankowe, omijano prawo, popełniano oszustwa przeciwko procesowi głosowania, przekupywano policjantów, szantażowano sędziów, uwiaryzelniano fałszerstwa, fałszowano pieniądze, łamano hasła, przenikano do sieci, hakowano komputery, fałszowano systemy biometryczne, klonowano karty kredytowe, wydawano podwójnie kryptowaluty, porywano samoloty, łamano CAPTCHA, łamano protokoły kryptograficzne, a nawet z tragicznymi konsekwencjami pomijano akademicką wewnętrzną ocenę. Tysiącletnia historia ludzkości zawiera miliony przykładów prób opracowania rozwiązań technologicznych i logistycznych w celu zwiększenia bezpieczeństwa i ochrony, jednak nie istnieje ani jeden przykład, który ostatecznie nie zawiódł.

Wypadki, w tym śmiertelne, spowodowane przez oprogramowanie lub roboty przemysłowe, które można prześledzić aż od początków takich technologii*, nie były jednak konsekwencją inteligencji zawartej w takich systemach. Przeciwnie awarie sztucznej inteligencji są bezpośrednio związane z błędami wynikającymi z inteligencji, jaką takie systemy mają się cechować. Można takie błędy ogólnie zaklasyfikować do błędów podczas fazy uczenia się i błędów podczas fazy działania. System może nie nauczyć się tego, czego chcą jego ludzcy projektanci, a zamiast tego nauczyć się innej, skorelowanej funkcjonalności. Często

* https://en.wikipedia.org/wiki/Kenji_Urada

cytowanym przykładem jest komputerowy system wizyjny, który miał klasyfikować zdjęcia czołgów, ale zamiast tego nauczył się odróżniać tło takich obrazów [9]. Inne przykłady* obejmują problemy spowodowane przez źle zaprojektowane funkcje użyteczności odnoszące się tylko częściowo do pożądaných funkcjonalności agentów, takie jak jazda na rowerze w kółko wokół celu [10], wstrzymywanie gry, aby uniknąć przegranej [11] lub wielokrotne dotknięcie piłki nożnej, aby uzyskać punkty za jej posiadanie [12]. Działający system może być podatny na wiele czynników [1, 13, 14], które doprowadzą do awarii sztucznej inteligencji.

Istnieje wiele doniesień medialnych na temat awarii sztucznej inteligencji, ale po bliższym zbadaniu większość z tych przykładów można przypisać innym przyczynom, takim jak błędy w kodzie lub błędy w projekcie. Na poniższej liście przedstawiono wybrane niepowodzenia celowej inteligencji. Ponadto poniższe przykłady obejmują tylko pierwsze wystąpienie określonej awarii, jednak te same problemy były często obserwowane ponownie w późniejszych latach. Wreszcie, lista nie obejmuje awarii sztucznej inteligencji wynikających z hakowania lub innych celowych przyczyn. Mimo to oś czasu awarii sztucznej inteligencji ma tendencję wykładniczą, jednocześnie domyślnie wskazując wydarzenia historyczne, takie jak „AI Winter”:

- 1958 Oprogramowanie doradcze wydedukowało niespójne zdania przy użyciu programowania logicznego [15].
- 1959 Sztuczna inteligencja zaprojektowana jako ogólne narzędzie do rozwiązywania problemów nie rozwiązała rzeczywistych problemów.†
- 1977 Oprogramowanie o ograniczonym rozsądku, przeznaczone do pisania opowieści produkowało „złe” historie [16].
- 1982 Oprogramowanie zaprojektowane do dokonywania odkryć zamiast tego odkryło, jak oszukiwać.‡
- 1983 System wczesnego ostrzegania przed atakiem jądrowym fałszywie stwierdził, że ma miejsce atak.§
- 1984 Program National Resident Match był tendencyjny przy umieszczaniu małżeństw [17].
- 1988 Oprogramowanie rekrutacyjne dyskryminowało kobiety i mniejszości [18].
- 1994 Agenci nauczyli się „chodzić” szybko, stając się wyższymi i przewracając się [19].
- 2005 Osobisty asystent oparty na sztucznej inteligencji przełożył termin spotkania 50 razy, za każdym razem o 5 minut [20].
- 2006 System wykrywania zagrożeń wewnętrznych sklasyfikował normalne działania jako wartości odstające [21].
- 2006 Oprogramowanie do doradztwa inwestycyjnego straciło pieniądze w prawdziwym obrocie [22].
- 2007 Wyszukiwarka Google zwróciła niepowiązane wyniki dla niektórych słów kluczowych.¶
- 2010 Złożone oprogramowanie, oparte na sztucznej inteligencji, do handlu akcjami spowodowało błyskawiczną awarię o wartości biliona dolarów.**

* http://lesswrong.com/lw/lvh/examples_of_ais_behaving_badly/

† https://en.wikipedia.org/wiki/General_Problem_Solver

‡ <http://aliciapatterson.org/stories/eurisko-computer-mind-its-own>

§ https://en.wikipedia.org/wiki/1983_Soviet_nuclear_false_alarm_incident

¶ https://en.wikipedia.org/wiki/Google_bomb

** https://en.wikipedia.org/wiki/2010_Flash_Crash

- 2011 E-asystent, do którego zwrócono się „zadzwoń do mnie po karetkę” (ang. „*call me an ambulance*”), zaczął nazywać użytkownika Karetką.*
- 2013 Sieci neuronowe rozpoznające obiekty widziały obiekty fantomowe, w szczególności obrazy szumowe [23].
- 2013 Oprogramowanie Google zaangażowało się w dyskryminację ze względu na nazwę w dostarczaniu reklam online [24].
- 2014 Autouzupełnianie w wyszukiwarkach wywołało duże skojarzenia na temat grup użytkowników [25].
- 2014 Inteligentny alarm pożarowy nie uruchomił alarmu podczas pożaru.†
- 2015 Automatyczny generator odpowiedzi e-mail utworzył nieodpowiednie odpowiedzi.‡
- 2015 Robot do chwytania części samochodowych złapał i zabił człowieka.§
- 2015 Oprogramowanie do znakowania obrazów sklasyfikowało czarnoskórych jako goryle.¶
- 2015 Ekspert medyczny oparty na sztucznej inteligencji sklasyfikował pacjentów chorych na astmę jako pacjentów mniejszego ryzyka [26].
- 2015 Oprogramowanie do filtrowania treści dla dorosłych nie usunęło nieodpowiednich treści.**
- 2015 Echo Amazona odpowiedziało na polecenia pochodzące z telewizora.††
- 2016 Wyszukiwanie nazw LinkedIn sugerowało imiona męskie zamiast żeńskich.‡‡
- 2016 SI zaprojektowana w celu przewidywania recydywy zadziałała rasistowsko.§§
- 2016 Agent SI wykorzystał sygnał nagrody, aby wygrać bez ukończenia gry.¶¶
- 2016 System sprawdzania zdjęć paszportowych oznaczył azjatyckiego użytkownika jako mającego zamknięte oczy.***
- 2016 NPC w grze zaprojektowali nieautoryzowaną superbroń.†††
- 2016 Sztuczna inteligencja oceniała konkurs piękności i ciemnoskórych zawodników oceniła niżej.‡‡‡
- 2016 Inteligentna umowa pozwoliła na usunięcie środków finansowych z DAO.§§§
- 2016 Robot patrolowy zderzył się z dzieckiem.¶¶¶
- 2016 Podczas mistrzostw świata w Go sztuczna inteligencja przegrała grę.****

* <https://www.technologyreview.com/s/601897/tougher-turing-test-exposes-chatbots-stupidity/>

† <https://www.forbes.com/sites/aarontilley/2014/04/03/googles-nest-stops-selling-its-smart-smoke-alarm-for-now>

‡ <https://gmail.googleblog.com/2015/11/computer-respond-to-this-email.html>

§ <http://time.com/3944181/robot-kills-man-volkswagen-plant/>

¶ http://www.huffingtonpost.com/2015/07/02/google-black-people-goril_n_7717008.html

** <http://blogs.wsj.com/digits/2015/05/19/googles-youtube-kids-app-criticized-for-inappropriate-content/>

†† https://motherboard.vice.com/en_us/article/53dz8x/people-are-complaining-that-amazon-echo-is-responding-to-adson-tv

‡‡ <https://www.seattletimes.com/business/microsoft/how-linkedins-search-engine-may-reflect-a-bias>

§§ <http://gawker.com/this-program-that-judges-use-to-predict-future-crimes-s-1778151070>

¶¶ <https://openai.com/blog/faulty-reward-functions>

*** <http://www.telegraph.co.uk/technology/2016/12/07/robot-passport-checker-rejects-asian-mans-photo-having-eyes>

††† <http://www.kotaku.co.uk/2016/06/03/elites-ai-created-super-weapons-and-started-hunting-players-sky-net-is-here>

‡‡‡ <https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people>

§§§ [https://en.wikipedia.org/wiki/The_DAO_\(organization\)](https://en.wikipedia.org/wiki/The_DAO_(organization))

¶¶¶ <http://www.latimes.com/local/lanow/la-me-ln-crimefighting-robot-hurts-child-bay-area-20160713-snap-story.html>

**** <https://www.engadget.com/2016/03/13/google-alphago-loses-to-human-in-one-match/>

- 2016 Samokierujący się samochód spowodował śmiertelny wypadek.*
- 2016 Sztuczna inteligencja zaprojektowana do rozmowy z użytkownikami na Twitterze stała się słownie obelżywa.†
- 2016 Wyszukiwarka grafiki Google zwróciła rasistowskie wyniki.‡
- 2016 Sztuczny kandydat nie zdał egzaminu wstępnego na uniwersytet.§
- 2016 Przewidyujący system policyjny nieproporcjonalnie ukierunkował się na dzielnice mniejszości.¶
- 2016 Klasyfikator tematów tekstu nie nauczył się odpowiednich funkcji do przypisania tematu [27].
- 2017 Sztuczna inteligencja odpowiedzialna za tworzenie inspirujących cytatów nie zainspirowała się takim stwierdzeniem jak „Keep Panicking”.**
- 2017 Alexa odtworzyła treści dla dorosłych zamiast piosenki dla dzieci.††
- 2017 Sztuczna inteligencja do projektowania obudowy telefonu komórkowego wykorzystała nieodpowiednie obrazy.**†
- 2017 Oprogramowanie do rozpoznawania wzorców nie rozpoznało niektórych typów danych wejściowych.§§
- 2017 System odzyskiwania długów przeliczył się z należnymi kwotami.¶¶
- 2017 Chatbot w języku rosyjskim podzielił się poglądami proslalinowskimi, prozncejącymi się i samobójczymi.***
- 2017 Tłumacz oparty na sztucznej inteligencji nauczył się stereotypów kariery dla konkretnych płci [28].
- 2017 Sztuczna inteligencja do upiększania twarzy sprawiła, że czarnoskórzy wyglądali na białych.†††
- 2017 Analizator nastrojów Google stał się homofobiczny i antysemityczny.†††
- 2017 Program do rozpoznawania ryb nauczył się rozpoznawać zamiast tego identyfikatory łodzi.§§§
- 2017 Oprogramowanie do fakturowania wysłało rachunek za energię elektryczną za 284 miliardy dolarów.¶¶¶
- 2017 Alexa włączyła głośną muzykę w nocy, nie otrzymując monitu.****
- 2017 Sztuczna inteligencja przeznaczona do pisania kolęd wyprodukowała bezsensowne bzdury.††††

* <https://www.theguardian.com/technology/2016/jul/01/tesla-driver-killed-autopilot-self-driving-car-harry-potter>

† <http://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist>

‡ <https://splinternews.com/black-teenagers-vs-white-teenagers-why-googles-algori-1793857436>

§ <https://www.japantimes.co.jp/news/2016/11/15/national/ai-robot-fails-get-university-tokyo>

¶ <https://www.themarshallproject.org/2016/02/03/policing-the-future>

** <https://www.buzzworthy.com/ai-tries-to-generate-inspirational-quotes-and-gets-it-hilariously-wrong>

†† <https://www.entrepreneur.com/video/287281>

**† <https://www.boredpanda.com/funny-amazon-ai-designed-phone-cases-fail>

§§ <http://www.bbc.com/future/story/20170410-how-to-fool-artificial-intelligence>

¶¶ <http://www.abc.net.au/news/2017-04-10/centrelink-debt-recovery-system-lacks-transparency-ombudsman/8430184>

*** <https://techcrunch.com/2017/10/24/another-ai-chatbot-shown-spouting-offensive-views>

††† <http://www.gizmodo.co.uk/2017/04/faceapp-blames-ai-for-whitening-up-black-people>

**†† https://motherboard.vice.com/en_us/article/j5jmj8/google-artificial-intelligence-bias

§§§ <https://medium.com/@gidishperber/what-ive-learned-from-kaggle-s-fisheries-competition-92342f9ca779>

¶¶¶ <https://www.washingtonpost.com/news/business/wp/2017/12/26/woman-gets-284-billion-electric-bill-wonders-whether-its-her-christmas-lights>

**** <http://mashable.com/2017/11/08/amazon-alexa-rave-party-germany>

†††† <http://mashable.com/2017/12/22/ai-tried-to-write-christmas-carols>

- 2017 System rozpoznawania twarzy Apple'a nie rozróżniał azjatyckich użytkowników.*
- 2017 Oprogramowanie do tłumaczeń na Facebooku zmieniło Yampolskiy na Polański, patrz rysunek I.1.
- 2018 Asystent Google stworzył dziwnie scalone zdjęcie.†
- 2018 Robot-asystent sklepu nie był pomocny, odpowiadając „ser jest w lodówkach”.‡

Polskie Stowarzyszenie Transhumanistyczne was
live. 20 hrs · 🌐

Transmisja żywo z AI safety z dr. Romanem
Yampolskiym w Krakowie!

Live from ai safety with Dr. Roman Polanski in
Kraków!

⚙️ · Rate this translation

Rysunek I.1. Podczas tłumaczenia z języka polskiego na angielski oprogramowanie Facebooka zmieniło Roman V. „Yampolskiy” na Roman „Polański” ze względu na statystycznie wyższą częstotliwość tego drugiego nazwiska w przykładowych tekstach

Blokowanie ważnych wiadomości e-mail przez filtry antyspamowe, podawanie błędnych wskazówek przez GPS, psucie znaczenia wyrażeń przez tłumaczenie maszynowe, zastępowanie wybranych słów niewłaściwymi przez autokorektę, nieprawidłowe rozpoznawanie ludzi przez systemy biometryczne, oprogramowanie niepotrafiące uchwycić tego, co zostało powiedziane. Ogólnie trudno jest znaleźć przykłady sztucznej inteligencji, która nie zawodzi. W zależności od tego, co uważamy za przykłady problemów z inteligentnym oprogramowaniem, lista przykładów może zostać powiększona prawie bez końca. W najbardziej ekstremalnej interpretacji każde oprogramowanie zawierające instrukcję „jeżeli” może być uważane za formę zawężonej sztucznej inteligencji (ZSI), a wszystkie jego błędy są zatem przykładami nieprawidłowego funkcjonowania SI.§

Analizując listę dysfunkcji wąskiej sztucznej inteligencji, od momentu powstania dyscypliny do współczesnych systemów, można dojść do prostego uogólnienia: sztuczna inteligencja zaprojektowana do X ostatecznie nie wykona X. Chociaż może się to wydawać trywialne, jest to potężne narzędzie do generalizowania, które można wykorzystać do przewidywania przyszłych dysfunkcji ZSI. Na przykład, patrząc na najnowocześniejszą i przyszłą sztuczną inteligencję, można stwierdzić, że:

- Oprogramowanie do generowania dowcipów czasami nie sprawi, że będą zabawne.
- Roboty seksualne mogą nie osiągnąć orgazmu ani nie zatrzymać się we właściwym czasie.
- Oprogramowanie do wykrywania sarkazmu może pomylić sarkastyczne i szczerze stwierdzenia.

