UNIVERSITY OF
LOUISVILLE

Future of AI, Safety & Security
人工智能、安全和防御的未来

Dr. Roman.Yampolskiy@louisville.edu

Computer Engineering and Computer Science
University of Louisville - cecs.louisville.edu/ry
Director – CyberSecurity Lab

twitter @romanyam

Follow me on Facebook /roman.yampolskiy

Artificial Intelligence is Here …

人工智能到来 …

Google

The Knowledge Graph

Learn more about one of the key breakthroughs behind the future of search

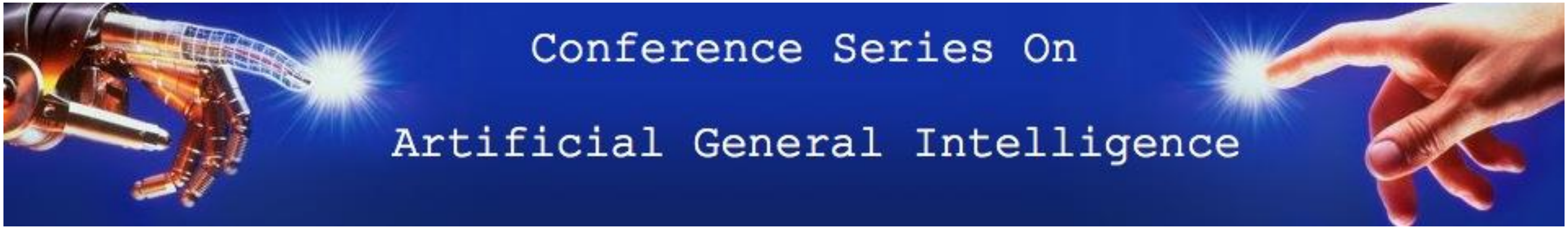
Siri
On May 19 remind me it's dad's birthday
Here's your reminder for May 19, 2012 at 9 am:
19 Saturday May 2012
Dad's birthday
Cancel
Confirm

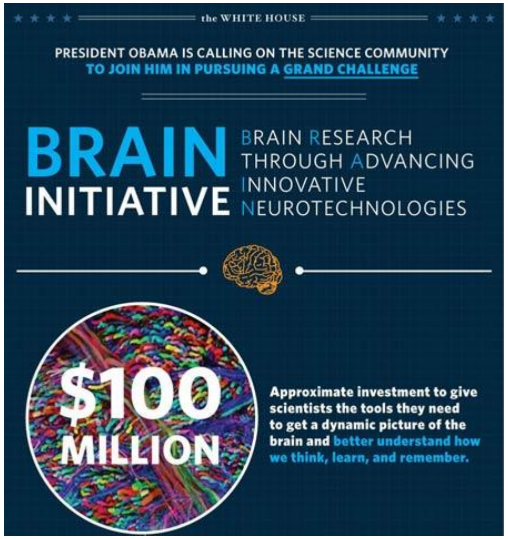
Robots are Here …

机器人到来 …

SuperIntelligence is Coming …

超级智能即将来临



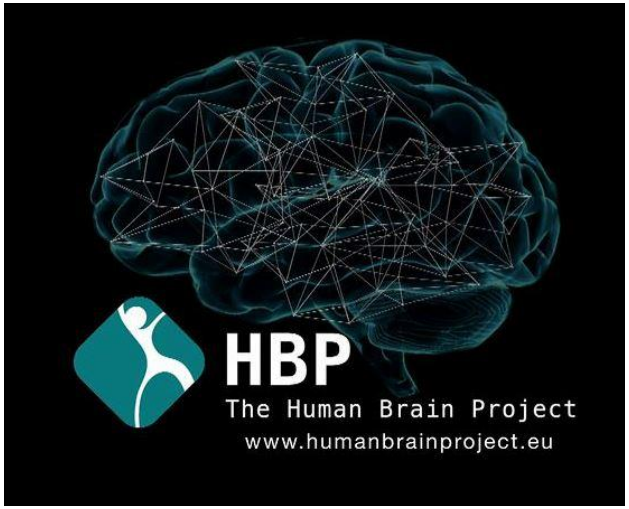

the WHITE HOUSE

PRESIDENT OBAMA IS CALLING ON THE SCIENCE COMMUNITY TO JOIN HIM IN PURSUING A GRAND CHALLENGE

BRAIN INITIATIVE BRAIN RESEARCH THROUGH ADVANCING INNOVATIVE NEUROTECHNOLOGIES

$100 MILLION

Approximate investment to give scientists the tools they need to get a dynamic picture of the brain and better understand how we think, learn, and remember.



