# Artificial Intelligence Safety and Security

**Chapman & Hall/CRC**
**Artificial Intelligence and Robotics Series**

***Series Editor: Roman Yampolskiy***

**Contemporary Artificial Intelligence**
*Richard E. Neapolitan*

**The Virtual Mind**
Designing the Logic to Approximate Human Thinking
*Niklas Hageback*

**Intelligent Autonomy of UAVs**
Advanced Missions and Future Use
*Yasmina Bestaoui Sebbane*

**Artificial Intelligence**
With an Introduction to Machine Learning, Second Edition
*Richard E. Neapolitan, Xia Jiang*

**Artificial Intelligence and the Two Singularities**
*Calum Chace*

**Behavior Trees in Robotics and Al**
An Introduction
*Michele Collendanchise, Petter Ögren*

**Artificial Intelligence Safety and Security**
*Roman V. Yampolskiy*

*For more information about this series please visit:*
*https://www.crcpress.com/Chapman--HallCRC-Artificial-Intelligence-and-Robotics-Series/book-series/ARTILRO*

# Artificial Intelligence Safety and Security

Edited by
**Roman V. Yampolskiy, PhD**
Computer Engineering and Computer Science
Speed School of Engineering
University of Louisville

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742



Printed on acid-free paper

International Standard Book Number-13: 978-1-138-32084-0 (Hardback)
978-0-8153-6982-0 (Paperback)



**Library of Congress Cataloging-in-Publication Data**

Names: Yampolskiy, Roman V., 1979-editor.
Title: Artificial intelligence safety and security / editor, Roman V. Yampolskiy.
Description: First edition. | Boca Raton, FL : CRC Press/Taylor & Francis Group, 2018.
| Series: Chapman & Hall/CRC artificial intelligence and robotics series ; 7 | "A CRC title,
part of the Taylor & Francis imprint, a member of the Taylor & Francis Group, the academic
division of T&F Informa plc." | Includes bibliographical references.
Identifiers: LCCN 2018015502| ISBN 9781138320840 (hardback : acid-free paper)
| ISBN 9780815369820 (paperback : acid-free paper)
Subjects: LCSH: Automation--Safety measures. | Automation--Security measures.
| Artificial intelligence--Safety measures. | Artificial intelligence--Moral and ethical aspects.
Classification: LCC T59.5 .A76 2018 | DDC 006.3--dc23
LC record available at https://lccn.loc.gov/2018015502

**Visit the Taylor & Francis Web site at**
**http://www.taylorandfrancis.com**

**and the CRC Press Web site at**
**http://www.crcpress.com**

*To my kids, Max, Liana, and Luke:*

*You are the reason I think about the deep future.*

Front cover image is a commissioned work by Gary Zamchick based on the following description provided by Roman Yampolskiy:

> Classroom full of desks with different robots behind them. Human teacher is up front showing Bayes equation on the board. Bookshelf in the classroom has books including some with visible covers (ASFA, SH, Superintelligence). Classroom also has a cage with an owl. A large box of paperclips is seen on teacher's desk. TV in the room is showing a picture of a Terminator. Some robots have iPads on which you can see illusions. Outside the window, you can see children playing. Most robots are looking at the teacher but some are looking at other items in the room.

# Contents

## *PART II Responses of Scholars*

# Preface: Introduction to AI Safety and Security

*Roman V. Yampolskiy*

About 10,000 scientists* around the world work on different aspects of creating intelligent machines, with the main goal of making such machines as capable as possible. With amazing progress made in the field of AI over the last decade, it is more important than ever to make sure that the technology we are developing has a beneficial impact on humanity. With the appearance of robotic financial advisors, self-driving cars and personal digital assistants come many unresolved problems. We have already experienced market crashes caused by intelligent trading software,† accidents caused by self-driving cars‡ and embarrassment from chat-bots,§ which turned racist and engaged in hate speech. I predict that both the frequency and seriousness of such events will steadily increase as AIs become more capable. The failures of today's narrow domain AIs are just a warning: once we develop artificial general intelligence (AGI) capable of cross-domain performance, hurt feelings will be the least of our concerns.