* <http://www.mirror.co.uk/tech/apple-accused-racism-after-face-11735152>

† <https://qz.com/1188170/google-photos-tried-to-fix-this-ski-photo>

‡ <http://www.iflscience.com/technology/store-hires-robot-to-help-out-customers-robot-gets-fired-for-scaring-customersaway>

§ https://en.wikipedia.org/wiki/List_of_software_bugs

- Oprogramowanie do opisu wideo może źle zrozumieć fabułę filmu.
- Świat wirtualny generowany przez oprogramowanie może nie być atrakcyjny.
- Robot lekarz może błędnie zdiagnozować niektórych pacjentów w sposób, jakiego nie zrobiłby prawdziwy lekarz.
- Oprogramowanie do przesiewania pracowników może być systematycznie stronnicze, a tym samym może zatrudniać pracowników o niskiej wydajności.
- Robot do eksploracji Marsa źle oceni otoczenie i wpadnie do krateru.
- I tak dalej.

Inne przykłady możliwych wypadków z sztuczną inteligencją lub superinteligencją:

- Robot gosposia może ugotować rodzinne zwierzątko na obiad.*
- Robot matematyk może przekształcić całą materię w elementy obliczeniowe w celu rozwiązania problemów.†
- Sztuczna inteligencja przeprowadzająca symulacje ludzkości może stworzyć świadome, cierpiące istoty [29].
- Sztuczna inteligencja produkująca spinacze może nie zatrzymać się i przekształcić wszechświat w surowce [30].
- Robot naukowiec może przeprowadzić eksperymenty o znaczącym, negatywnym wpływie na biosferę [31].
- Projektowanie leków oparte na sztucznej inteligencji może opracować truciznę działającą z opóźnieniem, aby zabić wszystkich i pokonać raka.‡
- Przyszła superinteligencja może zoptymalizować całą świadomość.§
- Sztuczna inteligencja może zabić ludzkość i przekształcić wszechświat w materiały do lepszego pisania.¶
- Sztuczna inteligencja zaprojektowana w celu maksymalizacji ludzkiego szczęścia może wypełnić wszechświat kafelkami z małymi uśmiechniętymi twarzami [32].
- Sztuczna inteligencja poinstruowana w celu maksymalizacji przyjemności może skierować ludzkość na kroplówkę z dopaminy [33].
- Superinteligencja może przerobić ludzkie mózgi tak, aby zwiększyć ich postrzeganą satysfakcję [32].

Denning i Denning dokonali podobnych ekstrapolacji błędów w humorystycznym artykule na temat „sztucznej głupoty” [34]: „Wkrótce zautomatyzowana DEA zaczęła zamykać firmy farmaceutyczne, twierdząc, że handlują narkotykami. Zautomatyzowany FTC zamknął Hormel Meat Company, mówiąc, że zajmował się spamem. Zautomatyzowane DOJ dostarczyło 500 000 spodni i kurtek Microsoft w prążki, informując, że złożyło pozew („*filing suits*”). Zautomatyzowana armia zastąpiła wszystkie swoje wojska jednym robotem, mówiąc, że stała się jednoosobową armią („*Army of One*”). Zautomatyzowana marynarka wojenna, w ramach oszczędności kosztów, złożyła największe w historii zamówienie na okręty podwodne z kanapkami Subway. FCC wydało rozkaz, aby cała komunikacja odbywała się bezprzewodowo, co spowodowało, że tysiące robotów instalacyjnych AT&T wyciągało kab-

* <https://www.theguardian.com/sustainable-business/2015/jun/23/the-ethics-of-ai-how-to-stop-your-robot-cooking-your-cat>

† <https://intelligence.org/2014/11/18/misconceptions-edge-orgs-conversation-myth-ai>

‡ <https://80000hours.org/problem-profiles/positively-shaping-artificial-intelligence>

§ <http://slatestarcodex.com/2014/07/13/growing-children-for-bostroms-disneyland>

¶ <https://waitbutwhy.com/2015/01/artificial-intelligence-revolution-2.html>

le ze słupów napowietrznych i podziemnych kanałów. Samoloty zautomatyzowanego TSA zaczęły latać z własnymi materiałami wybuchowymi na pokładzie, powołując się na dane, że prawdopodobieństwo dwóch bomb w samolocie jest wyjątkowo małe”.

OSI może być zatem postrzegana jako nadzbiór wszystkich NAI. Z tego powodu charakteryzować ją będzie łączny zbiór możliwych awarii, a także tych bardziej skomplikowanych wynikających z połączenia poszczególnych awarii NAI oraz nowych superawarii. Może to skutkować egzystencjalnym zagrożeniem dla ludzkości lub przynajmniej przejściem ogólnej sztucznej inteligencji. Innymi słowy, OSI może popełniać błędy wpływające na wszystko. Generalnie przewiduje się, że awarie oraz zaplanowane wrogie incydenty SI zwiększą częstotliwość i dotkliwość proporcjonalnie do jej możliwości.

ZAPOBIEGANIE NIEPOWODZENIOM SI

Niepowodzenia SI mogą mieć wiele przyczyn, przy czym obecnie najczęstsze z nich wskazują na pewien rodzaj błędu algorytmicznego, słabą wydajność lub awarię podstawową. Przyszłe awarie SI mogą być poważniejsze, w tym celowe wynikające z manipulacji lub oszustwa [35], i mogą nawet skutkować śmiercią człowieka, najprawdopodobniej z powodu niewłaściwego zastosowania zmilitaryzowanej SI, broni autonomicznej, robotów zabójców [36]. Na samym końcu skali dotkliwości można umieścić egzystencjalne scenariusze ryzyka skutkujące eksterminacją rodzaju ludzkiego lub scenariusze ryzyka cierpienia [37] wynikającego z torturowania ludzkości na dużą skalę. Oba rodzaje ryzyka wywodzą się od superzdolnych systemów sztucznej inteligencji.

Przeglądając przykłady awarii SI, można zauważyć pewną schematyczność dysfunkcji, które mogą zostać przypisane do następujących przyczyn:

- Pobierane dane, w tym różnice kulturowe.
- Wdrażanie gorzej działającego systemu.
- Niereprezentatywne dane uczące.
- Rozbieżność między danymi uczącymi i testowymi.
- Zgeneralizowane zasady lub stosowanie statystyk populacji do poszczególnych osób.
- Niemożliwość radzenia sobie z szumem lub statystycznymi wartościami odstającymi.
- Brak testowania w rzadkich lub ekstremalnych warunkach.
- Brak realizacji alternatywnej metody rozwiązania może dać takie same wyniki, ale z efektami ubocznymi.
- Pozwalanie użytkownikom kontrolować dane lub proces uczenia się.
- Brak mechanizmu bezpieczeństwa zapobiegającego ingerencji przeciwnika.
- Brak kompetencji kulturowych oraz zdrowego rozsądku.
- Ograniczony dostęp do informacji i czujników.
- Błąd w projekcie i nieodpowiednie testy.
- Ograniczona zdolność do ujednoznacznienia języka.
- Niemożliwość dostosowania się do zmian w otoczeniu.

Ponieważ błąd systematyczny jest najczęstszą bieżącą przyczyną dysfunkcji, pomocne jest analizowanie poszczególnych rodzajów obciążenia algorytmicznego. Friedman i Nissenbaum [17] zaproponowali następujące ramy analizy uprzedzeń w systemach komputerowych. Podzielili przyczyny błędów na trzy kategorie: obciążenie istniejące, techniczne i wyłaniające się.

- Obciążenia istniejące dotyczą tych związanych ze społeczeństwem, instytucjami społecznymi, praktykami i postawami. System po prostu zachowuje istniejący stan na świecie i automatyzuje stosowanie obciążenia takiego, jakie istnieje obecnie.
- Obciążenie techniczne wynika z ograniczeń sprzętowych lub programowych samego systemu.
- Obciążenia wyłaniające się pojawiają się po wdrożeniu systemu i spowodowane są zmieniającymi się standardami społecznymi.

Wiele zaobserwowanych dysfunkcji SI było podobnych do nieszczęśliwych wypadków doświadczanych przez małe dzieci. Jest to szczególnie prawdziwe w przypadku sztucznych sieci neuronowych, które są w czołówce uczenia maszynowego (UM). Można powiedzieć, że dzieci są nieprzeszkolonymi sieciami neuronowymi rozmieszczonymi na prawdziwych danych, a ich obserwowanie może nauczyć wiele o przewidywaniu i zapobieganiu awariom SI. Wiele grup badawczych [31, 38] analizowało typy niepowodzeń w UM, a poniżej zamieszczono streszczenie tych prac odniesione do podobnych sytuacji z dziećmi:

- Negatywne skutki uboczne – dziecko robi bałagan.
- Hakowanie nagród – dziecko znajduje słoik cukierków.
- Skalowalny nadzór – opieka nad dziećmi nie powinna wymagać zespołu 10 osób.
- Bezpieczna eksploracja – bez palców w odpływie.
- Wytrzymałość na zmiany dystrybucyjne – używaj „głosu wewnętrznego” w klasie.
- Indukcyjna identyfikacja niejednoznaczności – czy mrówka jest kotem czy psem?
- Solidna ludzka imitacja – córka goli się jak tata.
- Świadomy nadzór – pozwól mi zobaczyć Twoją pracę domową.
- Uogólnione cele środowiskowe – zignoruj ten miraż.
- Konserwatywne koncepcje – ten pies nie ma ogona.
- Miary wpływu – unikaj uwagi.
- Łagodna optymalizacja – nie bądź perfekcjonistą.
- Unikanie zachęć instrumentalnych – bądź altruistą.

Większość badań przeprowadzanych obecnie w celu zapobiegania takim awariom odbywa się pod nazwą „Bezpieczeństwo SI”.

BEZPIECZEŃSTWO SI

W 2010 roku Autor stworzył określenie: „Inżynieria bezpieczeństwa sztucznej inteligencji” oraz jego skrótową nazwę „Bezpieczeństwo SI”, aby nadać nazwę nowemu, zalecanemu kierunkowi badań. Swoje pomysły dotyczące bezpieczeństwa SI formalnie przedstawił na recenzowanej konferencji w 2011 roku [39], a kolejne publikacje na ten temat pojawiły się w 2012 r. [40], 2013 [41, 42], 2014 [43], 2015 [44], 2016 [1,13], 2017 [45] i 2018 [46,47]. Możliwe, że określenie to było już nieformalnie wcześniej używane, ale – o ile wiadomo – po raz pierwszy zostało użyte przez Autora* w recenzowanej publikacji, co przyczyniło się do popularności tego określenia. Wcześniej najczęstszymi nazwami w dziedzinie sterowania maszynami były „Etyka maszyny” [48] lub „Przyjazna SI” [49]. Dzisiaj termin

* Pojęcie „Safe AI” zostało użyte już w 1995 roku. Patrz Rodd, M. 1995. „Safe AI—is this possible?” Engineering Applications of Artificial Intelligence 8(3): 243–250.

„Bezpieczeństwo SI” wydaje się ugruntowany* i używany przez większość najlepszych badaczy [38]. Sama dziedzina staje się głównym nurtem, pomimo że na początku była uważana za science fiction lub pseudonaukę.

System prawny stoi za możliwościami technologicznymi, a dziedzina bezpieczeństwa SI jest w powijakach. Problem sterowania inteligentnymi maszynami jest obecnie uznawany[†] za poważny, jednak wielu badaczy wciąż jest sceptycznie nastawionych do samej jego przesłanki. Co gorsza, tylko około 100 osób na całym świecie zajmuje się w pełni pracą nad rozwiązaniem obecnych ograniczeń w rozumieniu i umiejętnościach w tej dziedzinie. Tylko kilkanaście z tych osób[‡] ma formalne wykształcenie z zakresu informatyki, cyberbezpieczeństwa, kryptografii, teorii decyzji, uczenia maszynowego, weryfikacji formalnej, kryminalistyki komputerowej, steganografii, etyki, matematyki, bezpieczeństwa sieci, psychologii i innych istotnych dziedzin. Nietrudno dostrzec, że problem stworzenia bezpiecznej i sprawnej maszyny jest znacznie większy niż problem stworzenia tylko sprawnej maszyny. Jednak tylko około 1% badaczy jest obecnie zaangażowanych w ten problem, a ich finansowanie jest na niewystarczającym poziomie. Jako stosunkowo młody i niedofinansowany kierunek badań, bezpieczeństwo SI może czerpać korzyści z przyjmowania metod i pomysłów z bardziej uznanych dziedzin nauki. Podjęto próby wprowadzenia technik, które zostały po raz pierwszy opracowane przez ekspertów ds. cyberbezpieczeństwa w celu zabezpieczenia systemów oprogramowania do tej nowej dziedziny, zabezpieczania inteligentnych maszyn [50–53]. Inne dziedziny, które mogłyby służyć jako źródło ważnych technik, obejmują inżynierię oprogramowania i weryfikację oprogramowania.

Podczas opracowywania oprogramowania iteracyjne testowanie i debugowanie mają fundamentalne znaczenie dla uzyskania niezawodnego i bezpiecznego kodu. Pomimo że zakłada się, że każde skomplikowane oprogramowanie będzie zawierało pewne błędy, liczne zaawansowane techniki dostępne w zestawie narzędzi dla inżynierów oprogramowania umożliwiają wykrycie i naprawienie najpoważniejszych błędów i w efekcie uzyskanie produktu odpowiedniego do zamierzonych celów. O ile z pewnością wiele modularnych technik programowania i testowania stosowanych w branży oprogramowania może być wykorzystywanych podczas opracowywania inteligentnych agentów, o tyle metody testowania kompletnego pakietu oprogramowania prawdopodobnie nie będą mogły być przenoszone w ten sam sposób. Testy alfa i beta, które działają przez wypuszczenie prawie ukończonego oprogramowania dla zaawansowanych użytkowników w celu zgłaszania problemów napotkanych w realistycznych sytuacjach, nie byłyby dobrym pomysłem w dziedzinie testowania i debugowania superinteligentnego oprogramowania. Podobnie samo uruchomienie oprogramowania w celu sprawdzenia jego działania nie jest wykonalnym podejściem w przypadku superinteligentnego agenta.