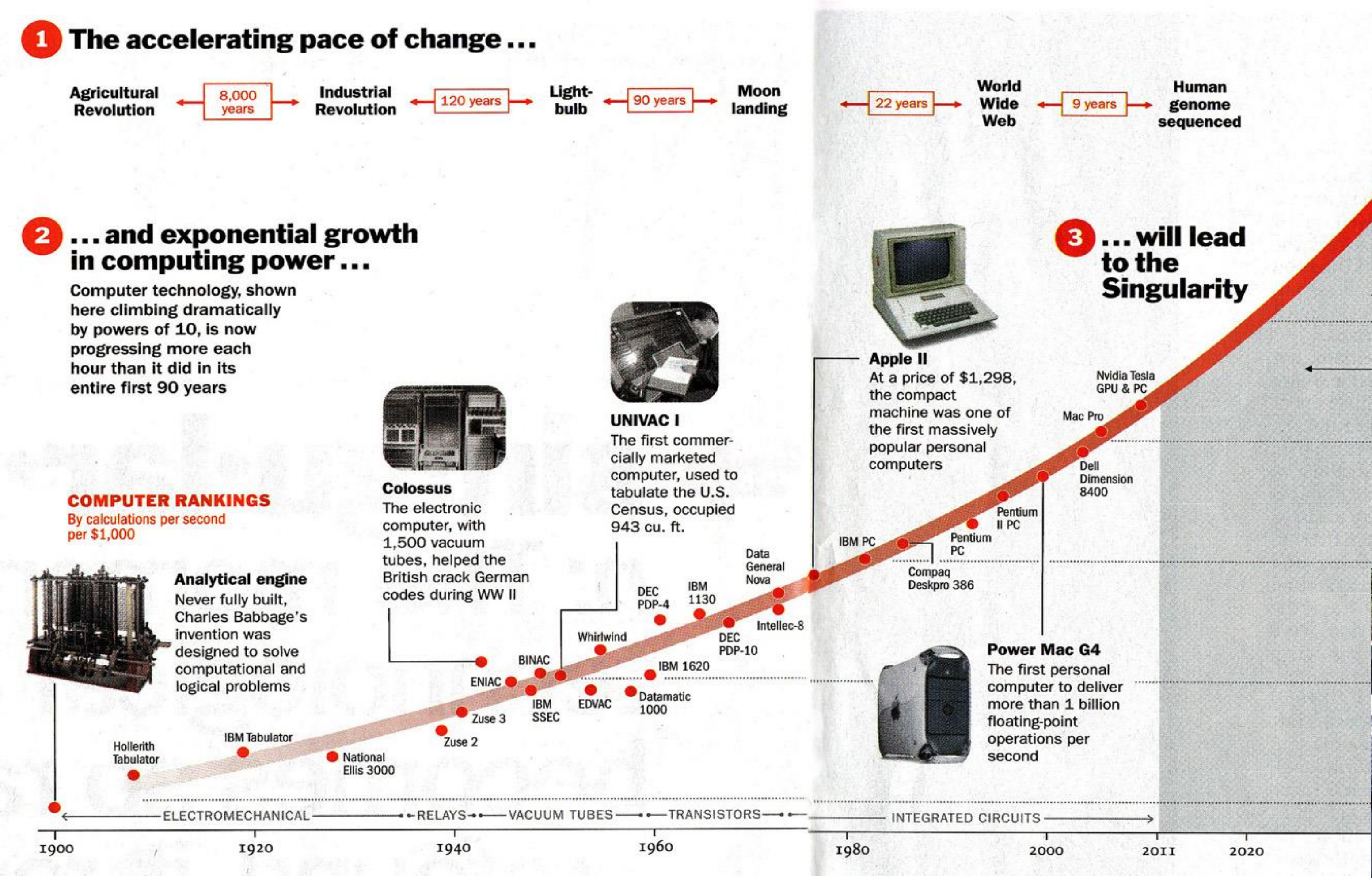
BRAIN RESEARCH THROUGH ADVANCING INNOVATIVE NEUROTECHNOLOGIES
BRAIN
INITIATIVE

vicarious

MetaMind

GoodAI

SuperSoon 超级临近

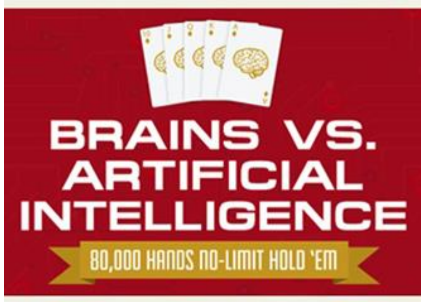


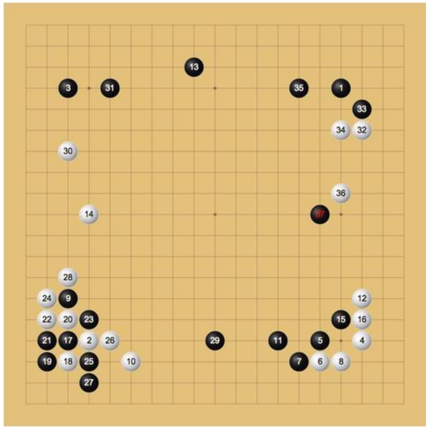
- Raymond Kurzweil 在《时代》杂志
- 2023-2045+

IBM Watson

SuperSmart
超级聪明

Google DeepMind


BRAINS VS. ARTIFICIAL INTELLIGENCE
80,000 HANDS NO-LIMIT HOLD 'EM



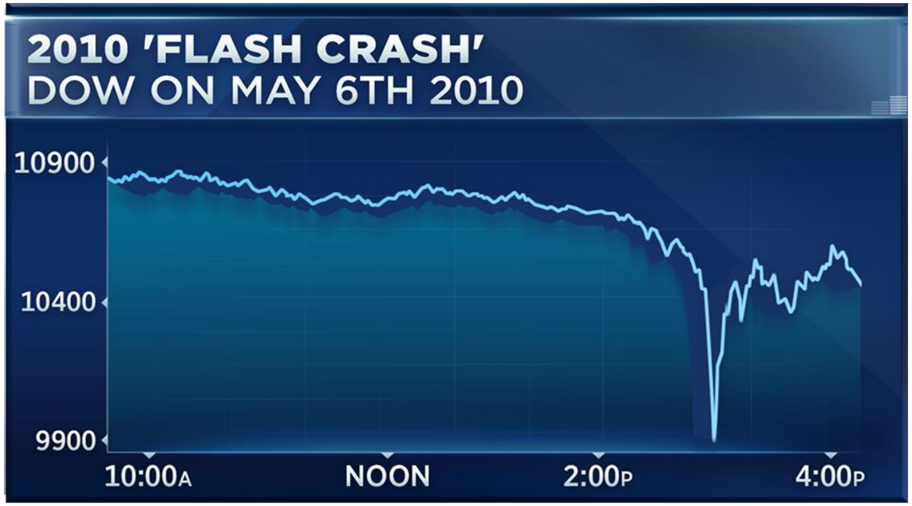
AlphaGo

SuperComplex
超级复杂

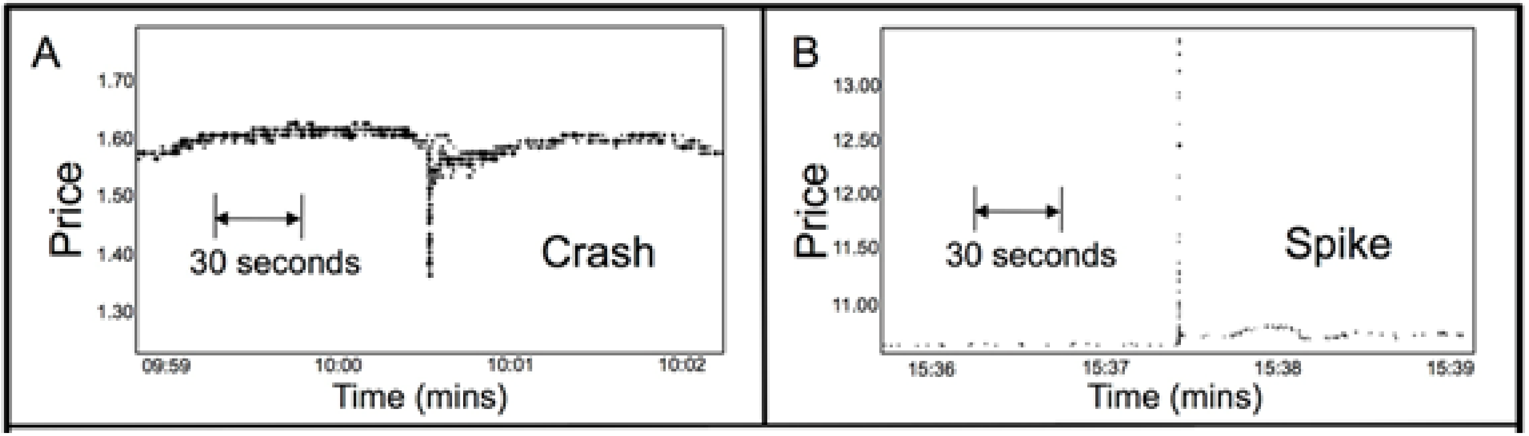
“飞机操作人员或飞行员中很少有人了解那部分软件。”

SuperFast 超级快速

PHOTO: DANIEL ACKER/BLOOMBERG VIA GETTY IMAGES



Ultrafast Extreme Events (UEEs)



Abrupt Rise of New Machine Ecology **Beyond Human Response Time**.
By Johnson et al. Nature. Scientific Reports 3, #2627 (2013)

SuperControlling 超级控制

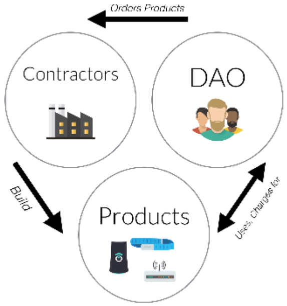
能源：核电站
Energy: Nuclear Power Plants

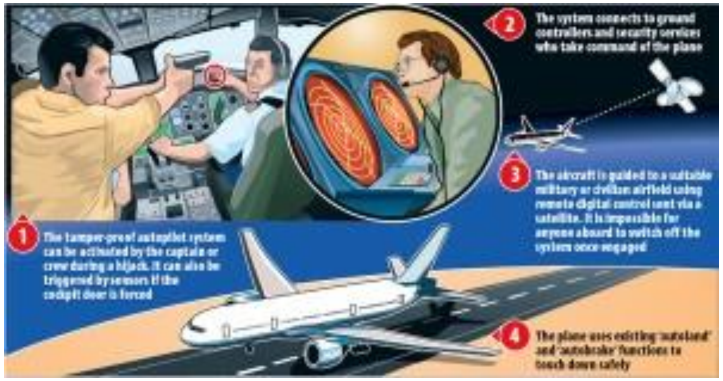
公共事业：水厂/电网
Utilities: Water Plants/Electrical Grid

军事：核武器
Military: Nuclear Weapons

Communications: Satellites

股市：75%以上的交易订单由自动交易系统产生
Stock Market: 75+% of all trade orders generated by Automated Trading Systems