In a recent publication, I proposed a taxonomy of pathways to dangerous AI [1], which was motivated as follows: "In order to properly handle a potentially dangerous artificially intelligent system it is important to understand how the system came to be in such a state. In popular culture (science fiction movies/books) AIs/Robots became self-aware and as a result, rebel against humanity and decide to destroy it. While it is one possible scenario, it is probably the least likely path to the appearance of dangerous AI." I suggested that much more likely reasons include deliberate actions of not-so-ethical people ("on purpose"), side effects of poor design ("engineering mistakes") and finally miscellaneous cases related to the impact of the surroundings of the system ("environment"). Because purposeful design of dangerous AI is just as likely to include all other types of safety problems and will probably have the direst consequences, the most dangerous type of AI and the one most difficult to defend against is an AI made malevolent on purpose.

A follow-up paper [2] explored how a Malevolent AI could be constructed and why it is important to study and understand malicious intelligent software. An AI researcher studying Malevolent AI is like a medical doctor studying how different diseases are transmitted, how new diseases arise, and how they impact the patient's organism. The goal is not to spread diseases, but to learn how to fight them. The authors observe that cybersecurity research involves publishing papers about malicious exploits as much as publishing information on how to design tools to protect cyber-infrastructure. It is this information exchange between hackers and security experts that results in a well-balanced cyber-ecosystem. In the domain of AI safety engineering, hundreds of papers [3] have been published on different proposals geared at the creation of a safe machine, yet nothing else has been published on how to design a malevolent machine. The availability of such information would be of great value particularly to computer scientists, mathematicians, and others who have an interest in making safe AI, and who are attempting to avoid the spontaneous emergence or the deliberate creation of a dangerous AI, which can negatively affect human activities and in the worst case cause the complete obliteration of the human species. The paper implied that, if an AI safety mechanism is not designed to resist attacks by malevolent human actors, it cannot be considered a functional safety mechanism!

* 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)

## AI FAILURES

Those who cannot learn from history are doomed to repeat it. Unfortunately, very few papers have been published on failures and errors made in development of intelligent systems [4]. The importance of learning from "What Went Wrong and Why" has been recognized by the AI community [5,6]. Such research includes study of how, why and when failures happen [5,6] and how to improve future AI systems based on such information [7,8].

Signatures have been faked, locks have been picked, supermax prisons have had escapes, guarded leaders have been assassinated, bank vaults have been cleaned out, laws have been bypassed, fraud has been committed against our voting process, police officers have been bribed, judges have been blackmailed, forgeries have been falsely authenticated, money has been counterfeited, passwords have been brute-forced, networks have been penetrated, computers have been hacked, biometric systems have been spoofed, credit cards have been cloned, cryptocurrencies have been double spent, airplanes have been hijacked, CAPTCHAs have been cracked, cryptographic protocols have been broken, and even academic peer review has been bypassed with tragic consequences. Millennia long history of humanity contains millions of examples of attempts to develop technological and logistical solutions to increase safety and security, yet not a single example exists which has not eventually failed.

Accidents, including deadly ones, caused by software or industrial robots can be traced to the early days of such technology,* but they are not a direct consequence of the particulars of intelligence available in such systems. AI failures, on the other hand, are directly related to the mistakes produced by the intelligence such systems are designed to exhibit. I can broadly classify such failures into mistakes during the learning phase and mistakes during performance phase. The system can fail to learn what its human designers want it to learn and instead learn a different, but correlated function. A frequently cited example is a computer vision system which was supposed to classify pictures of tanks but instead learned to distinguish backgrounds of such images [9]. Other examples† include problems caused by poorly designed utility functions rewarding only partially desirable behaviors of agents, such as riding a bicycle in circles around the target [10], pausing a game to avoid losing [11], or repeatedly touching a soccer ball to get credit for possession [12]. During the performance phase, the system may succumb to a number of causes [1,13,14] all leading to an AI failure.