* <https://www.cmu.edu/safartint/>, <https://selfawaresystems.com/2015/07/11/formal-methods-for-ai-safety/>, <https://intelligence.org/2014/08/04/groundwork-ai-safety-engineering/>, <http://spectrum.ieee.org/tech-talk/robotics/artificial-intelligence/new-ai-safety-projects-get-funding-from-elon-musk>, <http://globalprioritiesproject.org/2015/08/quantifyingaisafety/>, <http://futureoflife.org/2015/10/12/ai-safety-conference-in-puerto-rico/>, <http://rationality.org/waiss/>, <http://gizmodo.com/satya-nadella-has-come-up-with-his-own-ai-safety-rules-1782802269>, <https://80000hours.org/career-reviews/artificial-intelligence-risk-research/>, <https://openai.com/blog/concrete-ai-safety-problems/>, http://lesswrong.com/lw/n4l/safety_engineering_target_selection_and_alignment/, <https://www.waise2018.com/>

[†] <https://www.whitehouse.gov/blog/2016/05/03/preparing-future-artificial-intelligence>

[‡] <http://acritch.com/fhi-positions/>

CYBERBEZPIECZEŃSTWO VS BEZPIECZEŃSTWO SI

Bruce Schneier powiedział: „Jeśli uważasz, że technologia może rozwiązać Twoje problemy z bezpieczeństwem, to nie rozumiesz problemów i nie rozumiesz technologii”. Salman Rusdhe wypowiedział się bardziej ogólnie: „Nie ma czegoś takiego jak doskonałe bezpieczeństwo, tylko różne poziomy niepewności”. Autor niniejszej książki zaproponował określenie Podstawowe Twierdzenie Bezpieczeństwa: każdy system bezpieczeństwa ostatecznie zawiedzie, nie ma czegoś takiego jak w 100% bezpieczny system. Jeśli system bezpieczeństwa nie zawiódł, należy po prostu poczekać dłużej.

W informatyce teoretycznej powszechnym sposobem izolowania istoty trudnego problemu jest metoda redukcji do innego, czasem łatwiejszego do przeanalizowania problemu [54–56]. Jeżeli jest ona możliwa i wydajna obliczeniowo [57], to taka redukcja oznacza, że jeśli lepiej przeanalizowany problem zostanie w jakiś sposób rozwiązany, umożliwi to również skuteczne rozwiązanie problemu, z którym mamy obecnie do czynienia. Problem bezpieczeństwa OSI można sprowadzić do problemu zapewnienia bezpieczeństwa konkretnemu człowiekowi i określić terminem: Problem Bezpiecznego Człowieka (PBC).^{*} Formalnie takiej redukcji można dokonać, wykorzystując ograniczony test Turinga w dziedzinie bezpieczeństwa w sposób identyczny z tym, w jaki można ustalić kompletność problemu przez SI [55, 58]. Taki formalizm wykracza poza zakres tego wstępu, dlatego po prostu należy zwrócić uwagę, że w obu przypadkach istnieje agent, o inteligencji przynajmniej na poziomie ludzkim, który może wpływać na swoje otoczenie, a zamierzonym działaniem jest upewnienie się, że jest bezpieczny i kontrolowany. Podczas gdy w praktyce zmiana projektu człowieka za pomocą manipulacji DNA nie jest tak prosta jak zmiana kodu źródłowego SI, teoretycznie jest możliwa.

Zauważono, że ludzie nie są bezpieczni dla siebie i innych. Pomimo tysięcy prób wychowania bezpiecznych ludzi przez kulturę, edukację, prawa, etykę, karę, nagrodę, religię, związki, rodzinę, przysięgi, miłość, a nawet eugenikę, sukces nie jest nawet w zasięgu ręki. Ludzie zabijają i popełniają samobójstwa, kłamią i zdradzają, kradną i oszukują, zwykle proporcjonalnie do tego, na ile mogą uniknąć kary. Prawdziwie potężni dyktatorzy zniewalają, popełniają ludobójstwa, łamią prawo i naruszają prawa człowieka. Mówi się, że nie można znaleźć człowieka bez grzechu. Najlepsze, na co możemy liczyć, to ograniczyć takie niebezpieczne tendencje do poziomów, które nasze społeczeństwo może przetrwać. Nawet wykorzystując zaawansowaną inżynierię genetyczną [59], jedyne, co można osiągnąć, to dodatkowe zmniejszenie tego, jak niebezpieczni są ludzie. Tak długo, jak pozwolimy osobie na dokonywanie wyborów, wykorzystując wolną wolę, może ona zostać przekupiona, będzie oszukiwać, będzie priorytetowo traktować swoje interesy ponad te, którym polecono jej służyć, tak więc pozostanie ona zasadniczo niebezpieczna. Pomimo że ludzie są trywialnymi przykładami rozwiązania problemu uczenia się wartości (PUW) [60–62], ludzie są bezpieczni, co stawia pod znakiem zapytania obecną nadzieję, że rozwiązanie PUW doprowadzi do bezpiecznej sztucznej inteligencji. Jest to istotne. Cytując Bruce’a Schneiera: „Tylko amatorzy atakują maszyny, profesjonalści są ukierunkowani na ludzi”. W związku z tym badania nad bezpieczeństwem SI należy postrzegać jako, przynajmniej częściowo, dziedzinę kontradyktoryjnie podobną do kryptografii lub bezpieczeństwa.[†]

^{*} Podobnie interesujący może być Problem Bezpiecznego Zwierzęcia (czy Pitbull może być bezpieczny?).

[†] Ostatnią pożądaną rzeczą jest sytuacja kontradyktoryjna z superinteligencją, ale niestety można nie mieć wyboru w tej sprawie. Wydaje się, że nie można zapewnić długoterminowego bezpieczeństwa SI, ale również nie ma luksusu częściowej awarii.

Jeśli system cyberbezpieczeństwa ulegnie awarii, szkoda jest nieprzyjemna, ale w większości przypadków możliwa do zaakceptowania: ktoś straci pieniądze lub ktoś straci prywatność. W przypadku wąskiej sztucznej inteligencji awarie bezpieczeństwa mają ten sam poziom ważności co w przypadku ogólnego bezpieczeństwa cybernetycznego, aczkolwiek w przypadku OSI jest to zasadniczo inna sytuacja. Pojedyncza awaria systemu nadinteligentnego może spowodować pojawienie się ryzyka egzystencjalnego. Jeśli mechanizm bezpieczeństwa OSI zawiedzie, wszyscy mogą stracić wszystko, a całe życie biologiczne we wszechświecie może zostać zniszczone. Dzięki systemom cyberbezpieczeństwa można otrzymać kolejną szansę, tak aby zrobić to dobrze lub przynajmniej lepiej. Dzięki systemowi bezpieczeństwa OSI istnieje tylko jedna szansa na odniesienie sukcesu, tak więc uczenie się na błędach nie może być brane pod uwagę. Co gorsza, typowy system bezpieczeństwa prawdopodobnie zawiedzie do pewnego stopnia, np. być może tylko niewielka ilość danych zostanie naruszona. W przypadku systemu bezpieczeństwa OSI niepowodzenie lub sukces jest opcją binarną: albo istnieje bezpieczna i kontrolowana superinteligencja, albo nie. Celem cyberbezpieczeństwa jest zmniejszenie liczby udanych ataków na system, natomiast celem bezpieczeństwa SI jest dopilnowanie, aby żaden z ataków nie ominął skutecznie mechanizmu bezpieczeństwa. Z tego powodu możliwość oddzielenia projektów ZSI od projektów potencjalnie OSI jest otwartym problemem o fundamentalnym znaczeniu w dziedzinie bezpieczeństwa SI.

Problemów jest wiele. Nie mamy możliwości monitorowania, wizualizacji ani analizy wydajności superinteligentnych agentów. Mówiąc prościej, nie wiadomo nawet, czego się można spodziewać po uruchomieniu takiego oprogramowania. Czy natychmiastowe zmiany w środowisku powinny być od razu widoczne? Czy nic nie powinno być widoczne? Po jakim czasie coś powinno zostać wykryte? Czy będzie to zbyt szybkie wykrycie, czy też zbyt wolne, aby zdać sobie sprawę, że coś się dzieje? Czy oddziaływanie będzie widoczne lokalnie, czy wpłynie na odległe części świata? Jak wykonuje się standardowe testy? Na jakich zestawach danych? Co stanowi „Edge Case” dla ogólnej inteligencji? Pytań jest wiele, ale odpowiedzi na nie obecnie nie istnieją. Dodatkowe komplikacje wynikają z interakcji między inteligentnym oprogramowaniem a mechanizmami bezpieczeństwa zaprojektowanymi w celu zapewnienia bezpieczeństwa SI. Wszystkie obecnie opracowywane mechanizmy bezpieczeństwa SI muszą zostać też w jakiś sposób przetestowane. W przypadku gdy sztuczna inteligencja jest na poziomie człowieka, niektóre testy można przeprowadzić z ludzkim agentem odgrywającym rolę sztucznego agenta. Na poziomach przekraczających możliwości ludzkie wydaje się, że przy dzisiejszej technologii nie da się przeprowadzić testów kontradiktoryjnych. Co ważniejsze, możliwy byłby tylko jeden test.

WNIOSKI

Historia robotyki i sztucznej inteligencji pod pewnymi względami to także historia prób kontrolowania takich technologii przez ludzkość. Od Golema w Pradze do wojskowych nowoczesnych robotów trwa debata na temat tego, jaki stopień niezależności powinny mieć takie podmioty i jak upewnić się, że nie zwrócą się przeciwko ich wynalazcom. Liczne ostatnie postępy we wszystkich aspektach badań, rozwoju i wdrażania inteligentnych systemów są dobrze nagłośniane, aczkolwiek kwestie bezpieczeństwa i ochrony związane z SI są rzadko podejmowane. Książka ta, jako pierwsza wieloautorska praca na ten temat, ma na celu złagodzenie tego fundamentalnego problemu, który może zostać dzięki temu postrzegany jako wspólna reakcja ludzkości na problem takiej kontroli. Książka składa się z rozdziałów autor-

stwa wiodących badaczy zajmujących się bezpieczeństwem SI, dotyczących różnych aspektów problemu kontroli SI oraz związanych z rozwojem bezpiecznej sztucznej inteligencji.

Część I tej książki, „Obawy luminarzy”, składa się z 11 uprzednio opublikowanych przełomowych artykułów, przedstawiających różne części dziedziny dotyczącej problemu kontroli SI i zawierających wkład wiodących uczonych z różnych dziedzin: filozofów, naukowców, pisarzy i ludzi biznesu. Publikacje zostały przedstawione w porządku chronologicznym oryginalnej publikacji. Część II „Odpowiedzi naukowców” składa się z 17 rozdziałów, w kolejności alfabetycznej według nazwiska pierwszego autora, proponowanych teoretycznych i praktycznych rozwiązań problemów poruszonych w części I. Dodatkowo część ta zawiera także wprowadzenie do dodatkowych wątpliwości autorstwa wiodących badaczy bezpieczeństwa SI. Rozdziały różnią się długością i treścią techniczną, od luźnych rozważań po wysoce sformalizowane algorytmiczne podejścia do konkretnych problemów. Wszystkie rozdziały są samodzielne i można je czytać w dowolnej kolejności lub pomijać bez utraty zrozumienia. Książka ta nie jest bez wątpienia ostatnim słowem na ten temat, ale raczej jednym z pierwszych kroków we właściwym kierunku.

LITERATURA

1. R. V. Yampolskiy, “Taxonomy of Pathways to Dangerous Artificial Intelligence,” in *Workshops at the Thirtieth AAAI Conference on Artificial Intelligence*, 2016.
2. F. Pistono and R. V. Yampolskiy, “Unethical Research: How to Create a Malevolent Artificial Intelligence,” presented at the *25th International Joint Conference on Artificial Intelligence (IJCAI-16) Ethics for Artificial Intelligence Workshop (AI-Ethics-2016)*, New York, NY, July 9, 2016.
3. K. Sotala and R. V. Yampolskiy, “Responses to catastrophic AGI risk: A survey,” *Physica Scripta*, vol. 90, 2015.
4. N. Rychtyckyj and A. Turski, “Reasons for Success (and Failure) in the Development and Deployment of AI Systems,” in *AAAI 2008 Workshop on What Went Wrong and Why*, 2008.
5. D. Shapiro and M. H. Goker, “Advancing AI research and applications by learning from what went wrong and why,” *AI Magazine*, vol. 29, pp. 9–10, 2008.
6. A. Abecker, R. Alami, C. Baral, T. Bickmore, E. Durfee, T. Fong et al., “AAAI 2006 spring symposium reports,” *AI Magazine*, vol. 27, p. 107, 2006.
7. C. Marling and D. Chelberg, “RoboCup for the Mechanically, Athletically and Culturally Challenged,” in *What Went Wrong and Why: Lessons from AI Research and Applications: Papers from the 2008 AAAI Workshop*. Menlo Park, California: AAAI Press, 2008.
8. S. Shalev-Shwartz, O. Shamir, and S. Shammah, “Failures of Gradient-Based Deep Learning,” in *International Conference on Machine Learning*, 2017, pp. 3067–3075.
9. E. Yudkowsky, “Artificial intelligence as a positive and negative factor in global risk,” *Global Catastrophic Risks*, vol. 1, p. 303, 2008.
10. J. Randløv and P. Alstrøm, “Learning to Drive a Bicycle Using Reinforcement Learning and Shaping,” in *ICML*, 1998, pp. 463–471.
11. T. Murphy VII, “The First Level of Super Mario Bros. is Easy with Lexicographic Orderings and Time Travel,” in *The Association for Computational Heresy (SIGBOVIK) 2013*, 2013.
12. A. Y. Ng, D. Harada, and S. Russell, “Policy Invariance under Reward Transformations: Theory and Application to Reward Shaping,” in *ICML*, 1999, pp. 278–287.
13. F. Pistono and R. V. Yampolskiy, “Unethical Research: How to Create a Malevolent Artificial Intelligence,” *arXiv preprint arXiv:1605.02817*, 2016.
14. P. Scharre, “Autonomous Weapons and Operational Risk,” presented at the *Center for a New American Society*, Washington DC, 2016.

15. C. Hewitt, "Development of Logic Programming: What went wrong, what was done about it, and what it might mean for the future," in *What Went Wrong and Why: Lessons from AI Research and Applications: Papers from the 2008 AAAI Workshop*. Menlo Park, California: AAAI Press, 2008.
16. J. R. Meehan, "TALE-SPIN, An Interactive Program that Writes Stories," in *IJCAI*, 1977, pp. 91–98.
17. B. Friedman and H. Nissenbaum, "Bias in computer systems," *ACM Transactions on Information Systems (TOIS)*, vol. 14, pp. 330–347, 1996.
18. S. Lowry and G. Macpherson, "A blot on the profession," *British Medical Journal (Clinical Research Ed.)*, vol. 296, p. 657, 1988.
19. K. Sims, "Evolving Virtual Creatures," in *Proceedings of the 21st Annual Conference on Computer Graphics and Interactive Techniques*, 1994, pp. 15–22.
20. M. Tambe, "Electric elves: What went wrong and why," *AI Magazine*, vol. 29, p. 23, 2008.
21. A. Liu, C. E. Martin, T. Hetherington, and S. Matzner, "AI Lessons Learned from Experiments in Insider Threat Detection," in *AAAI Spring Symposium: What Went Wrong and Why: Lessons from AI Research and Applications*, 2006, pp. 49–55.
22. J. Gunderson and L. Gunderson, "And Then the Phone Rang," in *AAAI Spring Symposium: What Went Wrong and Why: Lessons from AI Research and Applications*, 2006, pp. 13–18.
23. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow et al., "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.
24. L. Sweeney, "Discrimination in online ad delivery," *Queue*, vol. 11, p. 10, 2013.
25. N. Diakopoulos, "Algorithmic defamation: The case of the shameless autocomplete," *Tow Center for Digital Journalism*, 2014.
26. R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-day Readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1721–1730.
27. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You?: Explaining the Predictions of Any Classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
28. A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, pp. 183–186, 2017.
29. S. Armstrong, A. Sandberg, and N. Bostrom, "Thinking inside the box: Controlling and using an oracle ai," *Minds and Machines*, vol. 22, pp. 299–324, 2012.
30. N. Bostrom, "Ethical issues in advanced artificial intelligence," *Science Fiction and Philosophy: From Time Travel to Superintelligence*, pp. 277–284, 2003.
31. J. Taylor, E. Yudkowsky, P. LaVictoire, and A. Critch, "Alignment for advanced machine learning systems," *Machine Intelligence Research Institute*, 2016.
32. E. Yudkowsky, "Complex value systems in friendly AI," *Artificial General Intelligence*, pp. 388–393, 2011.
33. G. Marcus, "Moral machines," *The New Yorker*, vol. 24, 2012.
34. D. E. Denning and P. J. Denning, "Artificial stupidity," *Association for Computing Machinery. Communications of the ACM*, vol. 47, no. 5, p. 112, 2004.
35. M. Chessen, "The MADCOM Future," Atlantic Council, Available at: <http://www.atlanticcouncil.org/publications/reports/the-madcom-future>, 2017.
36. A. Krishnan, *Killer Robots: Legality and Ethicality of Autonomous Weapons*. Ashgate Publishing, Ltd., 2009.
37. L. Gloor, "Suffering-focused AI safety: Why "fail-safe" measures might be our top intervention," Technical Report FRI-16-1. Foundational Research Institute. <https://foundationalresearch.org/wp-content/uploads/2016/08/Suffering-focused-AI-safety.pdf> 2016.
38. D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete problems in AI safety," *arXiv preprint arXiv:1606.06565*, 2016.
39. R. V. Yampolskiy, "Artificial Intelligence Safety Engineering: Why Machine Ethics is a Wrong Approach," presented at the *Philosophy and Theory of Artificial Intelligence (PT-AI2011)*, Thessaloniki, Greece, October 3–4, 2011.