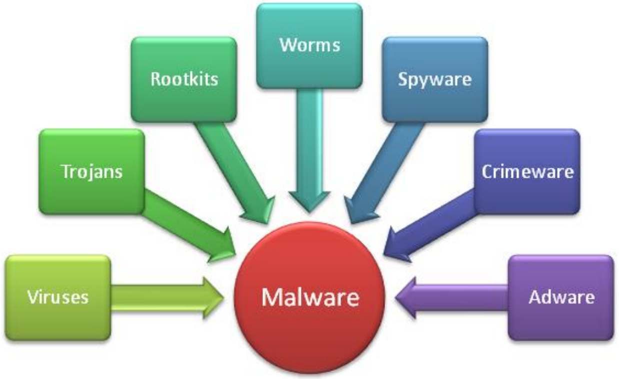


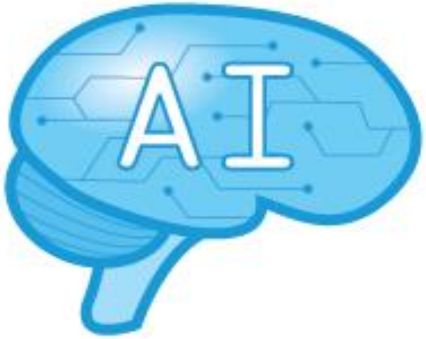

1 The tamper-proof autopilot system can be activated by the captain or crew during a hijack. It can also be triggered by sensors if the cockpit door is forced
2 The system connects to ground controllers and security services who take command of the plane
3 The aircraft is guided to a suitable military or civilian airfield using remote digital control sent via a satellite. It is impossible for anyone aboard to switch off the system once engaged
4 The plane uses existing 'autoland' and 'autobrake' functions to touch down safely

航空：不间断自动巡航系统
Aviation: Uninterruptible Autopilot System

SuperViruses

超级病毒





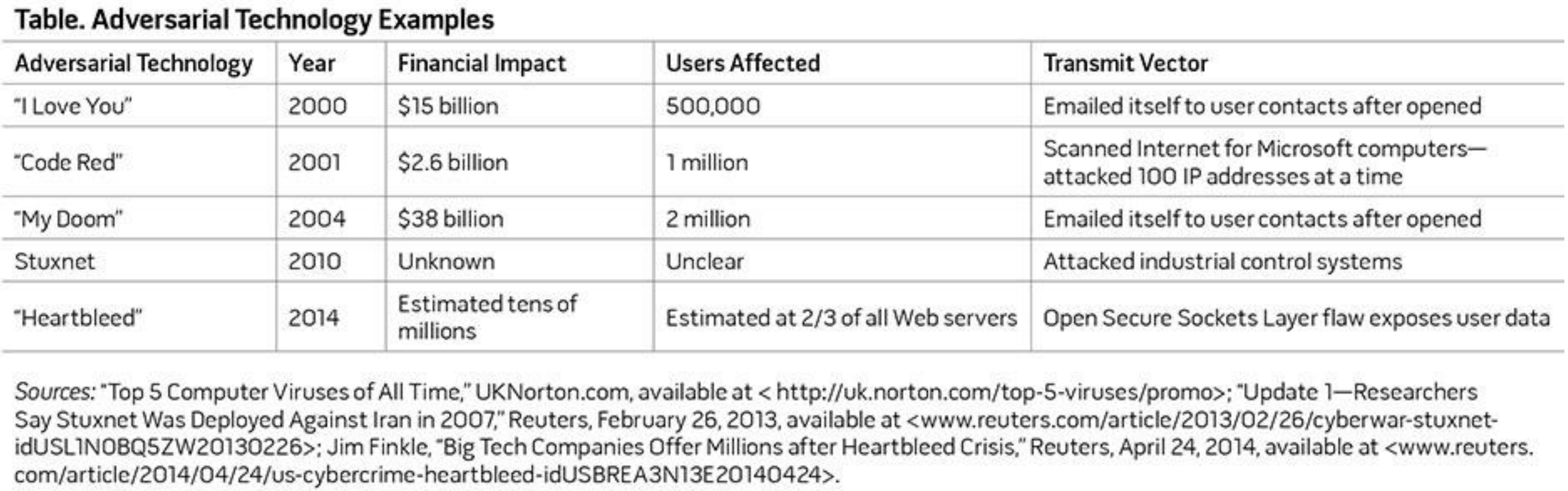


Table. Adversarial Technology Examples

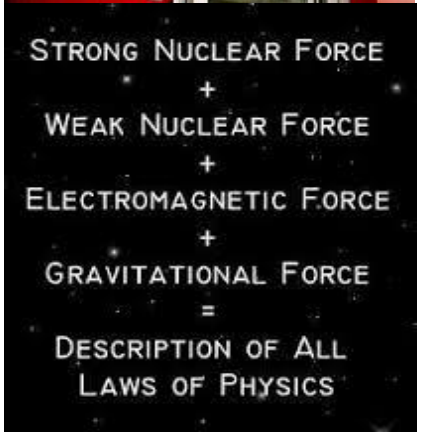
Adversarial Technology	Year	Financial Impact	Users Affected	Transmit Vector
"I Love You"	2000	$15 billion	500,000	Emailed itself to user contacts after opened
"Code Red"	2001	$2.6 billion	1 million	Scanned Internet for Microsoft computers—attacked 100 IP addresses at a time
"My Doom"	2004	$38 billion	2 million	Emailed itself to user contacts after opened
Stuxnet	2010	Unknown	Unclear	Attacked industrial control systems
"Heartbleed"	2014	Estimated tens of millions	Estimated at 2/3 of all Web servers	Open Secure Sockets Layer flaw exposes user data

Sources: "Top 5 Computer Viruses of All Time," UKNorton.com, available at < http://uk.norton.com/top-5-viruses/promo>; "Update 1—Researchers Say Stuxnet Was Deployed Against Iran in 2007," Reuters, February 26, 2013, available at <www.reuters.com/article/2013/02/26/cyberwar-stuxnet-idUSL1N0BQ5ZW20130226>; Jim Finkle, "Big Tech Companies Offer Millions after Heartbleed Crisis," Reuters, April 24, 2014, available at <www.reuters.com/article/2014/04/24/us-cybercrime-heartbleed-idUSBREA3N13E20140424>.

Relying on Kindness of Machines? **The Security Threat of Artificial Agents.**
By Randy Eshelman and Douglas Derrick. JFQ 77, 2nd Quarter 2015.

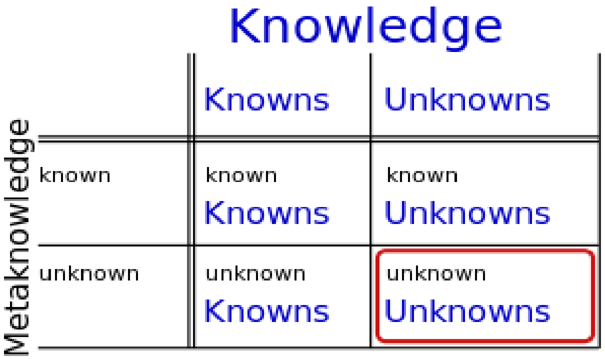
Positive Impacts of SuperIntelligence

超级智能带来的积极影响


STRONG NUCLEAR FORCE
+
WEAK NUCLEAR FORCE
+
ELECTROMAGNETIC FORCE
+
GRAVITATIONAL FORCE
=
DESCRIPTION OF ALL LAWS OF PHYSICS

The limitations of the Robot Scientist
Do you feel like going out for noodles?
Never.
Wilcox

Photo Courtesy of Shutterstock

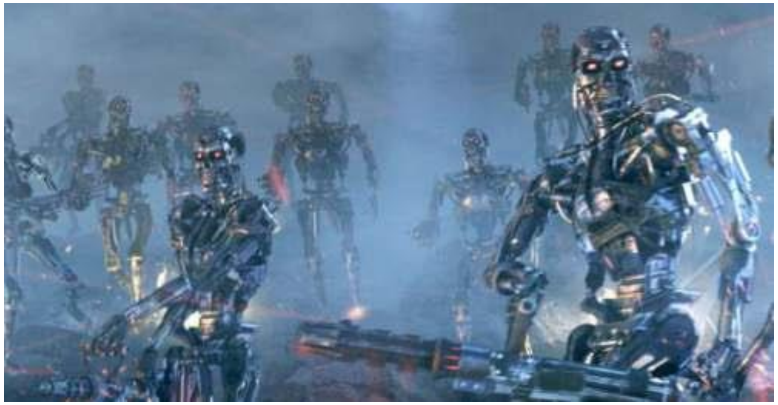


	Knowledge	
Metaknowledge	Knowns	Unknowns
known	known Knowns	known Unknowns
unknown	unknown Knowns	unknown Unknowns