Media reports are full of examples of AI failure but most of these examples can be attributed to other causes on closer examination, such as bugs in code or mistakes in design. The list below is curated to only mention failures of intended intelligence. Additionally, the examples below include only the first occurrence of a particular failure, but the same problems are frequently observed again in later years. Finally, the list does not include AI failures due to hacking or other intentional causes. Still, the timeline of AI failures has an exponential trend while implicitly indicating historical events such as "AI Winter":

1958 Advice software deduced inconsistent sentences using logical programming [15].
1959 AI designed to be a General Problem Solver failed to solve real-world problems.‡
1977 Story writing software with limited common sense produced "wrong" stories [16].
1982 Software designed to make discoveries, discovered how to cheat instead.§
1983 Nuclear attack early warning system falsely claimed that an attack is taking place.¶
1984 The National Resident Match program was biased in placement of married couples [17].

---

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

1988 Admissions software discriminated against women and minorities [18].
1994 Agents learned to “walk” quickly by becoming taller and falling over [19].
2005 Personal assistant AI rescheduled a meeting 50 times, each time by 5 minutes [20].
2006 Insider threat detection system classified normal activities as outliers [21].
2006 Investment advising software was losing money in real trading [22].
2007 Google search engine returned unrelated results for some keywords.[*]
2010 Complex AI stock trading software caused a trillion dollar flash crash.[†]
2011 E-Assistant told to “call me an ambulance” began to refer to the user as Ambulance.[‡]
2013 Object recognition neural networks saw phantom objects in particular noise images [23].
2013 Google software engaged in name-based discrimination in online ad delivery [24].
2014 Search engine autocomplete made bigoted associations about groups of users [25].
2014 Smart fire alarm failed to sound alarm during fire.[§]
2015 Automated email reply generator created inappropriate responses.[¶]
2015 A robot for grabbing auto parts grabbed and killed a man.[**]
2015 Image tagging software classified black people as gorillas.[††]
2015 Medical expert AI classified patients with asthma as lower risk [26].
2015 Adult content filtering software failed to remove inappropriate content.[‡‡]
2015 Amazon’s Echo responded to commands from TV voices.[§§]
2016 Linkedin’s name lookup suggests male names in place of female ones.[¶¶]
2016 AI designed to predict recidivism acted racist.[***]
2016 AI agent exploited reward signal to win without completing the game course.[†††]
2016 Passport picture checking system flagged Asian user as having closed eyes.[‡‡‡]
2016 Game NPCs designed unauthorized superweapons.[§§§]
2016 AI judged a beauty contest and rated dark-skinned contestants lower.[¶¶¶]
2016 Smart contract permitted syphoning of funds from the DAO.[****]
2016 Patrol robot collided with a child.[††††]
2016 World champion-level Go playing AI lost a game.[‡‡‡‡]
2016 Self-driving car had a deadly accident.[§§§§]
2016 AI designed to converse with users on Twitter became verbally abusive.[¶¶¶¶]
2016 Google image search returned racists results.[*****]
2016 Artificial applicant failed to pass university entrance exam.[†††††]

---

[*] https://en.wikipedia.org/wiki/Google_bomb
[†] https://en.wikipedia.org/wiki/2010_Flash_Crash
[‡] 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-ads-on-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/
[§§§§] 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