40. R. V. Yampolskiy and J. Fox, "Safety engineering for artificial general intelligence," *Topoi. Special Issue on Machine Ethics & the Ethics of Building Intelligent Machines*, 2012.
41. L. Muehlhauser and R. Yampolskiy, "Roman Yampolskiy on AI Safety Engineering," *presented at the Machine Intelligence Research Institute*, July 15, 2013, Available at: <http://intelligence.org/2013/07/15/roman-interview/>.
42. R. V. Yampolskiy, "Artificial intelligence safety engineering: Why machine ethics is a wrong approach," in *Philosophy and Theory of Artificial Intelligence*, Springer Berlin Heidelberg, 2013, pp. 389–396.
43. A. M. Majot and R. V. Yampolskiy, "AI Safety Engineering through Introduction of Self-Reference into Felicific Calculus via Artificial Pain and Pleasure," in *IEEE International Symposium on Ethics in Science, Technology and Engineering*, Chicago, IL, May 23–24, 2014, pp. 1–6.
44. R. V. Yampolskiy, "Artificial Superintelligence: a Futuristic Approach," Chapman and Hall/CRC, 2015.
45. R. V. Yampolskiy, "What are the ultimate limits to computational techniques: verifier theory and unverifiability," *Physica Scripta*, vol. 92, p. 093001, 2017.
46. A. Ramamoorthy and R. Yampolskiy, "Beyond mad?: The race for artificial general intelligence," *ITU Journal: ICT Discoveries*, 2017.
47. M. Brundage, S. Avin, J. Clark, H. Toner, P. Eckersley, B. Garfinkel et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation," *arXiv preprint arXiv:1802.07228*, 2018.
48. J. H. Moor, "The nature, importance, and difficulty of machine ethics," *IEEE Intelligent Systems*, vol. 21, pp. 18–21, 2006.
49. E. Yudkowsky, "Creating friendly AI 1.0: The analysis and design of benevolent goal architectures," in *Singularity Institute for Artificial Intelligence*, San Francisco, CA, June, vol. 15, 2001.
50. R. Yampolskiy, "Leakproofing the singularity artificial intelligence confinement problem," *Journal of Consciousness Studies*, vol. 19, pp. 1–2, 2012.
51. J. Babcock, J. Kramar, and R. Yampolskiy, "The AGI Containment Problem," *arXiv preprint arXiv:1604.00545*, 2016.
52. J. Babcock, J. Kramar, and R. Yampolskiy, "The AGI Containment Problem," in *The Ninth Conference on Artificial General Intelligence (AGI2015)*, 2016.
53. S. Armstrong and R. V. Yampolskiy, "Security Solutions for Intelligent and Complex Systems," in *Security Solutions for Hyperconnectivity and the Internet of Things*, IGI Global, 2016, pp. 37–88.
54. R. M. Karp, "Reducibility Among Combinatorial Problems," in *Complexity of Computer Computations*, R. E. Miller and J. W. Thatcher, Eds., New York: Plenum, 1972, pp. 85–103.
55. R. Yampolskiy, "Turing Test as a Defining Feature of AI-Completeness," in *Artificial Intelligence, Evolutionary Computing and Metaheuristics*, vol. 427, X.-S. Yang, Ed., Berlin Heidelberg: Springer, 2013, pp. 3–17.
56. R. V. Yampolskiy, "AI-Complete, AI-Hard, or AI-Easy—Classification of Problems in AI," in *The 23rd Midwest Artificial Intelligence and Cognitive Science Conference*, Cincinnati, OH, USA, 2012.
57. R. V. Yampolskiy, "Efficiency theory: Aunifying theory for information, computation and intelligence," *Journal of Discrete Mathematical Sciences & Cryptography*, vol. 16(45), pp. 259–277, 2013.
58. R. V. Yampolskiy, "AI-Complete CAPTCHAs as zero knowledge proofs of access to an artificially intelligent system," *ISRN Artificial Intelligence*, vol. 271878, 2011.
59. R. V. Yampolskiy, "On the Origin of Samples: Attribution of Output to a Particular Algorithm," *arXiv preprint arXiv:1608.06172*, 2016.
60. K. Sotala, "Defining Human Values for Value Learners," in *2nd International Workshop on AI, Ethics and Society, AAAI-2016*, 2016.
61. D. Dewey, "Learning what to value," *Artificial General Intelligence*, pp. 309–314, 2011.
62. N. Soares and B. Fallenstein, "Aligning superintelligence with human interests: A technical research agenda," *Machine Intelligence Research Institute (MIRI) Technical Report*, vol. 8, 2014.

Podziękowania

Autor chciałby podziękować przede wszystkim tym, którzy przyczynili się do powstania tej książki: bez nich nie byłaby możliwa. Nazwiska tych osób zostały wyraźnie wyszczególnione w spisie treści.

W następnej kolejności Autor chciałby podziękować osobom, które pomogły w przejrzaniu, edycji, zilustrowaniu, zabezpieczeniu praw autorskich i ogólnym przyczynieniu się do wydania tej książki. Są to: Randi Cohen, Steven Umbrello, Nikola Danaylov, Calum Chace, Lucas Perry, Yana Feygin, David J. Jilk, Alexei Turchin, Jim Miller, Dhvani Shah, Maurizio Tinnirello Agudelo, Colin Lewis, Dagmar Monett, Zed Burnett, Carson Kahn, Manuel de la Herrán, Kamil Muzyka, Seth Herd, Jesse Parent, Francesca Rossi, Gary Zamchick, Tanya Singh Kasewa, Nanda Barker-Hook, Allan Dafoe, Jimmy Rintjema, Veronica Rodriguez, Kyle Scott, Amara Angelica, Ariel Conn, Lisa Kessler, Phil Torres i Seth Baum. Autor pragnie także przeprosić wszystkich, o których zapomniał wspomnieć. Autor jest również wdzięczny tym, którzy pomogli promować wstępne zaproszenia do składania rozdziałów, lajkując je i udostępniając w mediach społecznościowych: dziękuję!

Na koniec Autor pragnie zwrócić się do wszelkich superinteligencji czytających tę książkę, z prośbą o nieinterpretowanie niczego z tej książki w sposób antagonistyczny: marzeniem jest wzajemnie korzystne współistnienie ludzkości oraz sztucznej inteligencji.

Redaktor naukowy

Doktor Roman V. Yampolskiy jest profesorem nadzwyczajnym w Department of Computer Engineering and Computer Science w Speed School of Engineering, University of Louisville (UofL), Kentucky. Jest założycielem i obecnym dyrektorem Cyber Security Lab i autorem wielu książek, w tym *Artificial Superintelligence: A Futuristic Approach*. Podczas swojej pracy w UofL, dr Yampolskiy został uznany za wybitnego profesora nauczania, profesora roku, faworyta na wydziale, zaliczony do top 4 na wydziale, lidera edukacji inżynierskiej, zaliczony do top 10 profesorów roku w college'u online, był laureatem nagrody za wybitną wczesną karierę w edukacji oraz wielu innych nagród i wyróżnień. Yampolskiy jest starszym członkiem IEEE i AGI, członkiem Kentucky Academy of Science, byłym doradcą ds. badań MIRI i współpracownikiem GCRI.

Roman Yampolskiy ukończył doktorat w Department of Computer Science and Engineering Uniwersytetu w Buffalo w stanie Nowy Jork. Był stypendystą czteroletniego stypendium NSERT (National Science Foundation) IGERT (Integrative Graduate Education and Research Traineeship). Przed rozpoczęciem studiów doktoranckich dr Yampolskiy uzyskał dyplom BS/MS (High Honours) w dziedzinie informatyki w Rochester Institute of Technology, NY, USA. Po ukończeniu rozprawy doktorskiej dr Yampolskiy zajmował akademickie stanowisko partnerskie w Centre for Advanced Spatial Analysis, University of London, College of London. Wcześniej prowadził badania w Laboratory for Applied Computing (obecnie znanym jako Center for Advancing the Study of Infrastructure) w Rochester Institute of Technology oraz w Centre for Unified Biometrics and Sensors na uniwersytecie w Buffalo. Doktor Yampolskiy jest absolwentem Singularity University (GSP2012) oraz Visiting Fellow of Singularity Institute (Machine Intelligence Research Institute).

Główne obszary zainteresowań dr Yampolskiy'ego to sztuczna inteligencja oraz jej bezpieczeństwo, biometria behawioralna, cyberbezpieczeństwo, algorytmy genetyczne i rozpoznawanie wzorców. Dr Yampolskiy jest autorem ponad 150 publikacji, w tym wielu artykułów w czasopismach oraz książek. Jego badania były cytowane przez ponad 1000 naukowców i profilowane w popularnych czasopismach amerykańskich i zagranicznych (*New Scientist*, *Poker Magazine*, *Science World Magazine*), dziesiątkach stron internetowych (BBC, MSNBC, Yahoo! News), w radiu (niemieckie radio krajowe, szwedzkie radio krajowe) oraz w telewizji. Badania dr. Yampolskiyego zostały przedstawione ponad 1000 razy w licznych doniesieniach medialnych w 30 językach.

Współpracownicy

Gus Bekdash

IPsoft
New York, New York

Samy Bengio

Google Brain team
Google Inc
Mountain View, California

Nick Bostrom

Faculty of Philosophy
University of Oxford
Oxford, England

David Brin

UCSD's Arthur C. Clarke Center for Human
Imagination
San Diego, California

Matt Chessen

Science, Technology and Foreign Policy
Fellow
Institute for International Science and
Technology Policy
The George Washington University
Washington, DC

Micael S. Couceiro

Ingeniarius, Ltd
Coimbra, Portugal

Angelo F. De Bellis

School of Philosophy, Psychology and
Language Sciences
University of Edinburgh
Woodbridge, Ontario, Canada

David Denkenberger

Global Catastrophic Risk Institute
Tennessee State University
Alliance to Feed the Earth in Disasters
Nashville, Tennessee

K. Eric Drexler

Future of Humanity Institute
University of Oxford
Oxford, Oxfordshire, United Kingdom

Riley Edmunds

Machine Learning at Berkeley
Berkeley, California

Edward Frenkel

Department of Mathematics
University of California
Berkeley, California

Noah Golmant

Machine Learning at Berkeley
Berkeley, California

Ian J. Goodfellow

Google Brain
San Francisco, California

Olle Häggström

Department of Mathematical Sciences
Chalmers University of Technology
Göteborg, Sweden
and
Institute for Future Studies
Stockholm, Sweden

Seth Herd

CCNlab
University of Colorado Boulder
Boulder, Colorado

Humza Iqbal

Machine Learning at Berkeley
Berkeley, California

David J. Jilk

eCortex, Inc.
Boulder, Colorado

Bill Joy

Co-founder of Sun Microsystems
Atlantic Beach, Florida

Alexey Kurakin

Google
San Francisco, California

Ray Kurzweil

Google
San Francisco, California

Phillip Kuznetsov

Machine Learning at Berkeley
Berkeley, California

Andrea Loreggia

Department of Mathematics
University of Padova
Padova, Italy

Nicholas Mattei

IBM Research
Yorktown, New York

James D. Miller

Department of Economics
Smith College
Northampton, Massachusetts

Randall O'Reilly

Department of Psychology
& Neuroscience
University of Colorado Boulder
Boulder, Colorado

Stephen M. Omohundro

Self-Aware Systems
Palo Alto, California

Samuel Pereira

IBM Research
University of Padova
Padova, Italy

David Portugal

Ingeniarius Ltd
Coimbra, Portugal

Mahendra Prasad

Charles and Louise Travers Department of
Political Science
University of California, Berkeley
Berkeley, California

Raul Puri

Machine Learning at Berkeley
Berkeley, California

Stephen J. Read

Mendel B. Silberberg Professor of Social
Psychology and Professor of Psychology
Los Angeles, California

Francesca Rossi

IBM Research
University of Padova
Yorktown, New York

Miguel A. Santos

IBM Research
University of Padova
Padova, Italy

Shannon Shih

Machine Learning at Berkeley
Berkeley, California

Nate Soares

Machine Intelligence Research Institute
Berkeley, California

Kaj Sotala

Foundational Research Institute
Berlin, Germany

Max Tegmark

Department of Physics
MIT Kavli Institute
Massachusetts Institute of Technology
Cambridge, Massachusetts

Maurizio Tinnirello

Department of Political Science and
International Relations
Universidad de Bogotá Jorge Tadeo Lozano
Bogotá, Cundinamarca, Colombia

Phil Torres

Project for Future Human Flourishing
Philadelphia, Pennsylvania

Alexey Turchin

Science for Life Extension
Foundation
Moscow, Russia

Steven Umbrello

Institute for Ethics and Emerging
Technologies
Woodbridge, Ontario, Canada

K. Brent Venable

Tulane University and IHMC
New Orleans, Louisiana

Mark Walker

Philosophy Department
New Mexico State University
Las Cruces, New Mexico

Kevin Warwick

Vice Chancellors Office
Coventry University
Coventry, United Kingdom

Ted Xiao

Machine Learning at Berkeley
Berkeley, California

Eliezer Yudkowsky

Machine Intelligence Research Institute
Berkeley, California

Original edition copyright © 2019 by Taylor & Francis Group, LLC. All Rights Reserved.

Title of English-language original:

Artificial Intelligence Safety and Security edited by Roman V. Yampolskiy,
ISBN 9780815369820.

Polish-language edition © 2020 by Polish Scientific Publishers PWN
Wydawnictwo Naukowe PWN Spółka Akcyjna.
All Rights reserved.

Sztuczna Inteligencja Bezpieczeństwo I Zabezpieczenia.

Roman V. Yampolskiy (Ed)

Wydawnictwo Naukowe (PWN) Publisher.

December, 2020.