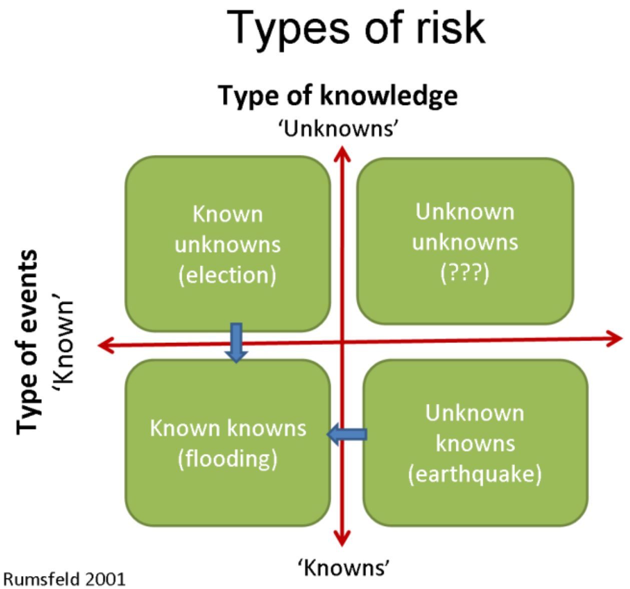
Negative Impacts of SuperIntelligence

超级智能带来的一些问题

© DESPAIR.COM
ADAPTATION
The Bad News is Robots Can Do Your Job Now. The Good News is We're Now Hiring Robot Repair Technicians. The Worse News is We're Working on Robot-Fixing Robots- and We Do Not Anticipate Any Further Good News.



LIVE
Optimus Prime Wins Election!
CNN
VOWS TO END DECEPTICON THREAT
HUNT THEM DOWN WHEREVER THEY HIDE
FUTURES NAS ▼5.00
LEAFLETS DROPPED IN SOUTHERN IRAQ CNN NEW DEATH


Types of risk

Type of knowledge

'Unknowns'

Type of events 'Known'

Known unknowns (election)

Unknown unknowns (???)

Known knowns (flooding)

Unknown knowns (earthquake)

'Unknown

'Knowns'

From Rumsfeld 2001

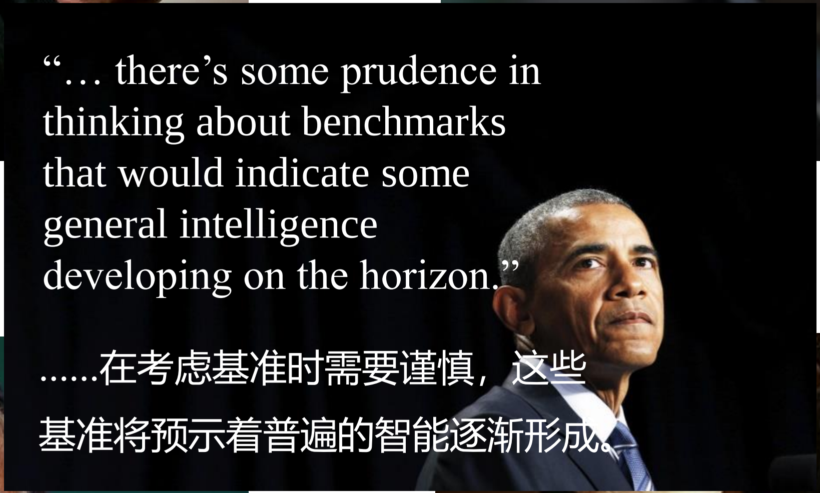
Concerns About A(G)I
关于人工智能的担忧

“The development of full artificial intelligence could spell the end of the human race.”

完全人工智能的发展可能意味着人类的终结。

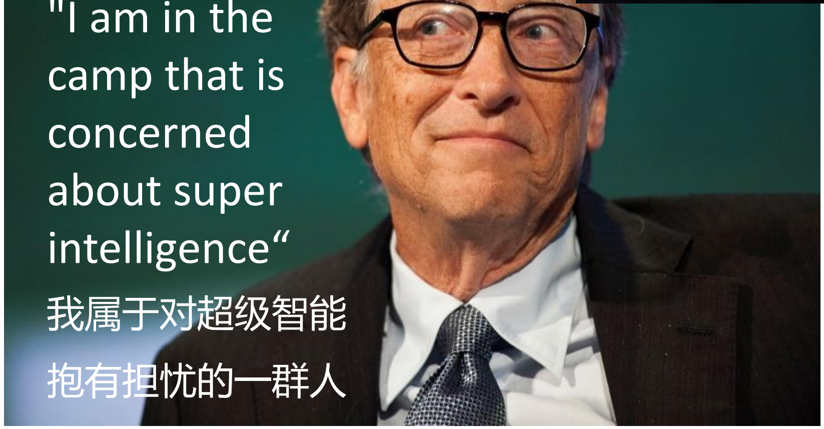
“I think we should be very careful about artificial intelligence”

我认为我们应该对人工智保持谨慎的态度



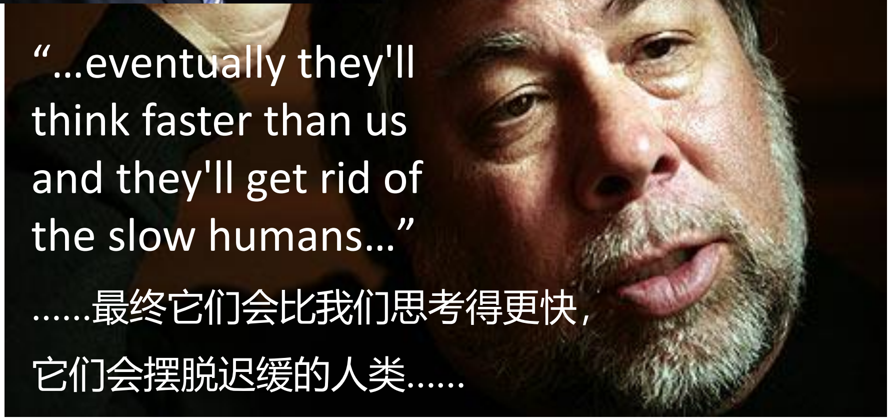
“… there’s some prudence in thinking about benchmarks that would indicate some general intelligence developing on the horizon.”

……在考虑基准时需要谨慎，这些基准将预示着普遍的智能逐渐形成。



"I am in the camp that is concerned about super intelligence“

我属于对超级智能抱有担忧的一群人



“…eventually they'll think faster than us and they'll get rid of the slow humans…”

……最终它们会比我们思考得更快，它们会摆脱迟缓的人类……

Taxonomy of Pathways to Dangerous AI

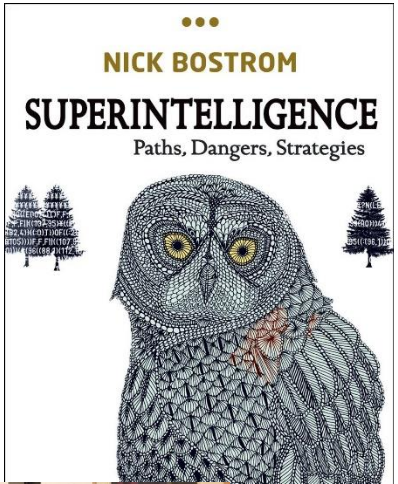
通往危险的人工智能的路径分类学

How and When did AI become Dangerous		External Causes			Internal Causes
		On Purpose	By Mistake	Environment	Independently
Timing	Pre-Deployment	a	c	e	g
	Post-Deployment	b	d	f	h

Roman V. Yampolskiy. Taxonomy of Pathways to Dangerous Artificial Intelligence. 30th AAAI Conference on Artificial Intelligence (AAAI-2016). 2nd International Workshop on AI, Ethics and Society (AIEthicsSociety2016). Phoenix, Arizona, USA. February 12-13th, 2016.