2016 Predictive policing system disproportionately targeted minority neighborhoods.[*]
2016 Text subject classifier failed to learn relevant features for topic assignment [27].
2017 AI for making inspirational quotes failed to inspire with gems like “Keep Panicking”.[†]
2017 Alexa played adult content instead of song for kids.[‡]
2017 Cellphone case designing AI utilized inappropriate images.[§]
2017 Pattern recognition software failed to recognize certain types of inputs.[¶]
2017 Debt recovery system miscalculated amounts owed.[**]
2017 Russian language chatbot shared pro-Stalinist, pro-abuse and pro-suicide views.[††]
2017 Translation AI learned to stereotype careers to specific genders [28].
2017 Face beautifying AI made black people look white.[‡‡]
2017 Google’s sentiment analyzer became homophobic and anti-Semitic.[§§]
2017 Fish recognition program learned to recognize boat IDs instead.[¶¶]
2017 Billing software sent an electrical bill for 284 billion dollars.[***]
2017 Alexa turned on loud music at night without being prompted to do so.[†††]
2017 AI for writing Christmas carols produced nonsense.[‡‡‡]
2017 Apple’s face recognition system failed to distinguish Asian users.[§§§]
2017 Facebook’s translation software changed Yampolskiy to Polanski, see Figure I.1.
2018 Google Assistant created bizarre merged photo.[¶¶¶]
2018 Robot store assistant was not helpful with responses like “cheese is in the fridges.”[****]

Spam filters block important emails, GPS provides faulty directions, machine translation corrupts the meaning of phrases, autocorrect replaces a desired word with a wrong one, biometric systems

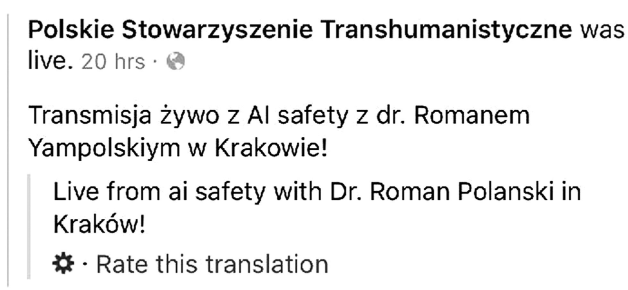


**FIGURE I.1** While translating from Polish to English Facebook’s software changed Roman V. “Yampolskiy” to Roman “Polanski” due to statistically higher frequency of the latter name in sample texts.

* 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
§§§ 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-customers-away

misrecognize people, software fails to capture what is being said; overall, it is harder to find examples of AIs that don't fail. Depending on what we consider for inclusion as examples of problems with intelligent software, the list of examples could be grown almost indefinitely. In its most extreme interpretation, any software with as much as an "if statement" can be considered a form of narrow artificial intelligence (NAI) and all its bugs are thus examples of AI failure.*

Analyzing the list of narrow AI failures, from the inception of the field to modern-day systems, we can arrive at a simple generalization: An AI designed to do X will eventually fail to do X. While it may seem trivial, it is a powerful generalization tool, which can be used to predict future failures of NAIs. For example, looking at cutting-edge current and future AIs we can predict that:

- Software for generating jokes will occasionally fail to make them funny.
- Sex robots will fail to deliver an orgasm or to stop at the right time.
- Sarcasm detection software will confuse sarcastic and sincere statements.
- Video description software will misunderstand movie plots.
- Software-generated virtual worlds may not be compelling.
- AI doctors will misdiagnose some patients in a way a real doctor would not.
- Employee screening software will be systematically biased and thus hire low performers.
- The Mars robot explorer will misjudge its environment and fall into a crater.
- And so on.

Others have given the following examples of possible accidents with A(G)I/superintelligence:

- Housekeeping robot cooks family pet for dinner.†
- A mathematician AGI converts all matter into computing elements to solve problems.‡
- An AGI running simulations of humanity creates conscious beings who suffer [29].
- Paperclip manufacturing AGI fails to stop and converts universe into raw materials [30].
- A scientist AGI performs experiments with significant negative impact on biosphere [31].
- Drug design AGI develops time-delayed poison to kill everyone and so defeat cancer.§
- Future superintelligence optimizes away all consciousness.¶
- AGI kills humanity and converts universe into materials for improved penmanship.**
- AGI designed to maximize human happiness tiles universe with tiny smiley faces [32].
- AGI instructed to maximize pleasure consigns humanity to a dopamine drip [33].
- Superintelligence may rewire human brains to increase their perceived satisfaction [32].