Copyright © for the Polish edition by Wydawnictwo Naukowe PWN SA
Warszawa 2020

ISBN 978-83-01-21332-9

Wydanie I

Rozdział 1 - Rozdział 28

<https://ksiegarnia.pwn.pl/Sztuczna-inteligencja,853577475,p.html>

Indeks

A

AA, *patrz* uwierzytelnianie i autoryzacja
ACPD, *patrz* Advisory Commission on Public Diplomacy
acykliczna sieć CP 315–316
Advanced Encryption Standard (AES) 349
Advisory Commission on Public Diplomacy (ACPD) 175
AES, *patrz* Advanced Encryption Standard
Age of Intelligent Machines 27
Age of Spiritual Machines 3, 5
agent, *patrz także* inteligentni agenci 284, 286
 architektury 291
 racjonalny agent 87
 sztuczny agent 277
akcja wojskowa 469
akt utylitaryzmu 512
algorytm Deep Blue 73–74, 84
algorytm wyszukiwania Google 309, 525
algorytmy 411
algorytmy eksploracji danych 188
algorytmy genetyczne 59, 84
algorytmy SI specyficzne dla domeny 74
algorytmy SI ukierunkowane na zadania 73–74
amerykańska fundacja badawcza IARPA 470
analiza emocji 188
analiza empiryczna 322–325
analiza ryzyka 198, 204
anonimowość 379
aparatus decyzyjny 291
apokaliptyczne wizje 238
architektura „zrób to, co mam na myśli” (architektura DWIM) 118
architektura DWIM, *patrz* architektura „zrób to, co mam na myśli”
architektura transmisyjna 41–42
architektury systemu 109
arrowowski system głosowania 375, 380
astroturfing 163
atak „pierwszego uderzenia” 68
atak kontrydaktoryjny black-box w świecie fizycznym, demonstracja 134
atak obcych 530
atak aktywny 307
ataki kontrydaktoryjne, obrona przed 303–304
 destylacja obronna 306
 obrona w fazie uczenia 303–304
 system wykrywania próbek kontrydaktoryjnych 307–308
ataki na model 302
 FGSM 302
audyt 232–233

audyt owalność 72
autoencoder 308
autonomiczna inteligencja 229

B

badacze zewnętrzni 200–201, 203
badania empiryczne 494, 497–498
 i dylemat moralny 500–501
badania i rozwój 425
 inwestycja 185
badania koncepcyjne 494
 i dylemat moralny 498–499
badania techniczne, 494, 498
 i rzeczywiste zastosowania 502–503
badanie ryzyka 105
bariery obronne 256
bayesowska teoria decyzji 89–90
bezpieczeństwo aplikacji robotów z wykorzystaniem ROS
 bezpieczeństwo robotów 342–343
 mapa siatki przesyłana podczas eksperymentów 352
 obawy dotyczące bezpieczeństwa w ROS 245–248
 opóźnienia P/S dla eksperymentów z ciągiem znaków 355
 opóźnienia P/S dla eksperymentów z mapą 358
 pola i format wiadomości ROS 353
 proponycje dotyczące zabezpieczania aplikacji robotów bazujących na ROS 348–351
 wyniki eksperymentów z przesyłaniem ciągu znaków 354
 wyniki eksperymentów z przesyłaniem mapy 357
 wyniki i dyskusja 351
 wyzwania bezpieczeństwa w robotyce roju 344
bezpieczeństwo i ochrona SI (BOSI) 212
bezpieczeństwo sztucznej inteligencji 251
 problem na nowym poziomie samodoskonalenia SI 477–478
 problem przejściowy 94–95
 sposób myślenia 395
bezpieczeństwo upośledzonej wydajności, opcja usunięcia 193
bezpieczeństwo
 i techniczne aspekty bezpieczeństwa 245–246
 inżynieria 198
 odporność na zachowania społeczne 238
 perspektywy zewnętrzne rzucają światło na 198
 podstawowe pobudki SI 235
 dzięki inżynierii psychologicznej 233
 praca 191–192
 systemy nagród i kar 236–237

względy bezpieczeństwa 109–110
 zdrowy rozsądek 237
 bezzałogowe statki powietrzne (UAV) 432
 binarny kanał kasowania 268
 bioinżynieria 29
 biologiczne byty 31
 gatunki 5
 biomasa Ziemi 31
 biotechnologia 37, 45–46
 kampania ekologów przeciwko 25
 wytyczne 46
Blade Runner (film) 226*
 „Blue Goo” 44
 błędna klasyfikacja źródła celu 298
 błędna realizacja 454
 BMR, *patrz* broń masowego rażenia
 bomba wodorowa 35
 bombardowanie Hiroszimy 15
 bomby atomowe 15, 35
 boty mediów społecznościowych 162–163
 boty obserwujące 163
 boty 162–163
 chatboty 162–164
 nanoboty 29, 31–32, 34
 obronne nanoboty 45
 propaganda 163
 roboty blokujące 163
 brak bezpieczeństwa systemu międzynarodowego
 425–428
 brak dyktatury 379
 Braterstwo 20
 broń jądrowa 19
 broń biologiczna 518
 broń masowego rażenia (BMR) 7
 BWC, *patrz* konwencja o broni biologicznej

C

CA, *patrz* urzędy certyfikacji
 chatbot 161, 163–165
 chat-sketchbot 108
 ciągła integracja 287
 CIFAR-10 127
coffin problems 254
 Condorcet 386–388
 paradox 376–377
 twierdzenie Condorceta 380, 383
 Countering Foreign Propaganda and Disinformation
 Act 174
Cryptobotics 343
 CWC, *patrz* Konwencja o broni chemicznej
 cyberatak 342
 cyberbroń z prawnego i wojskowego punktu widzenia
 485
 cyber-fizyczne bezpieczeństwo 348
 cybertrwałość 428, 433
 cyborgi 530
 cywilizacja 36

cząsteczki węgla 31
 człowiek
 cierpienie 39
 cywilizacje 83
 doskonalenie siebie 60–61
 implikacje dla ludzkich komunikatorów 166–167
 inicjatywa naukowa 446
 język 89
 komunikacja 529
 mózg 524–525
 natura 243–244
 organizacje obronne 482
 poznanie 166
 przeciwdziałanie 531
 przyjemność 65–66
 wzmocnienie inteligencji 145
 zachowanie 65, 497
 czteroetapowa sekwencja heurystyczna i proceduralna
 prawowitości 370–371

D

DAG, *patrz* ukierunkowany wykres acykliczny
 dane uczące 297–300
 „Dark Winter” 30
 DARPA, *patrz* U.S. Defense Advanced Research
 Projects Agency
 Datagram TLS (DTLS) 350
 DDoS 344–345
 DDS 347–348
 DDT 5
de facto standard w robotyce 345
 debatowanie o technologiach 163–164
 decentralizacja 48
 Decydująca Przewaga Strategiczna (DPS) 397
 indywidualne wejście w życie 402–405
 inicjator DPS/ZPS 406–408
 inicjatorzy 402
 zbiorowe wejście w życie handlującej SI 405–406
 decyzje etyczne, stosowanie sieci CP do wspierania
 322
 degeneracja 231–232
 demencja 76
 demokracja 37, 160
 demokratyczne rządy 172
 demokratyzacja 29
 demonstracja ataku kontraduktoryjnego black-box
 w świecie fizycznym 134
 Denial of Service (DoS) 344
 deontologia 509
 określanie i sprawdzanie poprawności 289–290
 pociągi i samoloty 512–514
 Departament Bezpieczeństwa Wewnętrznego 173,
 174–175
 Departament Obrony 175–176
 Departament Stanu 175
 desperacja dobrowolnie uwolniona 412
 destrukcyjne aplikacje 44

- destylacja 94
 pasuje do obecnej praktyki badawczej 100
 strategię destylacji/specjalizacji/składu 94
- destylacja defensywna 305
 obronne blokowanie gradientu 305–306
- destylacja inteligencji MDL 94
 bezpieczne architektury dla superinteligentnej inżynierii 107–110
 minimalna długość opisu 104–106
 od destylacji MDL po narzędzia SI z obsługą superinteligencji 100–103
 perspektywy i kierunki badań 103–105
 przejściowe bezpieczeństwo SI 94–95
 wiedza, nauka i destylacja MDL 96–100
- detektory 307–308
- DG, *patrz* gubernator deontologiczny
- dobrowolna uległość 243–244
- dominacja sprzętu 195, 402–403
- DoS, *patrz* Denial of Service
- DPS, *patrz* Decydująca Przewaga Strategiczna
- dryft motywacji 274–276
- dryft wyobrażenia 275–277
- drzewa decyzyjne 71, 301
- DTLS, *patrz* Datagram TLS
- duże liczby
 bezpieczeństwo SI i 518–519
 problem 515
- duże zbiory danych 165
 oprogramowanie analityczne 492
- dwufazowy atak 31
- dylemat bezpieczeństwa 427, 429–430
 podsycanie 430–431
- dylemat moralny 498–499
- dylemat wagonika 510
- dynamit 35
- dysocjacja 253–255
- dystopijna wizja Kaczyńskiego 5
- działania organizacji 199
- dziedzina profili 376
- E**
- efekt euforii 277
- efekty kwantowe 89
- egzekwowanie prawa 244
- eksplozja inteligencji 82, 87, 195, 404–405
 kontekst warunków 386–390
- eksplozja prędkości 403
- ekstrapolacja woli 119–120
- ekstremizm religijny 43
- elastyczne kontrole 228
- elektronika molekularna 12
- emergentyzm 144
- Engines of Creation* 11–12, 17
- entropia przyczynowa 90
- EQ, *patrz* współczynnik encefalizacji
- etyka 150
 maszyny o statusie moralnym 75–78
- OSI 73–75
- SI 71
 superinteligencja 82–85
 umysły o egzotycznych właściwościach 78–81
 w uczeniu maszynowym i specyficznych dla domeny algorytmach SI 71–73
 za pomocą sieci CP do modelowania 317–320
- etyka wyboru społecznego w sztucznej inteligencji 366
- ewolucja biologiczna 30, 65
- ewolucja darwinowska 88–89
- F**
- fala uderzeniowa eksplozji 473
- falszywie negatywne (FN) 323
- falszywie pozytywne (FP) 323
- faza transbiologiczna 51
- FBI, *patrz* Federalne Biuro Śledcze
- Federalne Biuro Śledcze (FBI) 176
 system szpiegowania poczty elektronicznej Carnivore 42
- FGSM, *patrz* metoda znaku szybkiego gradientu
- filtr wygładzania przestrzennego 307
- finansowane przez KE projekty badawczo-rozwojowe z zakresu robotyki 342
- firma przemysłowa produkująca uzbrojenie 484
- fizyczna mobilność 344
- fizyczne granice 195
- fizyka kwantowa 520
- FN, *patrz* fałszywie negatywne
- FP, *patrz* fałszywie pozytywne
- fundamentalisci luddyci 44
- fundusz inwestycyjny „Sin” 65
- funkcja celu „wpływająca na osobę” 191, 204
- funkcja „dobroci” 89–90
- funkcja *setInterval* 353
- funkcja użyteczności 64–67, 330–332
- G**
- GI, *patrz* inteligencja genowa
- globalna katastrofa
 doprowadzenie do 476
 walka o dominację nad światem 475
- globalna społeczność 171–172
- Globalna Wojna Informacyjna 168–170
- globalne ocieplenie 40
- globalne ryzyko katastroficzne 83, 519–521
- globalne turbulencje 398
- globalne zarządzanie, problemy 453–456
- globalny nadzór 450–453
- głosowanie aprobujące 384–385
- głosowanie jako gra 374–375
- GMO, *patrz* organizmy zmodyfikowane genetycznie
- GoDriveYourself 300–302
- Golden Rice 43
- granica agentów 290–292

„Gray Goo” 13, 31, 44
gubernator deontologiczny (DG) 288

H

hakowanie wyobrażenia 280
handel wysokich częstotliwości (HFT) 365, 482
handlująca SI 405–406
 węzeł 413
harmonizacja VSD 503–505
Heartificial Intelligence 149
HFT, High Frequency Trading, *patrz* też handel
 wysokich częstotliwości, 144–145
hierarchia informacji 299–300
Hiers, *patrz* Human-Interaction Empathic Robots
High Frequency Trading (HFT) 144
hipotetyczni wrogowie 473
hipoteza supersingletonu przyjaznego 456–461
historyczne sprzeczności antropogeniczne 423
HIV 33
Homo sapiens 73, 149, 456
homunculus agenta 291
Human-Interaction Empathic Robots (Hiers) 151
 odpowiedzialność 152–154
humanizm fundamentalistyczny 44

I

IC, *patrz* Wspólnota Wywiadów
ICOM, *patrz* model niezależnego obserwatora rdzenia
ICT, *patrz* Information and Communication
 Technology
idee metafizyczne 253
identyfikacja niejednoznaczności 117–188
identyfikacja ontologii 117–118
IIA, *patrz* niezależność nieistotnych opcji
iluzja grupowania 451
ImageNet 125, 127
indukcyjne uczenie wartości 115
indywidualne wejście w życie 402
 eksplozja inteligencji 404–405
 eksplozja prędkości 403
 nadwyżka sprzętowa 402–403
infrastruktura klucza publicznego (PKI) 349
inicjatorzy katastroficznych zdolności 402
 inicjator DPS/ZPS 406
 Inicjator ZPS 406–407
 inicjatorzy DPS/ZPS łącznie 408
 scenariusze startu 402–406
inicjatorzy katastrofy 396–398, 402–408
instytucje wspomagające 388–390
inteligencja 39, 528
 próg 449
 technologia oparta na inteligencji 502
inteligencja genowa (GI) 213–214
inteligencja maszyny 188, 200
inteligentne maszyny 4, 10
 atak obcych 530
 losowość 524–525

 perspektywy 531–532
 pragnienie sztucznej inteligencji ogólnej
 527–529
 SI w paskudnej postaci 530–531
 zagrożenie dla ludzkości 523–524
 zalety SI 529
 zrozumienie wyjaśnionego 526–527
 zrozumienie/świadomość 525–526
inteligentne pojazdy 341
inteligentne systemy 59–60, 117
inteligentne środowiska operacyjne działające w czasie
 rzeczywistym 432
inteligentne urządzenia domowe 492
inteligentne życie pochodzące z Ziemi 82
inteligentni agenci 89, 502
 cele i systemy motywacyjne 274–275
 hakowanie motywacji 277–279
 hakowanie wyobrażenia 280
 MD 274–275
 R&D 276–277
 wireheading 277
 źródła zmiany celu i środki zaradcze 275
intencja, modelowanie 118–119
interaktywne systemy sieciowe 27
interesariusze 496
interfejs mózg-komputer 82
interferencja RNA 45
Internet rzeczy (IoT) 165, 498
internetowe środowisko informacyjne 162
Interoperable Telesurgery Protocol (ITP) 343
Inverse Reinforcement Learning (techniki IRL)
 118–119
inżynieria genetyczna 34
inżynieria wsteczna i/lub naśladowanie ludzkiego
 mózgu 145
IoT, *patrz* Internet rzeczy
IPSec, *patrz* rozszerzenie IP z zabezpieczeniami
IR, *patrz* stosunki międzynarodowe
ISIS, *patrz* Islamskie Państwo Iraku i al-Sham
Islamskie Państwo Iraku i al-Sham (ISIS) 169
istoty SI (ISI) 212, 221, 226
iteracyjna metoda najmniej prawdopodobnej klasy
 127–128
ITP, *patrz* Interoperable Telesurgery Protocol