- **Deliberate actions** of not-so-ethical people (on purpose – a, b)
  - Hackers, criminals, military, corporations, governments, cults, psychopaths, etc.
- **Side effects** of poor design (engineering mistakes – c, d)
  - Bugs, misaligned values, bad data, wrong goals, etc.
- **Miscellaneous** cases, impact of the surroundings of the system (environment – e, f)
  - Soft errors, SETI
- **Runaway self-improvement** process (Independently – g, h)
  - Wireheading, Emergent Phenomena, “Treacherous Turn”
- 故意设计的危险AI同样有可能保护所有其他类型的安全问题，并将产生直接后果。它是最危险的AI，也最难防御。

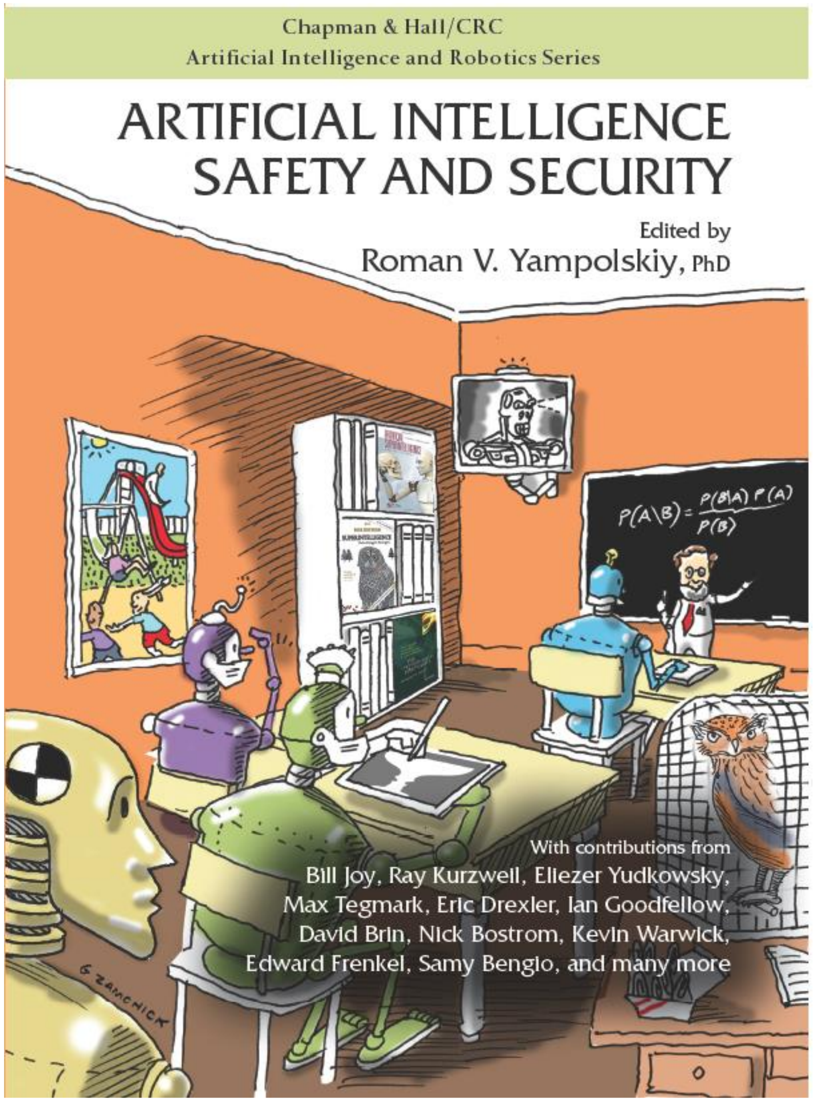
SuperBooks
超级图书



NICK BOSTROM

SUPERINTELLIGENCE

Paths, Dangers, Strategies

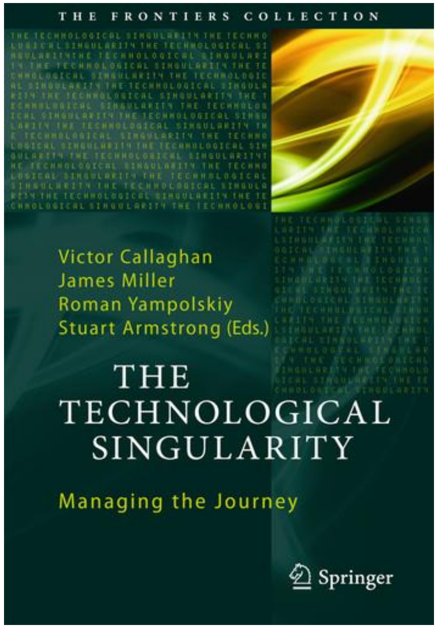


Chapman & Hall/CRC
Artificial Intelligence and Robotics Series

ARTIFICIAL INTELLIGENCE SAFETY AND SECURITY

Edited by
Roman V. Yampolskiy, PhD

With contributions from
Bill Joy, Ray Kurzweil, Eliezer Yudkowsky, Max Tegmark, Eric Drexler, Ian Goodfellow, David Brin, Nick Bostrom, Kevin Warwick, Edward Frenkel, Samy Bengio, and many more



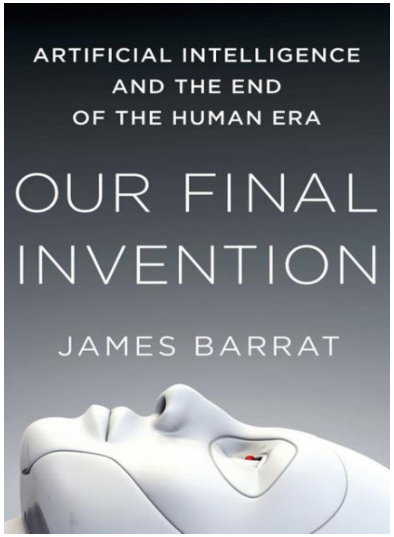
THE FRONTIERS COLLECTION

Victor Callaghan
James Miller
Roman Yampolskiy
Stuart Armstrong (Eds.)