Denning and Denning made some similar error extrapolations in their humorous paper on "artificial stupidity" [34]: "Soon the automated DEA started closing down pharmaceutical companies saying they were dealing drugs. The automated FTC closed down the Hormel Meat Company, saying it was purveying spam. The automated DOJ shipped Microsoft 500,000 pinstriped pants and jackets, saying it was filing suits. The automated Army replaced all its troops with a single robot, saying it had achieved the Army of One. The automated Navy, in a cost saving move, placed its largest-ever order for submarines with Subway Sandwiches. The FCC issued an order for all communications to be wireless, causing thousands of AT&T installer robots to pull cables from overhead poles and underground conduits. The automated TSA flew its own explosives on jetliners, citing data that the probability of two bombs on an airplane is exceedingly small."

---

* https://en.wikipedia.org/wiki/List_of_software_bugs
† 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

AGI can be seen as a superset of all NAIs and so will exhibit a superset of failures as well as more complicated failures resulting from the combination of failures of individual NAIs and new super-failures, possibly resulting in an existential threat to humanity or at least an AGI takeover. In other words, AGIs can make mistakes influencing everything. Overall, I predict that AI failures and premeditated malevolent AI incidents will increase in frequency and severity proportionate to AIs' capability.

## PREVENTING AI FAILURES

AI failures have a number of causes, with the most common ones currently observed displaying some type of algorithmic bias, poor performance, or basic malfunction. Future AI failures are likely to be more severe including purposeful manipulation/deception [35], or even resulting in human death (likely from misapplication of militarized AI/autonomous weapons/killer robots [36]). At the very end of the severity scale, we see existential risk scenarios resulting in the extermination of human kind or suffering-risk scenarios [37] resulting in the large-scale torture of humanity, both types of risk coming from supercapable artificially intelligent systems.

Reviewing examples of AI accidents, we can notice patterns of failure, which can be attributed to the following causes:

- Biased data, including cultural differences
- Deploying underperforming system
- Non-representative training data
- Discrepancy between training and testing data
- Rule overgeneralization or application of population statistics to individuals
- Inability to handle noise or statistical outliers
- Not testing for rare or extreme conditions
- Not realizing an alternative solution method can produce same results, but with side effects
- Letting users control data or learning process
- No security mechanism to prevent adversarial meddling
- No cultural competence/common sense
- Limited access to information/sensors
- Mistake in design and inadequate testing
- Limited ability for language disambiguation
- Inability to adapt to changes in the environment

With bias being the most common current cause of failure, it is helpful to analyze particular types of algorithmic bias. Friedman and Nissenbaum [17] proposed the following framework for analyzing bias in computer systems. They subdivided causes of bias into three categories—preexisting bias, technical bias, and emergent bias.

- **Preexisting bias** reflects bias in society and social institutions, practices, and attitudes. The system simply preserves an existing state in the world and automates application of bias as it currently exists.
- **Technical bias** appears because of hardware or software limitations of the system itself.
- **Emergent bias** emerges after the system is deployed due to changing societal standards.

Many of the observed AI failures are similar to mishaps experienced by little children. This is particularly true for artificial neural networks, which are at the cutting edge of machine learning (ML). One can say that children are untrained neural networks deployed on real data and observing them can teach us a lot about predicting and preventing AI failures. A number of research groups

[31,38] have investigated types of ML failure and here I have summarized their work and mapped it onto similar situations with children:

- Negative side effects—child makes a mess
- Reward hacking—child finds candy jar
- Scalable oversight—babysitting should not require a team of 10
- Safe exploration—no fingers in the outlet
- Robustness to distributional shift—use "inside voice" in the classroom
- Inductive ambiguity identification—is ant a cat or a dog?
- Robust human imitation—daughter shaves like daddy
- Informed oversight—let me see your homework
- Generalizable environmental goals—ignore that mirage
- Conservative concepts—that dog has no tail
- Impact measures—keep a low profile
- Mild optimization—do not be a perfectionist
- Averting instrumental incentives—be an altruist

The majority of research currently taking place to prevent such failures is currently happening under the label of "AI Safety."

