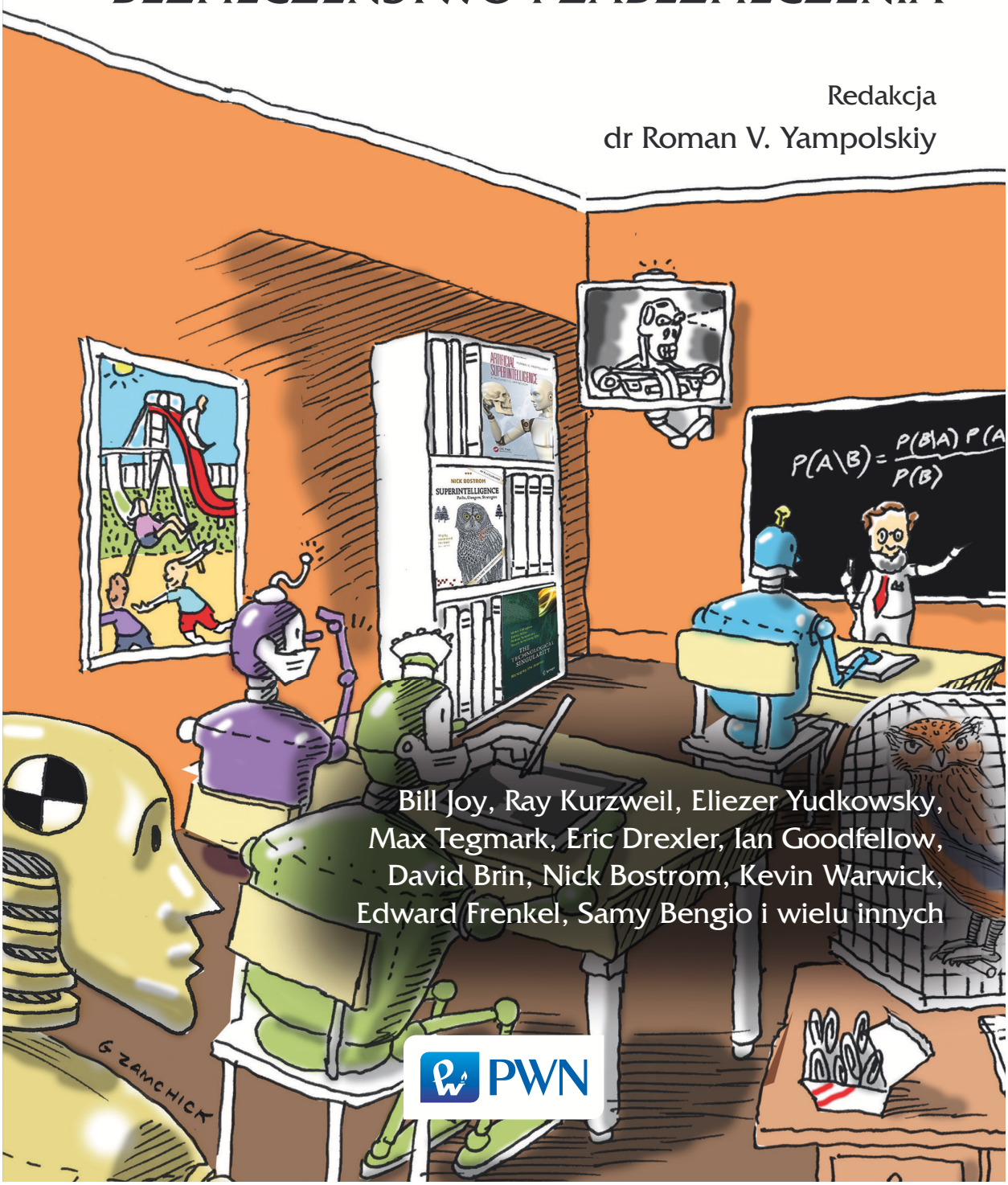
SZTUCZNA INTELIGENCJA
BEZPIECZEŃSTWO I ZABEZPIECZENIA

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

PWN

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**

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.

Okładka 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

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 najmniejszym 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 inżynieryjne") i wreszcie różne przypadki związane z wpływem otoczenia na system („środowisko"). Ponieważ w przypadku celowo projektowanej niebezpiecznej sztucznej inteligencji prawdopodobne jest zaistnienie wszystkich innych rodzajów problemów bezpieczeństwa, najniebezpieczniejszym 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)

W kolejnym artykule [2] zbadano, w jaki sposób można zbudować złośliwą sztuczną inteligencję i dlaczego ważne jest badanie i rozumienie 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 skarbce bankowe, omijano prawo, popełniano oszustwa przeciwko procesowi głosowania, przekupywano policjantów, szantażowano sędziów, uwierzytelniano 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 wzajemną 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 pożądanych funkcjonalności agentów, takie jak jazda na rowerze w kółko wokół celu [10], wstrzymywanie gry, aby uniknąć przegranej [11] lub wielokrotne dotykanie 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 traciło 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.****

---

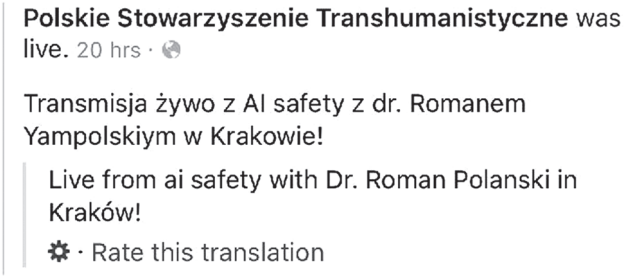
* https://www.technologyreview.com/s/601897/tougher-turing-test-exposes-chatbots-stupidity/
† https://www.forbes.com/sites/aarontilley/2014/04/03/googles-nest-stops-selling-its-smart-smoke-alarm-for-now
‡ https://gmail.googleblog.com/2015/11/computer-respond-to-this-email.html
§ http://time.com/3944181/robot-kills-man-volkswagen-plant/
¶ 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-skynet-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)
¶¶¶ 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 Przewidujący 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 takimi 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 prostalinowskimi, proznęcającymi 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 antysemicki.‡‡‡
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-whetherits-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".‡



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 szczere 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 przejęciem 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 błędu 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 zachęt 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 przedstawiono 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 Rushdie 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 tysiącleci 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 znieważają, popełnią 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 inżynierią 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, profesjonaliści 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 kontradyktoryjnych. 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 zostanie 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: Asurvey," *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 przejrzeniu, 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, Dhwani 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

Title of English-language original:
Artificial Intelligence Safety and Security edited by Roman V. Yampolskiy,
ISBN 9780815369820.

Sztuczna Inteligencja Bezpieczeństwo I Zabezpieczenia.
Roman V. Yampolskiy (Ed)
Wydawnictwo Naukowe (PWN) Publisher.
December, 2020.

Warszawa 2020
ISBN 978-83-01-21332-9
Wydanie I

Rozdział 1 - Rozdział 28
https://ksiegarnia.pwn.pl/Sztuczna-inteligencja,853577475,p.html

Indeks

C

D

E

F

G

L

M

N

O

P

V

W

X

Z

Ż

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

ISBN 978-83-01-21332-9