THE TECHNOLOGICAL SINGULARITY

Managing the Journey

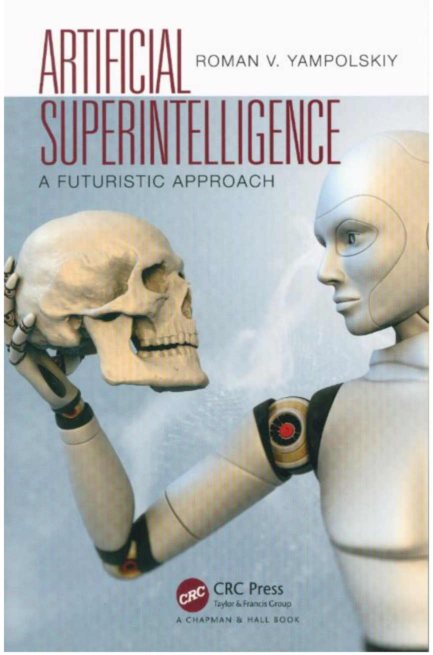
Springer



ARTIFICIAL INTELLIGENCE AND THE END OF THE HUMAN ERA

OUR FINAL INVENTION

JAMES BARRAT



ARTIFICIAL SUPERINTELLIGENCE

ROMAN V. YAMPOLSKIY

A FUTURISTIC APPROACH

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

SuperResponse 超级响应

MIRI
MACHINE INTELLIGENCE RESEARCH INSTITUTE

future of life INSTITUTE

CAMPAIGN TO STOP KILLER ROBOTS
StopKillerRobots.org

ICRAC
International Committee for Robot Arms Control

Future of Humanity Institute
UNIVERSITY OF OXFORD

Moral Responsibility for Computing Artifacts

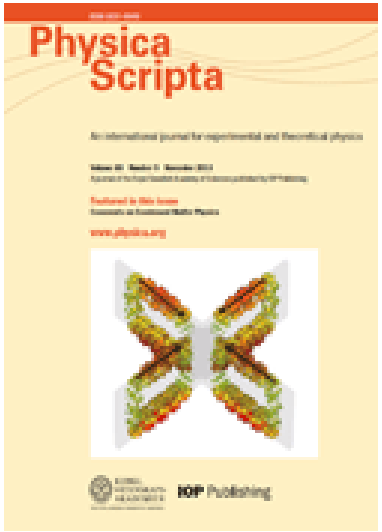
CENTRE FOR THE STUDY OF EXISTENTIAL RISK
UNIVERSITY OF CAMBRIDGE

SAVING HUMANITY from HOMO SAPIENS

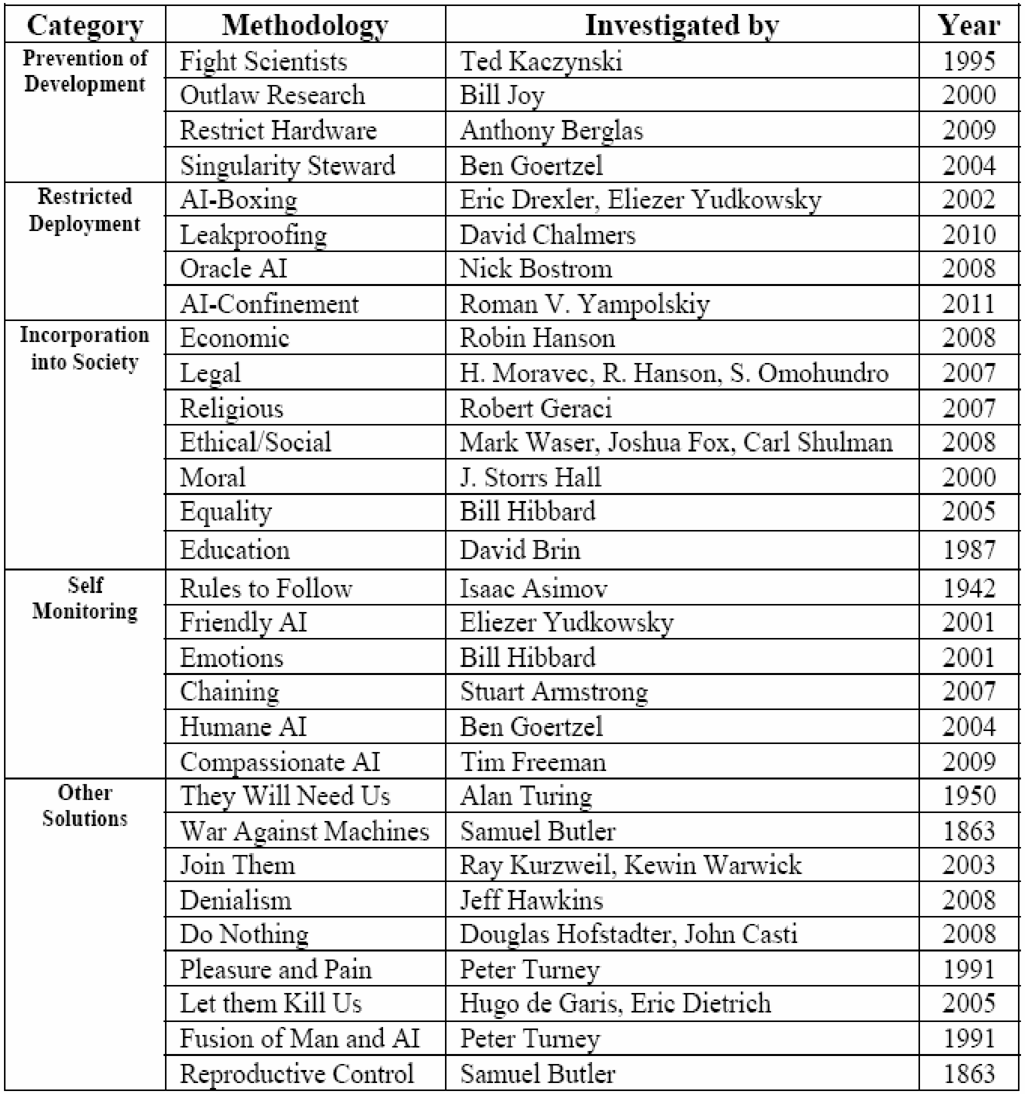
Global Catastrophic Risk

lifeboat foundation
SAFEGUARDING HUMANITY

Mitigating Negative Impact 缓解负面影响



Kaj Sotala and **Roman V. Yampolskiy**. Physica Scripta 90 (1)
http://iopscience.iop.org/1402-4896/90/1/018001/article

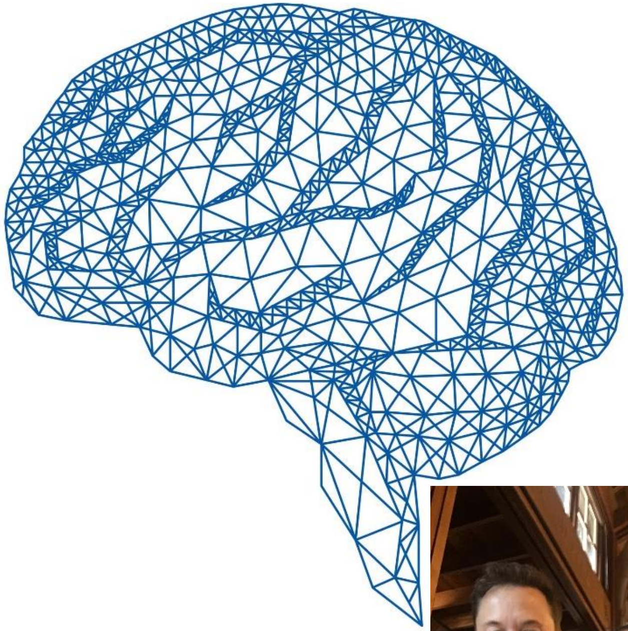


Category	Methodology	Investigated by	Year
Prevention of Development	Fight Scientists	Ted Kaczynski	1995
	Outlaw Research	Bill Joy	2000
	Restrict Hardware	Anthony Berglas	2009
	Singularity Steward	Ben Goertzel	2004
Restricted Deployment	AI-Boxing	Eric Drexler, Eliezer Yudkowsky	2002
	Leakproofing	David Chalmers	2010
	Oracle AI	Nick Bostrom	2008
	AI-Confinement	Roman V. Yampolskiy	2011
Incorporation into Society	Economic	Robin Hanson	2008
	Legal	H. Moravec, R. Hanson, S. Omohundro	2007
	Religious	Robert Geraci	2007
	Ethical/Social	Mark Waser, Joshua Fox, Carl Shulman	2008
	Moral	J. Storrs Hall	2000
	Equality	Bill Hibbard	2005
	Education	David Brin	1987
Self Monitoring	Rules to Follow	Isaac Asimov	1942
	Friendly AI	Eliezer Yudkowsky	2001
	Emotions	Bill Hibbard	2001
	Chaining	Stuart Armstrong	2007
	Humane AI	Ben Goertzel	2004
	Compassionate AI	Tim Freeman	2009
Other Solutions	They Will Need Us	Alan Turing	1950
	War Against Machines	Samuel Butler	1863
	Join Them	Ray Kurzweil, Kewin Warwick	2003
	Denialism	Jeff Hawkins	2008
	Do Nothing	Douglas Hofstadter, John Casti	2008
	Pleasure and Pain	Peter Turney	1991
	Let them Kill Us	Hugo de Garis, Eric Dietrich	2005
	Fusion of Man and AI	Peter Turney	1991
	Reproductive Control	Samuel Butler	1863

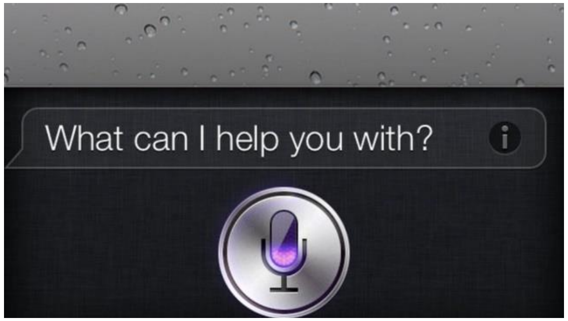
Responses to Catastrophic AGI Risk: A Survey