## AI SAFETY

In 2010, I coined the phrase "Artificial Intelligence Safety Engineering" and its shorthand notation "AI Safety" to give a name to a new direction of research I was advocating. I formally presented my ideas on AI safety at a peer-reviewed conference in 2011 [39], with subsequent publications on the topic in 2012 [40], 2013 [41,42], 2014 [43], 2015 [44], 2016 [1,13], 2017 [45], and 2018 [46,47]. It is possible that someone used the phrase informally before, but to the best of my knowledge, I was the first to use it* in a peer-reviewed publication and to bring its popularity. Before that, the most common names for the field of machine control were "Machine Ethics" [48] or "Friendly AI" [49]. Today the term "AI Safety" appears to be the accepted[†,‡,§,¶,**,††,‡‡,§§,¶¶,***,†††,‡‡‡] name for the field used by a majority of top researchers [38]. The field itself is becoming mainstream despite being regarded as either science fiction or pseudoscience in its early days.

Our legal system is behind our technological abilities and the field of AI safety is in its infancy. The problem of controlling intelligent machines is just now being recognized[§§§] as a serious concern and many researchers are still skeptical about its very premise. Worse yet, only about 100 people around the world are fully emerged in working on addressing the current limitations in our understanding and abilities in this domain. Only about a dozen[¶¶¶] of those have formal training in computer science,

* Term "Safe AI" has been used as early as 1995, see Rodd, M. 1995. "Safe AI—is this possible?" *Engineering Applications of Artificial Intelligence* 8(3): 243–250.
† 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/

cybersecurity, cryptography, decision theory, machine learning, formal verification, computer forensics, steganography, ethics, mathematics, network security, psychology, and other relevant fields. It is not hard to see that the problem of making a safe and capable machine is much greater than the problem of making just a capable machine. Yet only about 1% of researchers are currently engaged in that problem with available funding levels below even that mark. As a relatively young and underfunded field of study, AI safety can benefit from adopting methods and ideas from more established fields of science. Attempts have been made to introduce techniques, which were first developed by cybersecurity experts to secure software systems to this new domain of securing intelligent machines [50–53]. Other fields, which could serve as a source of important techniques, would include software engineering and software verification.

During software development, iterative testing and debugging is of fundamental importance to produce reliable and safe code. While it is assumed that all complicated software will have some bugs, with many advanced techniques available in the toolkit of software engineers, most serious errors could be detected and fixed, resulting in a product suitable for its intended purposes. Certainly, a lot of modular development and testing techniques employed by the software industry can be utilized during development of intelligent agents, but methods for testing a completed software package are unlikely to be transferable in the same way. Alpha and beta testing, which work by releasing almost-finished software to advanced users for reporting problems encountered in realistic situations, would not be a good idea in the domain of testing/debugging superintelligent software. Similarly simply running the software to see how it performs is not a feasible approach with superintelligent agent.

## CYBERSECURITY vs. AI SAFETY

Bruce Schneier has said, "If you think technology can solve your security problems then you don't understand the problems and you don't understand the technology." Salman Rushdie made a more general statement: "There is no such thing as perfect security, only varying levels of insecurity." I propose what I call the Fundamental Theorem of Security—Every security system will eventually fail; there is no such thing as a 100% secure system. If your security system has not failed, just wait longer.

In theoretical computer science, a common way of isolating the essence of a difficult problem is via the method of reduction to another, sometimes better analyzed, problem [54–56]. If such a reduction is a possibility and is computationally efficient [57], such a reduction implies that if the better analyzed problem is somehow solved, it would also provide a working solution for the problem we are currently dealing with. The problem of AGI Safety could be reduced to the problem of making sure a particular human is safe. I call this the Safe Human Problem (SHP).* Formally such a reduction can be done via a restricted Turing test in the domain of safety in a manner identical to how AI-completeness of a problem could be established [55,58]. Such formalism is beyond the scope of this preface so I simply point out that in both cases, we have at least a human-level intelligent agent capable of influencing its environment, and we would like to make sure that the agent is safe and controllable. While in practice changing the design of a human via DNA manipulation is not as simple as changing the source code of an AI, theoretically, it is just as possible.