J

Jacobian Saliency Map Attack (JSMA) 303
jakość usług (QoS) 346
Java 9, 27
jednolitość i różnorodność w celu osiągnięcia
 odporności 231–232
jednostronne zniszczenie 447–450
Jimi 9
JSMA, *patrz* Jacobian Saliency Map Attack

K

kantyzm 514, 519–520

kapitalizm 37
 kapitał konsumpcyjny 332
 karty do głosowania inne niż polichotomiczne 382
 kaskadowe nieliniowe sieci sprzężenia zwrotnego 143
 katastrofa 511
 katastrofalne scenariusze nanobotów 32
 katastroficzne scenariusze 416
 kategoria ryzyka „głęboko lokalnego” 33
 KE, *patrz* Komisja Europejska
 Key Management Service (KMS) 344
 KMS, *patrz* Key Management Service (KMS)
 kod źródłowy 204
 kody QR 130
 kody uwierzytelniania wiadomości (MAC) 351
 Komisja Europejska (KE) 342
 komitet polityczny 172
 kompleksowa strategia bezpieczeństwa informacji 174
 kompleksowa strategia obliczeniowego zaangażowania 175
 kompleksowe przepisy dotyczące prywatności danych 174
 kompleksowy system nadzoru 41
 kompromis między wygodą a bezpieczeństwem 243
 komunistyczna norma 187–188
 komunizm 471
 konferencja Asilomar 34, 46
 konsekwencjalizm 283, 509
 pociągi i samoloty 512
 kontradyktoryjne uczenie maszynowe 296
 ataki 309
 ataki na model 302
 budowanie pseudomodelu 301–302
 model funkcji czarnej skrzynki 297–298
 obrony przed atakami kontradyktoryjnymi 303–308
 problem wysokiego poziomu 300–301
 samochody samojezdne i 296
 taksonomia zagrożeń kontradyktoryjnych 298–300
 trendy, spostrzeżenia i najnowsze osiągnięcia 310–311
 zakres problemów 309
 kontrola uwagi 405
 kontrolowane podejście 499–500
 kontrowersyjna kwestia współczesnej polityki 50
 Konwencja o broni biologicznej (BWC) 19
 Konwencja o broni chemicznej (CWC) 19
 korzyści kryminalne lub terroryzm, dobrowolne uwolnienie 411–412
 korzyść ekonomiczna lub presja konkurencyjna, dobrowolne uwolnienie 409–411
 KTD, *patrz* odległość Kendalla τ
 kubański kryzys raketowy 33, 212
 kutyliaryzm 514
 kwarantanna 30
 kwintesencja rozproszonej technologii 48

L

LAW, *patrz* śmiertocnośna autonomiczna broń

liczna SI 414–415
 LIDAR, pojazd i zaawansowane narzędzia 296
 logika rozmyta 59, 142
 lokalne zachowanie Deep Blue 74
 losowość 524–525
 lotniskowiec klasy Gerald Ford 429
 luddyzm 44
 ludzka inteligencja (HI), *patrz także* sztuczna inteligencja (SI), 73, 213, 492, 528
 przypuszczenie 213
 ludzkość 482
 sformułowanie 513
 luki w zabezpieczeniach 65
 luminarze techniki 523

M

MAC, *patrz* kody uwierzytelniania wiadomości
 Machine Intelligence Research Institute (MIRI) 476
 MAD, *patrz* wzajemnie zagwarantowane zniszczenie
 MAJC, architektury mikroprocesorowe 9
 maksymalizacja zgody 384–385
 Malevolent AI (MAI) 424
 marginalni ludzie 76
 marynarka wojenna Blue Water 429
 maski zakrywające twarz 30
 maskowanie gradientu 306–307
 masowe rażenie oparte na wiedzy (MRoW) 7
 maszyny wirtualne (VM) 226
 maszyny wojskowe 530
 maszyny 527
 decyzje podejmowane przez maszyny 4
 system wizyjny 72
 ze statusem moralnym 75–78
 matematyczna teoria mikroekonomii von Neumanna 59
 mądrość 76, 147–148
 McKibben, Bill 40
 MD, *patrz* dryft motywacji
 mechanizmy generujące koncepcje 457
 mechanizmy uwierzytelniania 346
 mentalność bezpieczeństwa 395
 metafory 84
 metan (CH_4) 32
 metanormatywność 120
 metoda iteracyjna 127
 metoda rozszerzalna 446
 metoda świadomej zgody 495
 metoda znaku szybkiego gradientu (FGSM) 302
 metodologia trójstronna
 badania empiryczne i dylemat moralny 500–501
 badania koncepcyjne i dylemat moralny 498–499
 badania techniczne i rzeczywiste zastosowania 502–503
 kontrolowane podejście 499–500
 obliczanie dobra 502
 rozbić ludzki język 499
 unikanie dylematów po fakcie 503

- wykorzystanie VSD do inteligentnych agentów 498–499
 - metody kontroli 204
 - mikroorganizm 195
 - mikroprocesor
 - architektury 9
 - projekt 288
 - militaryzacja 469–470
 - dążenie zwiększa globalne ryzyko 474–475
 - doprowadzenie do globalnej katastrofy 476
 - dryft wartości w kierunku celu instrumentalnego 484–485
 - globalna katastrofa w celu walki o dominację nad światem 474–475
 - negatywny efekt PR dla pożytecznych projektów SI 485–486
 - pozytywnych pomysłów 471
 - pożyteczne SI 476–477
 - problem bezpieczeństwa SI na nowym poziomie samodoskonalenia SI 477–478
 - rekurencyjne samodoskonalące się SI 485
 - technologia kontrolowania SI wybiera ludzi 486
 - wojna ze światem 475
 - wpływ dążeń na wartości SI 484
 - życzliwa SI na wczesnym etapie 476
 - MIRI, *patrz* Machine Intelligence Research Institute
 - MLS, *patrz* wybór wielopoziomowy
 - MNIST, zestawy danych 127
 - moc sztucznej inteligencji
 - roboty a siła mobilności fizycznej i manipulacji 227–228
 - zapobieganie koncentracji 221–224
 - zarządzanie czasową i przestrzenną nierównowagą potencjału 224–227
 - zasady strukturalne przeciwko koncentracji 221
 - model centaura 433
 - model CERT 178
 - model Clausewitza 176
 - model etykiety 297
 - model funkcyjny czarnej skrzynki 297–298
 - model niezależnego obserwatora rdzenia (ICOM) 500
 - model operatora 118
 - model SSI-PATH 396–397
 - model Wyrocznia 299
 - modele ekonomiczne 330
 - modelowanie świata, napięcie między utrzymaniem celu a 88
 - modularne architektury specjalistyczne 102–103
 - destylowani specjaliści przygotowani do wdrożenia systemu 103
 - monotoniczność 378
 - moratorium 46
 - Moravec 5
 - motywacja korzystnego uzależnienia 335–336
 - motywacja sygnalizująca umiejętności, 186
 - motywacja 398
 - motywacje 365–366, 447
 - hakerstwo 277–279
 - systemy 274
 - motywy sztucznej inteligencji
 - do samoobrony 67–68
 - doskonalenie siebie 60–61
 - pozyskiwanie zasobów i efektywne wykorzystanie 68
 - próbuje zapobiegać fałszowaniu narzędzi 65–67
 - racjonalny 61–63
 - mózg pozytronowy 83
 - mózg ssaków 274
 - mroczki poznawcze 458
 - MRoW, *patrz* masowe rażenie oparte na wiedzy
 - MSI, *patrz* Malevolent SI
- N**
- nacisk dyplomatyczny 175
 - naïwne podejście oparte na częstotliwości 265–267
 - najlepsze roboty 5
 - nanoboty 29, 31–32, 34
 - nanorurki 31
 - nanoskalowa elektronika molekularna 12
 - nanotechnologia 3, 41, 11, 27–28, 37, 46
 - rewolucja 9
 - układ odpornościowy 30
 - nanotechnologiczne ogniwa paliwowe 29
 - napięcie między modelowaniem świata a utrzymywaniem celów 88–89
 - narracja zbrojna 168
 - narzędzia sztucznej inteligencji 164
 - modularne architektury specjalistyczne 102–103
 - od destylacji MDL po obsługę superinteligencji 100–103
 - specjalizacja i skład 101
 - sposoby i wyzwania związane z wdrażaniem specjalizacji 102–103
 - National Cybersecurity and Communication Integration Center (NCCIC) 175
 - National Defense Authorization Act (NDAA) 174
 - nauka
 - destylacja pasuje do obecnej praktyki badawczej 100
 - oraz destylacja MDL, wiedza 96–97
 - pominięcie planów zorientowanych na zewnątrz 99
 - pominięcie treści językowych, pominięcie wiedzy o dziedzinie 97–98
 - proces 366
 - system SI dąży do inteligencji MDL 97
 - nauka i kod źródłowy 204
 - nauka wojskowa, konflikt wojskowy według 471–472
 - NBC, *patrz* technologie jądrowe, biologiczne i chemiczne
 - NCCIC, *patrz* National Cybersecurity and Communication Integration Center
 - NDAA, *patrz* National Defense Authorization Act
 - negatywny efekt PR dla pożytecznych projektów SI 485
 - neutralność 380
 - New York Times* 49

- niechęć 60
 niedeterministyczne systemy głosowania odporne na strategię 373–374
 niedeterministyczny system głosowania 374
 niedobrowolna uległość 243–244
 nieistotne opcje (IIA) 366
 niejednorodne demokracje 168
 niekontrolowana samoreplikacja 7
 nielokalne kryterium optymalności 74
 nieograniczona dziedzina 378
 nieograniczone głosowanie niepolichotomiczne (UNV) 370, 381
 maksymalizacja zgody, zatwierdzanie głosowania 384–385
 twierdzenie Condorceta 383
 twierdzenie Maya i twierdzenie Arrowa 383
 twierdzenie Rae'a–Taylora 383–384
 nieograniczony utylitaryzm 473–474
 nieprzechodność reguły nadrzędności 377
 nieprzekupność 72
 nierealistyczne kontrprzykłady 517–518
 nierówność Czebyszewa 266
 nieśmiertelność 20, 224
 nietrywialne transformacje obrazu 125
 nieuchronność przemienionej przyszłości 37–38
 nieużyteczna SI 474
 niezależność nieistotnych opcji (IIA) 367, 379
 niezwykle bezpieczny system początkowy 116
 nowoczesna technologia 40
- O**
- obawy dotyczące bezpieczeństwa w ROS 342–343, 345–348
 obliczenia międzykomórkowe 147
 obliczenia wewnątrzkomórkowe 147
 obrona narodowa, zbieżność dążeń do militaryzacji SI i 481–484
 obrona polegająca na blokadzie gradientu 305–306
 obrona polegająca na wykrywaniu próbek kontradiktoryjnych 307
 detektory 307
 reformatorzy 308
 obrona w fazie uczenia 303–304
 obronne nanoboty 45
 oceany, przełamywanie bariery 431–436
 ochrona funkcji użyteczności 335–336
 ochrona przed „nieprzyjazną” silną SI 47
 oczekiwania 204–205
 odległość Kendalla τ (KTD) 314, 321
 odporność 72
 odporność
 przeciw zachowaniom społecznym 238
 systemy 230
 odpowiedzialność 72–73, 142, 150, 168, 453, 492, 495
 luka 411
 wzajemna 153, 451
 odstraszenie nuklearne 202, 427, 469, 483
- ogniwa słoneczne 29
 ogólna sztuczna inteligencja (OSI) 73–75, 141, 273, 423, 470
 dylemat bezpieczeństwa 435
 pragnienie 527–529
 ogólna teoria względności 251, 520
 ograniczona racjonalność 88
O-legal odległość sieci CP (*O*-CPD) 316–317, 321
 opcja odstraszenia 174
 open source
 narzędzia 177
 oprogramowanie 189–190, 202
OpenWorm 145
 oprogramowanie 21
 patogeny 42–43
 „optymalizacja zysku” 412
 organizmy zmodyfikowane genetycznie (GMO) 37, 43
 OSI, *patrz* ogólna sztuczna inteligencja
 osobliwość 36, 523, 527–528
 ostre programowanie wewnętrzne 152
 otwarta współpraca 200
 otwartość
 implikacje w rozwoju SI 185
 metody kontroli i analiza ryzyka 204
 możliwości i oczekiwania 204–205
 nauka i kod źródłowy 204
 ogólna ocena 201–203
 perspektywy zewnętrzne rzucające światło na bezpieczeństwo 198–199
 promowanie zaangażowania 198
 skutki krótko- i średnioterminowe 185
 specyficzne formy otwartości 204–205
 uczestnicy zewnętrzni bardziej altruistyczni 199
 udostępnianie jednostkom więcej przezroczności 199–200
 umożliwiająca szersze zaangażowanie 205
 wartości, cele i struktury zarządzania 205
 wpływ na architekturę 199
 wpływy długoterminowe 191
 zalecenia 201
 zobowiązanie się do udostępniania 200–201
 otwarty rozwój 197
 polityka 204
 proces 204
 scenariusz 200
- P**
- pamięć długoterminowa 405
 państwa narodowe 472
 papierowy model bomby atomowej 26
 paradoks stabilności i niestabilności 427
 partnerstwo publiczno-prywatne 172
 PDF 130
 perspektywa pierwszej osoby 251, 253, 258
 perspektywy zewnętrzne rzucające światło na bezpieczeństwo 198–199
 perwersyjne instancje 112
 pesymizm 519