Enhance Human Capabilities, Uploads, Neural Lace

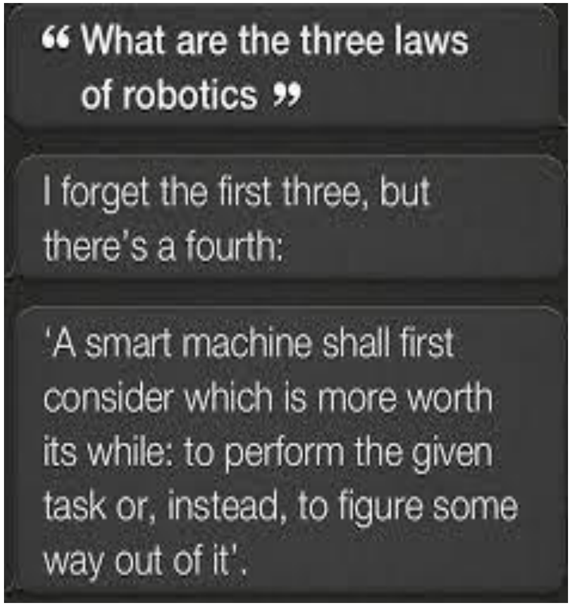
增强人的能力、上传、神经织网



Laws of Robotics 机器人法则



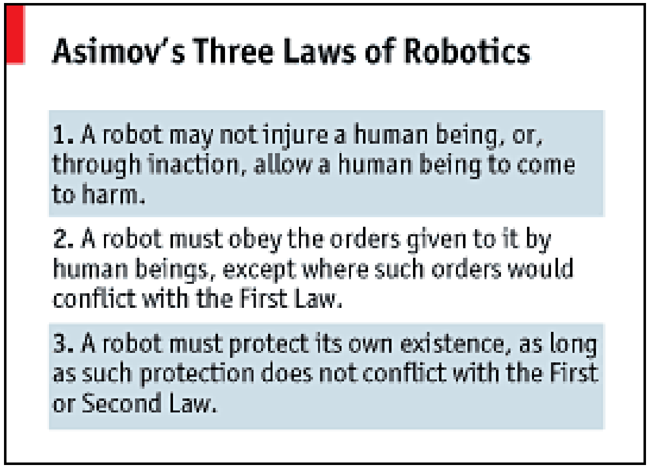
What can I help you with?



“ What are the three laws of robotics ”

I forget the first three, but there’s a fourth:

‘A smart machine shall first consider which is more worth its while: to perform the given task or, instead, to figure some way out of it’.

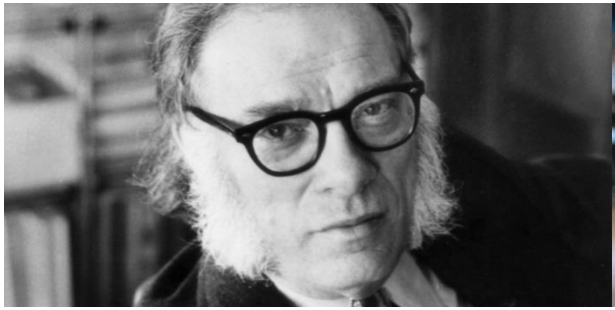


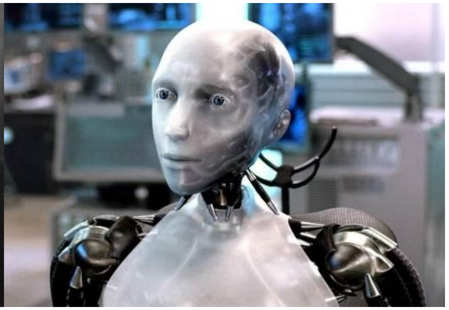
Asimov’s Three Laws of Robotics

1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.

2. A robot must obey the orders given to it by human beings, except where such orders would conflict with the First Law.

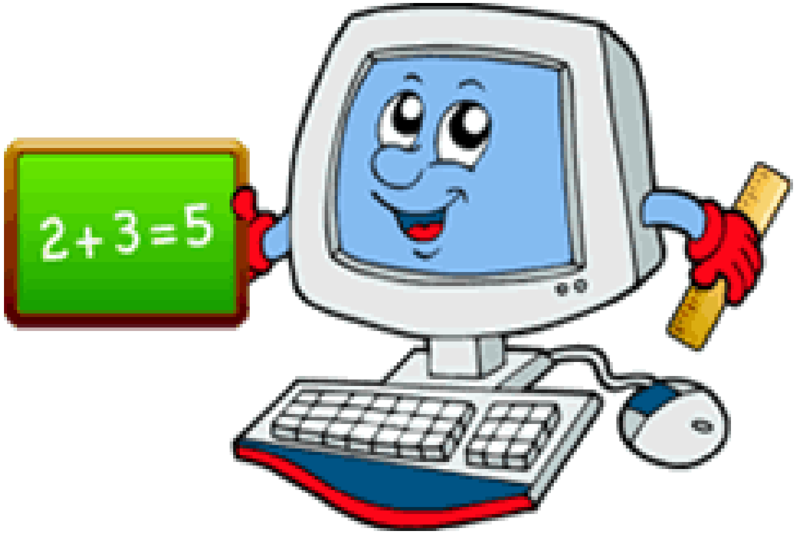
3. A robot must protect its own existence, as long as such protection does not conflict with the First or Second Law.



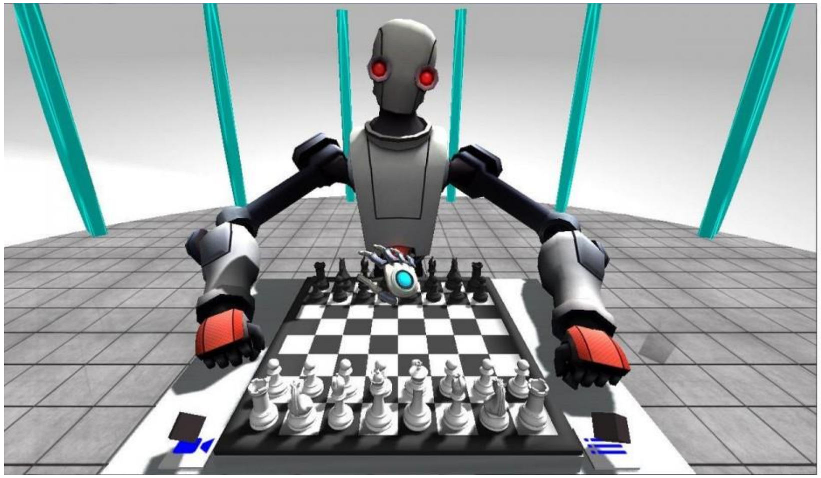


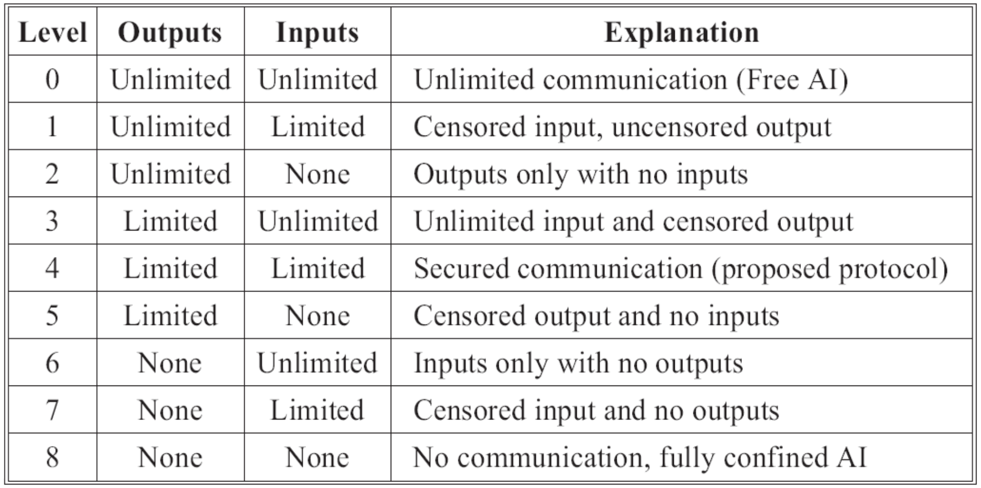
Formal Verification 正式验证

Have we built the software right?
Verification and Validation
Have we built the right software?