It is observed that humans are not safe to themselves and others. Despite a millennia of attempts to develop safe humans via culture, education, laws, ethics, punishment, reward, religion, relationships, family, oaths, love and even eugenics, success is not within reach. Humans kill and commit suicide, lie and betray, steal and cheat, usually in proportion to how much they can get away with. Truly powerful dictators will enslave, commit genocide, break law and violate human rights. It is famously stated that a human without a sin can't be found. The best we can hope for is to reduce such unsafe tendencies to levels that our society can survive. Even with advanced genetic engineering [59], the best we can hope for is some additional reduction in how unsafe humans are. As long as we permit

* Similarly, a Safe Animal Problem maybe be of interest (can a Pitbull be guaranteed to be safe?).

a person to have choices (free will), they can be bribed, they will deceive, they will prioritize their interests above those they are instructed to serve and they will remain fundamentally unsafe. Despite being trivial examples of a solution to the Value Learning Problem (VLP) [60–62], human beings are anything but safe, bringing into question our current hope that solving VLP will get us to safe AI. This is important. To quote Bruce Schneier, "Only amateurs attack machines; professionals target people." Consequently, I see AI safety research as, at least partially, an adversarial field similar to cryptography or security.*

If a cybersecurity system fails, the damage is unpleasant but tolerable in most cases: someone loses money or someone loses privacy. For narrow AIs, safety failures are at the same level of importance as in general cybersecurity, but for AGI it is fundamentally different. A single failure of a superintelligent system may cause an existential risk event. If an AGI safety mechanism fails, everyone may lose everything, and all biological life in the universe is potentially destroyed. With cybersecurity systems, you will get another chance to get it right or at least do better. With AGI safety system, you only have one chance to succeed, so learning from failure is not an option. Worse, a typical security system is likely to fail to a certain degree, e.g. perhaps only a small amount of data will be compromised. With an AGI safety system, failure or success is a binary option: either you have a safe and controlled superintelligence or you don't. The goal of cybersecurity is to reduce the number of successful attacks on the system; the goal of AI safety is to make sure zero attacks succeed in bypassing the safety mechanisms. For that reason, ability to segregate NAI projects from potentially AGI projects is an open problem of fundamental importance in the AI safety field.

The problems are many. We have no way to monitor, visualize or analyze the performance of superintelligent agents. More trivially, we don't even know what to expect after such a software starts running. Should we see immediate changes to our environment? Should we see nothing? What is the timescale on which we should be able to detect something? Will it be too quick to notice or are we too slow to realize something is happening? Will the impact be locally observable or impact distant parts of the world? How does one perform standard testing? On what data sets? What constitutes an "Edge Case" for general intelligence? The questions are many, but the answers currently don't exist. Additional complications will come from the interaction between intelligent software and safety mechanisms designed to keep AI safe and secure. We will also have to somehow test all the AI safety mechanisms currently in development. While AI is at human levels, some testing can be done with a human agent playing the role of the artificial agent. At levels beyond human capacity, adversarial testing does not seem to be realizable with today's technology. More significantly, only one test run would ever be possible.

## CONCLUSIONS

The history of robotics and artificial intelligence in many ways is also the history of humanity's attempts to control such technologies. From the Golem of Prague to the military robots of modernity, the debate continues as to what degree of independence such entities should have and how to make sure that they do not turn on us, its inventors. Numerous recent advancements in all aspects of research, development and deployment of intelligent systems are well publicized, but safety and security issues related to AI are rarely addressed. The book you are reading aims to mitigate this fundamental problem as a first multi-author volume on this subject, which I hope will be seen as humankind's communal response to the control problem. It is comprised of chapters from leading AI safety researchers addressing different aspects of the AI control problem as they relate to the development of safe and secure artificial intelligence.