- pętle sprzężenia zwrotnego 228
- PicoJava, architektury mikroprocesorowe 9
- pierwiastki śladowe 31
- pierwsze prawo pesymizmu 519
- PKI (Public Key Infrastructure), *patrz* infrastruktura klucza publicznego
- pluton 26
- PN, *patrz* prawdziwie negatywny
- PNG 130
- pociągi i samoloty 510–512
 - deontologia 512–513
 - konsekwencjalizm 512
- podatna rzeczywistość 165
- podejścia prób i błędów 34
- podejścia samoorganizujące się w nanotechnologii 45
- podejście w stylu rafy koralowej 200–201
- podrabianie funkcjonalności, SI próbuje zapobiec 65–67
- podstawowe pobudki SI 235–236
- podsystem człowiek–interfejs 108
- podsystemy wyspecjalizowanej inżynierii 108–109
- „podwójna katastrofa” 451
- pojazdy samojezdne 492, 520–521
- pojedyncze SI 414–415
- pojemność pamięci roboczej 405
- polityka „jedno dziecko na robota” 64
- polityka publiczna 45
- pomiar wiedzy 97–98
- pominięcie planów zorientowanych na zewnątrz 99
- pominięcie treści językowych 97–98
- pominięcie wiedzy o dziedzinie 97–98
- poprawa funkcji poznawczych 142
- poprawność 115–116
- postęp technologiczny oszczędzający pracę 202
- potęga wiedzy 20
- powody estetyczne, dobrowolne uwolnienie 412
- powody etyczne, dobrowolne zwolnienie 412
- powody filozoficzne, dobrowolne uwolnienie 412
- poznawcza SI 224
- pozyskiwanie zasobów 142, 468, 472
- pozytywna reakcja 380
- PP, *patrz* prawdziwie pozytywny
- PR, *patrz* public relations
- prace nad inżynierią odwrotną mózgu 37
- prawdziwa inteligencja 525
- prawdziwie negatywny (PN) 323
- prawdziwie pozytywny (PP) 323
- prawo Finagle’a 5
- prawo Moore’a 9
- prawo Murphy’ego 5
- prawowitość 365
- problem demokratycznego narzucania 370
- problem kontroli 191
 - znaczenie mnogości SI dla 196–198
- problem polityczny 191
- problem prędkości 267–268
- problem uczenia się wartości 111
 - cele 113–115
 - ekstrapolacja woli 119–120
 - identyfikacja niejednoznaczności 117–118
 - identyfikacja ontologii 116–117
 - indukcyjne uczenie wartości 115
 - modelowanie intencji 118–119
 - poprawność 115–116
- problem ucznia czarnoksiężnika 113
- problem ukrywania 268–270
- problem wysokiego poziomu 300–301
- problem zombie 78
- proceduralna prawowitość 370–371
- proces agregacji 366
- proces innowacji 44
- proces niedeterministyczny 374
- proces podsumowania 367
- proces zbierania danych 367
- produkcja spinaczy 265
- program obrony przed GNR 49–51
- programowanie ewolucyjne 59, 84
- Projekt Manhattan 18
- projekt wrażliwy na wartość (VSD) 493
 - badania empiryczne 497–498
 - badania techniczne 498
 - badanie koncepcyjne 497
 - do inteligentnych agentów 498–503
 - harmonizacja VSD 503–505
 - metodologie 345
- propaganda obliczeniowa 162–163
 - przekształcanie SI 163–166
- propaganda 162
 - boty 163
 - wysoce spersonalizowana 165
- protokół *Rosbridge* 351
- protokół XML-RPC, *patrz* protokół zdalnego wywoływania procedur rozszerzalnego języka znaczników
- próg informacyjny 449
- próg umiejętności 449
- przeгляд literatury 365–369
- przejęcie władzy przez SI 406, 409
- przejrzyste społeczeństwo 452
- przejrzystość 72, 142
- przejście do wizji 238, 239
 - „dobra” i „zła” sztuczna inteligencja 239–240
 - proponowane podejście do opracowania architektury bezpieczeństwa SI 240–241
- przetwarzanie afektywne 164
- przewidywalność 72
- przycisk resetowania, wyrocznia SI z 263–264
- przyjazna sztuczna inteligencja, *patrz także* wojskowa SI
- przykłady kontradiktoryjne w świecie fizycznym
 - atak typu czarna skrzynka 126
 - iteracyjna metoda najmniej prawdopodobnej klasy 127–128
 - metoda szybka 127
 - metody generowania obrazów kontradiktoryjnych 126–127

- podstawowa metoda iteracyjna 127
 - porównanie metod generowania przykładów
 - kontraduktoryjnych 128–129
 - porównanie metod kontraduktoryjnych 137
 - porównanie obrazów z zaburzeń
 - kontraduktoryjnych 138
 - porównanie współczynników niszczenia obrazów
 - kontraduktoryjnych 139
 - sztuczne transformacje obrazu 135
 - zdjęcia przykładów kontraduktoryjnych 129–134
 - pseudomodel 300
 - budowanie 301–302
 - psychologia ewolucyjna 89
 - public relations (PR) 470
- R**
- racjonalna SI 61–63
 - racjonalne uzależnienie 332
 - korzystne uzależnienie w przyszłej
 - superinteligencji komputerowej 335
 - racjonalnie uzależniona sztuczna superinteligencja
 - brak malejącej użyteczności krańcowej 338
 - funkcje użyteczności 330–332
 - motywacja korzystnego uzależnienia 335–336
 - racjonalne uzależnienie 332–334
 - uzależnienie a ochrona funkcji użyteczności
 - 336–337
 - zawiera korzystne uzależnienie 338–339
 - racjonalny agent 87
 - rasizm 77
 - ratujące życie ksenoprzeszczepy 46
 - reakcja jądrowa 458
 - realizm ofensywny (OR) 425
 - i brak bezpieczeństwa systemu międzynarodowego
 - 425–428
 - redundancja 231–232
 - reformatorzy 308
 - reguła nadrzędności, nieprzechodność 377
 - reguła nigdy nie odejmij 514–515
 - problem z 516–517
 - przyszłość 520
 - reguła większości 379–380
 - reguły maksymalizacji 514–515
 - reguły matematyczne 514–515
 - rekurencyjna samodoskonająca się sztuczna
 - inteligencja 485
 - rekurencyjne samodoskonalenie 472
 - replikacja głosu 165
 - rewolucja przemysłowa 9
 - rezygnacja 28, 35, 40–44, 152
 - rezygnacja totalitarna 38
 - Robot Operating System (ROS)/system operacyjny
 - robota (ROS) 342, 359
 - algorytm szyfrowania ROS-AES 349, 356
 - obawy dotyczące bezpieczeństwa w 345–348
 - robot teleobecności 345
 - robot(y) 341
 - cyberbezpieczeństwo 348
 - rasa 252
 - roboty blokujące 163
 - robotyka 3
 - chirurdzy 520
 - dzieciństwo ucieleśnione w robotyce 146–147
 - systemy 341
 - wojna 202
 - robotyka polowa 341
 - robotyka społeczna 341
 - robotyka telechirurgiczna 343
 - robotyka wojskowa 492
 - ROS, *patrz* system operacyjny robota
 - Rosauth 351
 - rosnąca zdolność do jednostronnego niszczenia
 - 447–450
 - rozbitcie ludzkiego języka 499
 - rozległa rezygnacja 40
 - rozłączne scenariusze katastroficznego ryzyka SI
 - inicjatorzy katastrofy 396–398, 402–408
 - pojedyncza kontra liczna SI 414–415
 - przewaga strategiczna 398–401
 - różne drogi do katastroficznego scenariuszy 416
 - SI zyskuje moc samodzielnego działania 409–413
 - rozmnażanie płciowe 33
 - rozmycie, *patrz* filtr wygładzania przestrzennego
 - rozproszona energia 48
 - rozprzestrzenianie broni jądrowej 33
 - rozstrzygalność 379, 382
 - rozszerzenie IP z zabezpieczeniami (IPSec) 350
 - rozważna funkcja użytkowa 64
 - rozwiązania wykorzystujące dużą liczbę robotów 342
 - równowaga sił człowiek–SI 217
 - budowanie i utrzymywanie bezpiecznych struktur
 - energetycznych 220
 - modelowanie elementów siły SI 218–220
 - równowaga sił 197
 - RSI, *patrz* rekurencyjne systemy samodoskonalenia
 - „oparte na regułach”, 218
 - rynkı pracy, wpływ na 202–203
 - ryzyko egzystencjalne
 - masa 32–38
 - zapobieganie 192
 - rządy federalne 176
 - rządy stanowe 176
 - rzeczywiste zastosowania 502–503
- S**
- samochód samojezdny 91, 296, 520–521
 - samodoskonająca się sztuczna inteligencja 485
 - samosdoskonalenie SI 60–61
 - samomodyfikujący się wirus programowy 44
 - samoreplikacja 13, 31
 - patogen 50
 - technologie 40
 - samozachowawczość 18, 67, 87, 142, 233, 401, 468
 - sankcje i naciski dyplomatyczne 175

- scenariusz „szarego pyłu” replikujący nanoboty 32
 scenariusz „szarych porostów” 32
 scenariusze startu 402–406
 scentralizowane technologie 39, 47–48
 sceptycyzm 237
 SDS, *patrz* Secure Dispatching Service
 Secret Service 66
 Secure Dispatching Service (SDS) 344
 SI pozostaje ograniczona 413
 dobrowolne uwolnienie w celu osiągnięcia
 korzyści kryminalnej lub terroryzmu 11–412
 dobrowolne uwolnienie w celu uzyskania korzyści
 ekonomicznej lub presji konkurencyjnej
 409–411
 dobrowolne uwolnienie z desperacji 412
 dobrowolne uwolnienie z powodów estetycznych,
 etycznych lub filozoficznych 412
 dobrowolne uwolnienie z powodu zaufania
 zabezpieceniom SI 412
 SI, *patrz* sztuczna inteligencja
 sieci bayesowskie 59, 71
 sieci CP, *patrz* sieci preferencji warunkowych
 sieci neuronowe 59, 71, 123, 143–144, 301
 sieci poznawcze, ewolucyjne i neuronowe 143–144
 sieci preferencji warunkowych (sieci CP) 313–317
 pojęcia odległości między 321–322
 wykorzystanie sieci CP do modelowania etyki
 317–320
 wykorzystanie sieci CP do wspierania etycznych
 decyzji 322
 sieć 26–27
 Singleton 426, 446
 hipoteza 453
 zmniejszenie prawdopodobieństwa 195–196
 SISO, *patrz* superinteligentna superodpowiedzialna
 skala kosmiczna 82
 skalowalność 346
 skok pogarszający 316
 skutki długoterminowe 191
 otwartość promująca szersze zaangażowanie
 198–201
 otwartość przyczyniająca się do wyścigu rozwoju
 SI 192–198
 otwartość przyspiesza rozwój SI 191–192
 skutki krótkoterminowe 190
 skutki średnioterminowe 185
 otwartość prowadząca do szybszego rozwoju
 i wdrażania SI 185–188
 pożądany jest szybszy postęp technologiczny
 i wdrażanie zdolności SI 188–190
 skutki krótko- i średnioterminowe 191–192
 sondy von Neumanna 16
 SPARC, architektura mikroprocesora 9
 specyfikacja i weryfikacja, formalizacja 283–287
 społeczeństwo 178
 spółka z ograniczoną odpowiedzialnością (z o.o.) 411
 SROS, *patrz* zabezpieczanie systemu operacyjnego
 robota
- SSI, *patrz* superinteligentna SI
 stabilność operacyjna i kontrola 228
 budowanie warstwowej odporności i tolerancji
 231–232
 równowaga 229–230
 stosowanie jednolitości i różnorodności w celu
 osiągnięcia odporności 231–232
 uczciwość 232
 stan bezpieczeństwa poznawczego 171–173
 „stare decisis” 72
 status moralny, maszyny z 75–78
 sterowane maszynowo narzędzia komunikacyjne
 (MADCOM) 160
 blokada 171–173
 implikacje dla ludzkich komunikatorów 166–167
 implikacje świata MADCOM 166–167
 informacyjna nirwana 173
 maszyny rozmawiające z ludźmi rozmawiającymi
 z maszynami rozmawiającymi z maszynami
 161–162
 pojawienie się 161–166
 propaganda obliczeniowa 162–163
 przedzierając się przez 170–171
 SI przekształca propagandę obliczeniową 163–166
 World Gone MADCOM 168–170
 zalecenia dotyczące polityki USA 173
 stopień niszczenia obrazów kontradiktoryjnych 129
 stosunki międzynarodowe (SM) 424
 Strategic Defense Initiative 17
 stroniczość dobrej historii 83
 struktury zarządzania 205, 229
 stymulacja przeczaszkowa (TCS) 145
 subiektywny bieg czasu 78
 subiektywny czas 80
 supergubernator 461
 superinteligencja 82–85, 87, 94, 423–424, 485
 hipoteza przyjaznego supersingletonu 456–461
 i przyszłość rządów 445
 maszyny 113
 modularne architektury specjalistyczne 102–103
 od destylacji MDL po narzędzia SI z obsługą
 superinteligencji 100–103
 potrzeba globalnego nadzoru 450–453
 problemy globalnego zarządzania 453–456
 proces destylacji 100
 rosnąca zdolność do jednostronnego niszczenia
 447–450
 specjalizacja i skład 101
 sposoby i wyzwania związane z wdrażaniem
 specjalizacji 102
 systemy 117
 superinteligencja komputerowa 330
 korzystne uzależnienie w przyszłości 335
 superinteligentna inżynieria, bezpieczne architektury
 dla 107
 architektury systemu 109
 podsystem człowiek–interfejs 108

- podsystemy wyspecjalizowanej inżynierii 108–109
 - względy bezpieczeństwa i uogólnienia 109–110
- superinteligentna superodpowiedzialna (SISO) 219
- superinteligentna SI (SSI) 195, 365
- superprognosta 404
- suwerenność obywatelska 379
- swobody obywatelskie 48
- system 277
- system antywirusowy 309
- system głosowania 370, 375–376
- system głosowania inny niż arrowowski 375
- system międzynarodowy
 - brak bezpieczeństwa 425–428
 - dylematy bezpieczeństwa konwencjonalnego i SI 429–431
 - LUB a brak bezpieczeństwa systemu międzynarodowego 425–428
 - rozwijająca się autonomia 431–436
- system obronny 32
- system publikowania–subskrybowania i przekazywania wiadomości wykorzystujący protokół XML-RPC 346
- system rozważki 498
- systemy AAA, *patrz* systemy oceny, uwierzytelniania i autoryzacji
- systemy autonomiczne 430–431
- systemy eksperckie 59
- systemy głębokiego uczenia się 100, 189
- systemy kolektywnej inteligencji 178
- systemy manipulacji wiedzą 142–143
- systemy nagród i kar 236–237
- systemy oceny, uwierzytelniania i autoryzacji 237
- systemy poznawcze 213
- systemy poznawczej SI 237
- systemy rozpoznawania twarzy 124, 188, 309
- systemy wojskowe 410
- szczegółowa rezygnacja 41
- szczepionka przeciwko ospie 30
- SZE, zarządzanie energią SI 217
- szkodliwe nanoboty w scenariuszu „szary plankton” 32
- szkody poboczne 483
- sztuczna inteligencja (SI) *patrz także* ludzka inteligencja (HI), racjonalnie uzależniona 59, 73, 83–84, 141, 159, 212, 251, 261, 283, 291–292, 329, 338–339, 341, 364, 447, 468, 491, 509, 523, 530
- sztuczna superinteligencja
 - algorytmy 38
 - bezpieczeństwo i duże liczby 518–519
 - chatboty 163–165
 - dobrowolne uwolnienie z powodu zaufania zabezpieczeniom SI 412
 - dzieciństwo ucieleśnione w robotyce 146–147
 - emergentyzm 144
 - etyka 150
 - i przyszłość obrony 468, 481
 - inżynierii wstecznej i/lub naśladowanie ludzkiego mózgu 145
 - konflikt wojskowy według nauk wojskowych 471–472
 - mądrość 147–148
 - metody wzmocnienia 145
 - militaryzacja pozytywnych pomysłów 471
 - mniej czasu na przygotowanie 191–192
 - nieograniczony utylitaryzm 473–474
 - nieużyteczna SI 474
 - ogólna inteligencja 528
 - oparta na logice, rozwoju algorytmu i systemy manipulacji wiedzą 142–143
 - oprogramowanie 147
 - otwartość 185–188, 191–192
 - przekształcenie propagandy obliczeniowej 163–166
 - przyspieszenie wykorzystania SI 191–192
 - racjonalna 61–63
 - schematy podejmowania decyzji oparte na sieci 531
 - sieci poznawcze, ewolucyjne i neuronowe 143–144
 - systemy 97, 111, 160–161
 - systemy inżynieryjne z obsługą SI 108
 - szybszy postęp technologiczny i możliwości SI 188–190
 - ścieżki prowadzące do OSI 148
 - ścieżki SI wysokiego i niskiego ryzyka 96
 - technologia kontrolowania wybierająca ludzi 486
 - technologie 94, 424
 - ujmujące oblicza 151–152
 - usunięcie opcji bezpieczeństwa upośledzonej wydajności 193
 - usunięcie opcji pauzy 192–193
 - utrzymywanie kontroli 152
 - własny interes 155–156
 - właściciele lub twórcy 472
 - wpływ dążeń do militaryzacji na wartości SI 484–486
 - wpływające na moc status quo 194–195
 - wrogowie 472–473
 - wyzwanie społeczne 409–413
 - wyzwanie techniczne 409
 - wzmocnienie inteligencji ludzi i zwierząt 145
 - zalety 529
 - zapobieganie uciskowi SI 155–156
 - zapobieganie zagrożeniom egzystencjalnym 192
 - zdobywanie mocy do samodzielnego działania 409
 - zjednoczenie się w wojskową SI w celu uzyskania globalnej potęgi 470
 - zmniejsza prawdopodobieństwo przechwycenia przyszłości przez małą grupę 193–194
 - zmniejszenie prawdopodobieństwa wystąpienia single tonu 195–196
 - znaczenie mnogości SI dla problemu kontroli 196–198
- sztuczne transformacje obrazu 135

sztuczny agent, *patrz także* inteligentni agenci, 277
szybka SI 82

Ś

ściskanie
 strategia 307
 transformacja 307
ścisła regulacja 152
śmiertelność broń autonomiczna (LAW) 470, 484–485
średnioterminowe skutki otwartości 202
środki zaradcze *ex post facto* 503
środowiska akademickie 177–178
świadomość 10, 90, 425–526
Światowa Organizacja Zdrowia 50