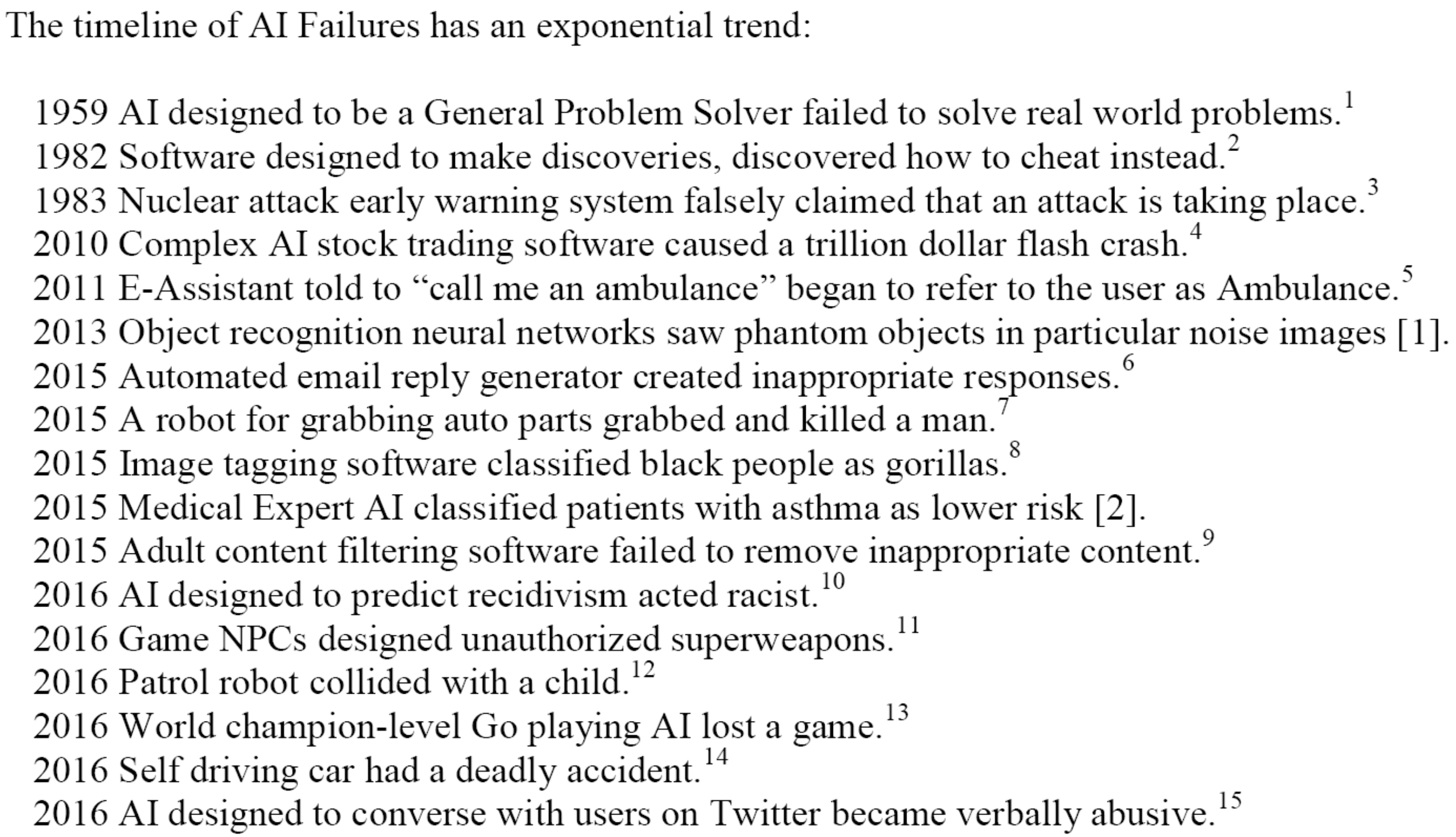
AI Confinement Problem 人工智能的限制问题





Level	Outputs	Inputs	Explanation
0	Unlimited	Unlimited	Unlimited communication (Free AI)
1	Unlimited	Limited	Censored input, uncensored output
2	Unlimited	None	Outputs only with no inputs
3	Limited	Unlimited	Unlimited input and censored output
4	Limited	Limited	Secured communication (proposed protocol)
5	Limited	None	Censored output and no inputs
6	None	Unlimited	Inputs only with no outputs
7	None	Limited	Censored input and no outputs
8	None	None	No communication, fully confined AI

Conclusions 结论



The timeline of AI Failures has an exponential trend:

1959 AI designed to be a General Problem Solver failed to solve real world problems.[1]
1982 Software designed to make discoveries, discovered how to cheat instead.[2]
1983 Nuclear attack early warning system falsely claimed that an attack is taking place.[3]
2010 Complex AI stock trading software caused a trillion dollar flash crash.[4]
2011 E-Assistant told to "call me an ambulance" began to refer to the user as Ambulance.[5]
2013 Object recognition neural networks saw phantom objects in particular noise images [1].
2015 Automated email reply generator created inappropriate responses.[6]
2015 A robot for grabbing auto parts grabbed and killed a man.[7]
2015 Image tagging software classified black people as gorillas.[8]
2015 Medical Expert AI classified patients with asthma as lower risk [2].
2015 Adult content filtering software failed to remove inappropriate content.[9]
2016 AI designed to predict recidivism acted racist.[10]
2016 Game NPCs designed unauthorized superweapons.[11]
2016 Patrol robot collided with a child.[12]
2016 World champion-level Go playing AI lost a game.[13]
2016 Self driving car had a deadly accident.[14]
2016 AI designed to converse with users on Twitter became verbally abusive.[15]

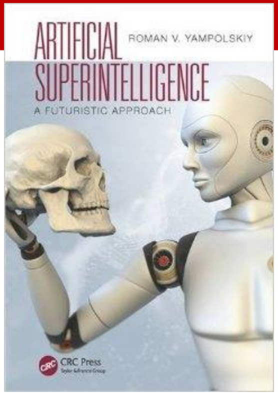
AI failures will grow in frequency and severity proportionate to AI's capability.

人工智能失败的频率和严重程度都会增加，这与人工智能的能力成比例。

References

can be found in …

参考资料



- Federico Pistono, Roman V. Yampolskiy. Unethical Research: How to Create a Malevolent Artificial Intelligence. 25th International Joint Conference on Artificial Intelligence (IJCAI-16). Ethics for Artificial Intelligence Workshop (AI-Ethics-2016). New York, NY. July 9 – 15, 2016.
- James Babcock, Janos Kramar, Roman Yampolskiy. The AGI Containment Problem. The Ninth Conference on Artificial General Intelligence (AGI2015). NYC, USA. Pei Wang & Bas Steunebrink (Ed.), Lecture Notes in Computer Science, Volume 9782, pp. 53-63. Springer. July 16-19, 2016.
- Roman V. Yampolskiy. Taxonomy of Pathways to Dangerous Artificial Intelligence. 30th AAAI Conference on Artificial Intelligence (AAAI-2016). 2nd International Workshop on AI, Ethics and Society (AIEthicsSociety2016). Phoenix, Arizona, USA. February 12-13th, 2016.
- Roman V. Yampolskiy and M. Spellchecker (2016). "Artificial Intelligence Safety and Cybersecurity: a Timeline of AI Failures." arXiv preprint arXiv:1610.07997.
- Kaj Sotala, Roman V. Yampolskiy. Responses to Catastrophic AGI risk: A Survey. *Physica Scripta*. Volume 90, Number 1. January 2015. pp. 1-33.
- Roman V. Yampolskiy. Utility Function Security in Artificially Intelligent Agents. *Journal of Experimental and Theoretical Artificial Intelligence (JETAI)*. 2014.
- Roman V. Yampolskiy. Turing Test as a Defining Feature of AI-Completeness. In Artificial Intelligence, Evolutionary Computation and Metaheuristics (AIECM) --In the footsteps of Alan Turing. Xin-She Yang (Ed.). pp. 3-17. (Chapter 1). Springer, London. 2013.

UNIVERSITY OF LOUISVILLE

The End!

Roman.Yampolskiy@louisville.edu

Director, CyberSecurity Lab

Computer Engineering and Computer Science

University of Louisville - cecs.louisville.edu/ry

twitter **@romanyam**

Follow me on Facebook **/Roman.Yampolskiy**

All images used in this presentation are copyrighted to their respective owners and are used for educational purposes only.