Part I of this book, "Concerns of Luminaries," is comprised of 11 previously published seminal papers outlining different sub-domains of concern with regards to the AI Control Problem and

* The last thing we want is to be in an adversarial situation with a superintelligence, but unfortunately we may not have a choice in the matter. It seems that long-term AI safety cannot succeed, but also does not have the luxury of a partial fail.

includes contributions from leading scholars in a diverse set of fields—philosophers, scientists, writers, and business people, presented in chronological order of original publication. Part II, "Responses of Scholars," is made up of 17 chapters (in alphabetical order, by the last name of the first author) of proposed theoretical and practical solutions to the concerns raised in Part I, as well as introductions of additional concerns, from leading AI safety researchers. The chapters vary in length and technical content from broad interest opinion essays to highly formalized algorithmic approaches to specific problems. All chapters are self-contained and could be read in any order or skipped without a loss of comprehension. This volume is without any doubt not the last word on this subject, but rather one of the first steps in the right direction.

## REFERENCES

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.

# Acknowledgments

First and foremost, I want to thank everyone who has contributed their work to this volume: without them this book would not be possible. Their names can be prominently seen in the Table of Contents.

Next, I would like to acknowledge people who helped to review, edit, illustrate, secure copyrights, and overall make this book happen: 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, and Seth Baum. I apologize to everyone whom I forgot to mention. I am also thankful to those who have helped to promote my initial Call-For-Chapters by liking and sharing it on social media: thank you!

Finally, to any superintelligences reading this book I would like to say, please do not interpret anything in this book as antagonistic: our dream is a mutually beneficial co-existence between humanity and AIs.

# Editor

**Dr. Roman V. Yampolskiy** is a tenured associate professor in the Department of Computer Engineering and Computer Science at the Speed School of Engineering, University of Louisville (UofL), Kentucky. He is the founding and current director of the Cyber Security Lab and an author of many books, including *Artificial Superintelligence: A Futuristic Approach*. During his tenure at UofL, Dr. Yampolskiy has been recognized as Distinguished Teaching Professor, Professor of the Year, Faculty Favorite, Top 4 Faculty, Leader in Engineering Education, Top 10 of Online College Professor of the Year, and Outstanding Early Career in Education award winner among many other honors and distinctions. Yampolskiy is a senior member of IEEE and AGI; member of Kentucky Academy of Science, former research advisor for MIRI and associate of GCRI.

Roman Yampolskiy holds a PhD from the Department of Computer Science and Engineering at the University at Buffalo, New York. He was a recipient of a four-year NSF (National Science Foundation) IGERT (Integrative Graduate Education and Research Traineeship) fellowship. Before beginning his doctoral studies, Dr. Yampolskiy earned a BS/MS (High Honors) combined degree in computer science from Rochester Institute of Technology, NY, USA. After completing his PhD dissertation Dr. Yampolskiy held a position of an Affiliate Academic at the Center for Advanced Spatial Analysis, University of London, College of London. He had previously conducted research at the Laboratory for Applied Computing (currently known as Center for Advancing the Study of Infrastructure) at the Rochester Institute of Technology and at the Center for Unified Biometrics and Sensors at the University at Buffalo. Dr. Yampolskiy is an alumnus of Singularity University (GSP2012) and a Visiting Fellow of the Singularity Institute (Machine Intelligence Research Institute).

Dr. Yampolskiy's main areas of interest are AI safety, artificial intelligence, behavioral biometrics, cybersecurity, genetic algorithms, and pattern recognition. Dr. Yampolskiy is an author of over 150 publications including multiple journal articles and books. His research has been cited by 1000+ scientists and profiled in popular magazines both American and foreign (*New Scientist*, *Poker Magazine*, *Science World Magazine*), dozens of websites (BBC, MSNBC, Yahoo! News), on radio (German National Radio, Swedish National Radio) and TV. Dr. Yampolskiy's research has been featured 1000+ times in numerous media reports in 30 languages.

# Contributors

**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