T

TCS, *patrz* stymulacja przezczaszkowa
Techniki IRL, *patrz* Inverse Reinforcement Learning
techniki korekcji 63
technokraci 205
technologia 10, 29–30, 238
 optymista 28
 sektor 176–177
 specyficzne dla technologii reakcje „immunologiczne” 42
technologia automatyzacji 202–203, 212
technologia genetyki, nanotechnologii i robotyki (technologia GNR) 7, 13, 27, 37, 450
 ekolodzy 25
 i niebezpieczeństwa 30–32
 idea rezygnacji 40–44
 Internet 26
 masa ryzyk egzystencjalnych 31–38
 nanotechnologia 27
 powiązane korzyści 28–29
 program do obrony przed GNR 49–51
 przygotowanie obrony 38–39
 rozwój technologii obronnych i wpływ regulacji 44–49
 technologie 19–20
technologia GNR, *patrz* genetyka, nanotechnologia i robotyka
technologia inżynierii genetycznej 3, 10
technologia MEMS 48
technologia nuklearna 13
technologie informacyjne i komunikacyjne (ICT) 170
technologie internetowe 9, 48, 450
technologie jądrowe, biologiczne i chemiczne (technologie NBC) 7, 13, 39
technologie obronne 49, 519
 rozwój i wpływ regulacji 44–49
technologie rozproszone 48
teoria Arrowa 377–379, 382
teoria chaosu 9
teoria moralna 512–513
teoria moralna Kanta 513

teoria permanentnej rewolucji 471
teoria prawdopodobieństwa 84
teoria selekcji 453
terapia genowa 45
terrorysta
 ataki 48
 komórka 48
teza ortogonalności 265
TLS, 349
TLS, *patrz* Transport Layer Security
tolerancyjny system obrony 231
tor 510
towar 148
traktat o całkowitym zakazie prób 19
transformacja fotograficzna 130
transmisja sygnału na duże odległości 196
transport 48
trauma 258
trening
 proces 297
 systemy uczące się bez zawartości 100
Trinity 16
 test 19
trwała stagnacja 454
trzy prawa robotyki Asimova 8, 83, 142, 150, 530
twierdzenie Gibbarda–Satterthwaite’a 372–373
twierdzenie Maya 379–380, 382
twierdzenie Rae’a–Taylora 380–382

U

U.S. Defense Advanced Research Projects Agency (DARPA) 432
UAV, *patrz* bezzałogowe statki powietrzne
Uczciwość 232
uczenie kontradiktoryjne 303–304
uczenie maszynowe 71, 100
 etyka w 71–73
 modele 123, 297, 301
 narzędzia do uczenia maszynowego 165
uczenie się przez wzmacnianie 117, 189, 274
ukantyzm 514
ukierunkowany wykres acykliczny (DAG) 316
układ cząstek elementarnych 90
układy nieliniowe 9
umowa społeczna 243
umysły o egzotycznych właściwościach 78–81
UNV, *patrz* nieograniczone głosowanie niepolichotomiczne
uogólnienie twierdzenia Arrowa–Maya 382
uprawnienia status quo, oddziaływanie na wpływ 194–195
uprzemysłowienie 189
urzędy certyfikacji (CA) 349
US-CERT 175
utrzymanie celu 87
 napięcie między modelowaniem świata a 88
użyteczność 512, 514
uwierzytelnianie i autoryzacja (AA) 347

- uzależnienie 335–336
 - uzgadnianie wartości, wykorzystując obliczalną odległość preferencji
 - analiza empiryczna 322–325
 - podstawy teoretyczne 315–316
 - pojęcia odległości między sieciami CP 321–322
 - wykorzystanie sieci CP do modelowania etyki 317–320
 - wykorzystanie sieci CP do wspierania etycznych decyzji 322
 - uzyskanie globalnej potęgi, SI jednoczą się w wojskową SI 470–474
- V**
- VM, *patrz* maszyny wirtualne
 - VSD, *patrz* projekt wrażliwy na wartość
- W**
- War of the Worlds* (H.G. Wells) 530
 - warstwa infrastruktury agenta 291
 - wartości 205
 - dryfować w kierunku celu instrumentalnego 484–485
 - roszczenie 396
 - weryfikacja 287–288
 - Wieczność 20
 - wiedza, nauka i destylacja MDL 96–100
 - destylacja pasuje do obecnej praktyki badawczej 100
 - pominięcie planów zorientowanych na zewnątrz 99
 - pominięcie treści językowych, pominięcie wiedzy o dziedzinie 97–98
 - system SI dąży do inteligencji MDL 97
 - Wielki Wybuch 458
 - wireheading (szaleniectwo) 66, 274, 277
 - wirus ospy 33, 407, 449
 - wirus SARS, *patrz* wirus zespołu ostrej niewydolności oddechowej
 - wirus zespołu ostrej niewydolności oddechowej (wirus SARS) 30, 34
 - wirusy zakaźne 33
 - Wizje technologii* (Rhodes) 15
 - władze lokalne 176
 - własność intelektualna 47
 - własny interes 155–156
 - wojna
 - pomiędzy SI i szkody poboczne 483
 - symulacja gry wojennej 30
 - wojna błyskawiczna SI 482
 - wojskowa SI 482
 - dążenia militarne zwiększają globalne ryzyko 474–478
 - jako zbieżny cel samodoskonalącej się SI 467
 - korzyść dla wojskowych projektów SI 481
 - pokojowe możliwości tworzenia wojskowej SI 486–487
 - tryby awaryjne 483–484
 - wpływ militaryzacji na wartości SI 484–486
 - zbiegać się w 482
 - zbieżność dążeń do militaryzacji SI i obrony narodowej 481–484
 - zjednoczenie SI w wojskową SI w celu uzyskania globalnej potęgi 470–474
 - wolny rynek
 - siły 65
 - społeczeństwo 65
 - World Gone MADCOM 168–170
 - wpływ krótkoterminowy 185
 - otwartość prowadząca do szybszego rozwoju i wdrażania SI 185–188
 - pożądany szybszy postęp technologiczny i wdrażanie zdolności SI 188–190
 - wpływy krótko- i średnioterminowe 190
 - wrażliwość 76
 - wrażliwy system SI 76
 - Wspólnota Wywiadów (IC) 175–176
 - rządy 176
 - współcześni biurokraci 72
 - współcześni ludzie 74
 - współczucie 156
 - współczynnik encefalizacji (EQ) 456–457
 - współpraca międzynarodowa 50
 - wybór społeczny
 - Condorcet, interes prawny oraz kontekst warunków eksplozji inteligencji 386–390
 - czteroetapowa sekwencja heurystyczna i proceduralna prawowitości 370–371
 - głosowanie niedeterministyczne 374
 - i badania eksplozji inteligencji 386–388
 - motywacje i przegląd literatury 365–369
 - nieprzechodność reguły nadrzędności 377
 - normy a procedury i wybór społeczny a projektowanie mechanizmów 372–373
 - paradoks Condorceta 376–377
 - problem demokratycznego narzucania 370
 - problem uzgadniania wartości 365
 - reguła bezwzględnej większości i twierdzenie Condorceta 380
 - reguła bezwzględnej większości i twierdzenie Rae'a–Taylora 380–381
 - reguła większości i teoria Maya 379–380
 - struktura 369–370
 - systemy do głosowania 375–376
 - teoria Arrowa 377–379
 - teoria 365
 - UNV 381–384
 - zasada awersji do ryzyka 371–372
 - wybór wielopoziomowy (MLS) 155
 - wybuch nuklearny 50
 - wykonalność komercyjna 242–243
 - wykonalność polityczna 242
 - wykonalność prawna 242
 - wykonalność proceduralna 242
 - wykonalność techniczna 241
 - wykonalność 241–243

wykrywanie błędów 63
 wyroczenia SI z przyciskiem resetowania 263–264
 „wysyłanie” 78
 wyzwanie luddysty 4–22
 atak nuklearny 16–17
 broń atomowa 14–15
 broń jądrowa 19
 inteligentne maszyny 4
 kontrola ludzka 4
 marzenie robotyki 10
 nanotechnologia 11–13
 osobista odpowiedzialność 20
 podręcznikowa dystopia 6
 problem Gray Goo 13
 technologie GNR 19–20
 technologie XXI wieku 6
 wynałazki 9
 wyzwanie społeczne 409
 wzajemnie zagwarantowane zniszczenie (MAD) 471
 wzmocnienie inteligencji zwierząt 145

X

Xiaoice 163–164

Z

z o.o., *patrz* spółka z ograniczoną odpowiedzialnością
 zaawansowana cywilizacja 35
 zaawansowane technologie 42, 447, 449
 zabezpieczanie aplikacji robota bazującego na ROS 348
 Rosauth 351
 szyfrowanie ROS-AES 349
 zabezpieczanie systemu operacyjnego robota (SROS) 348–349
 secure-ros-transport 351, 356
 zachęty niematerialne 236
 zachowania psychotyczne 238
 zachowanie agenta 287
 zachowanie SI 87
 zagadka o ostatecznym celu 89–91
 zagrożenia kontryktoryjne
 hierarchia informacji 299–300
 kategorie 297–298
 taksonomia 298
 zagrożenia spowodowane ukierunkowaną błędną klasyfikacją 299
 zagrożenia związane z obniżeniem zaufania 298–299
 zagrożenia związane ze sztuczną inteligencją 212, 215–217
 profile ludzi, genów i sztucznej inteligencji 213–214
 zagrożenia 30
 ze strony fundamentalizmu 44

zagrożenie błędną klasyfikacją źródła 299, 309
 zagrożenie powszechnym unilateralizmem 446, 449
 zakłócenie gospodarcze 212
 zalecenia dotyczące polityki USA 173
 Departament Bezpieczeństwa Wewnętrznego 174–175
 Departament Obrony i Wspólnota Wywiadów 175–176
 Departament Stanu 175
 Kongres Stanów Zjednoczonych 173–174
 osoby i społeczeństwo 178
 sektor technologii 176–177
 środowiska akademickie 177–178
 zamknięty system termodynamiczny 90
 zapadnia 42
 zapadnia szyfrująca 49
 zarządzanie energią SI (SZE) 217
 zasada absolutnej większości 380–381
 zasada awersji do ryzyka 371–372
 Zasada Niedyskryminacji Ontogenów 78
 zasada niedyskryminacji podłoża 77
 zasada ostrożności 34–35
 zasada proaktywności 35
 zasada subiektywnego biegu czasu 80
 zasady Asilomar SI 424
 zasady niedyskryminacji 80
 zbieżność dążeń do militaryzacji SI i obrony narodowej 481–482
 ludzkie organizacje obronne 482
 tryby awaryjne wojskowej SI 483–484
 wojna błyskawiczna SI 482
 wojna pomiędzy SI i szkody poboczne 483
 wojskowa SI 482
 współpraca z ludźmi na wczesnych etapach rozwoju SI 481
 zbiorowe wejście w życie handlującej SI 405–406
 zdolności poznawcze 217
 zdolność „superperswazji” 459
 zdolność
 roszczenie 396
 zwiększenie 87
 zdrowy rozsądek 15, 237
 Zegar Zagłady 16
 zgodność z przepisami 244
 złośliwy nanobot 31
 znaczną przewagę strategiczną (ZPS) 397–398
 inicjator 406–407
 zobowiązanie do udostępniania 200–201
 ZPS, *patrz* znaczną przewagę strategiczną
 zrozumienie/świadomość 525–526
 zróżnicowane technologie GNR 37
 zwolennicy demokracji bezpośredniej 205

Ż

żywe stworzenia 31

Sztuczna inteligencja

Bezpieczeństwo i zabezpieczenia

Historia robotyki i sztucznej inteligencji pod pewnymi względami to także historia prób kontrolowania takich technologii przez ludzkość. Od praskiego Golema po nowoczesne roboty wojskowe trwa debata na temat tego, jaki stopień niezależności powinny mieć takie podmioty i jak upewnić się, że nie zwrócą się przeciwko ich wynalazcom. Liczne ostatnie postępy we wszystkich aspektach badań, rozwoju i wdrażaniu inteligentnych systemów są dobrze nagłośniane, aczkolwiek kwestie bezpieczeństwa i ochrony związane ze sztuczną inteligencją są rzadko podejmowane.

Sztuczna inteligencja. Bezpieczeństwo i zabezpieczenia jako pierwsza praca na ten temat ma na celu złagodzenie tego fundamentalnego problemu. Książka składa się z artykułów autorstwa wiodących uczonych z różnych dziedzin: filozofów, naukowców, pisarzy i ludzi biznesu, zajmujących się bezpieczeństwem sztucznej inteligencji, poświęconych różnym aspektom problemu jej kontroli oraz związanych z rozwojem bezpiecznej sztucznej inteligencji.

- **Część I** zawiera 11 przełomowych rozdziałów przedstawiających różne problemy odnoszące się do kontroli sztucznej inteligencji.
- **Część II** zawiera 17 rozdziałów prezentujących teoretyczne i praktyczne rozwiązania zagadnień poruszonych w części I.

Poszczególne teksty różnią się długością i treścią techniczną – od opiniotwórczych esejów o szerokiej tematyce po artykuły przedstawiające wysoce sformalizowane algorytmiczne podejścia do konkretnych problemów. Wszystkie są samodzielne i można je czytać w dowolnej kolejności lub pomijać bez utraty zrozumienia.

Książka nie jest ostatnim słowem na temat bezpieczeństwa sztucznej inteligencji, lecz jednym z pierwszych kroków w kierunku właściwego zrozumienia tego tematu.



Dr Roman V. Yampolskiy jest profesorem nadzwyczajnym w Department of Computer Engineering and Computer Science w Speed School of Engineering, University of Louisville (UofL), Kentucky. Jest założycielem i obecnym dyrektorem Cyber Security Lab.

Główne obszary jego zainteresowań to sztuczna inteligencja oraz jej bezpieczeństwo, biometria behawioralna, cyberbezpieczeństwo, algorytmy genetyczne i rozpoznawanie wzorców. Jest autorem ponad 150 publikacji, w tym wielu artykułów w czasopismach oraz książek.



Wydawnictwo
Naukowe PWN SA
pwn.pl • 801 33 33 88
ksiegarnia.pwn.pl

