

UNIVERSITY OF
LOUISVILLE

World
Summit 



AI: Unexplainable, Unpredictable, Uncontrollable



Dr. Roman.Yampolskiy@louisville.edu

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

 [@romanyam](https://twitter.com/romanyam)



Follow me on
Facebook

[/roman.yampolskiy](https://www.facebook.com/roman.yampolskiy)



GPT-4 Technical Report

OpenAI*

Abstract

We report the development of GPT-4, a large-scale, multimodal model which can accept image and text inputs and produce text outputs. While less capable than humans in many real-world scenarios, GPT-4 exhibits human-level performance on various professional and academic benchmarks, including passing a simulated bar exam with a score around the top 10% of test takers. GPT-4 is a Transformer-based model pre-trained to predict the next token in a document. The post-training alignment process results in improved performance on measures of factuality and adherence to desired behavior. A core component of this project was developing infrastructure and optimization methods that behave predictably across a wide range of scales. This allowed us to accurately predict some aspects of GPT-4's performance based on models trained with no more than 1/1,000th the compute of GPT-4.

Exam	GPT-4	GPT-4 (no vision)	GPT-3.5
Uniform Bar Exam (MBE+MEE+MPT)	298 / 400 (~90th)	298 / 400 (~90th)	213 / 400 (~10th)
LSAT	163 (~88th)	161 (~83rd)	149 (~40th)
SAT Evidence-Based Reading & Writing	710 / 800 (~93rd)	710 / 800 (~93rd)	670 / 800 (~87th)
SAT Math	700 / 800 (~89th)	690 / 800 (~89th)	590 / 800 (~70th)
Graduate Record Examination (GRE) Quantitative	163 / 170 (~80th)	157 / 170 (~62nd)	147 / 170 (~25th)
Graduate Record Examination (GRE) Verbal	169 / 170 (~99th)	165 / 170 (~96th)	154 / 170 (~63rd)
Graduate Record Examination (GRE) Writing	4 / 6 (~54th)	4 / 6 (~54th)	4 / 6 (~54th)
USABO Semifinal Exam 2020	87 / 150 (99th - 100th)	87 / 150 (99th - 100th)	43 / 150 (31st - 33rd)
USNCO Local Section Exam 2022	36 / 60	38 / 60	24 / 60
Medical Knowledge Self-Assessment Program	75 %	75 %	53 %
Codeforces Rating	392 (below 5th)	392 (below 5th)	260 (below 5th)
AP Art History	5 (86th - 100th)	5 (86th - 100th)	5 (86th - 100th)
AP Biology	5 (85th - 100th)	5 (85th - 100th)	4 (62nd - 85th)
AP Calculus BC	4 (43rd - 59th)	4 (43rd - 59th)	1 (0th - 7th)
AP Chemistry	4 (71st - 88th)	4 (71st - 88th)	2 (22nd - 46th)
AP English Language and Composition	2 (14th - 44th)	2 (14th - 44th)	2 (14th - 44th)
AP English Literature and Composition	2 (8th - 22nd)	2 (8th - 22nd)	2 (8th - 22nd)
AP Environmental Science	5 (91st - 100th)	5 (91st - 100th)	5 (91st - 100th)
AP Macroeconomics	5 (84th - 100th)	5 (84th - 100th)	2 (33rd - 48th)
AP Microeconomics	5 (82nd - 100th)	4 (60th - 82nd)	4 (60th - 82nd)
AP Physics 2	4 (66th - 84th)	4 (66th - 84th)	3 (30th - 66th)
AP Psychology	5 (83rd - 100th)	5 (83rd - 100th)	5 (83rd - 100th)
AP Statistics	5 (85th - 100th)	5 (85th - 100th)	3 (40th - 63rd)
AP US Government	5 (88th - 100th)	5 (88th - 100th)	4 (77th - 88th)
AP US History	5 (89th - 100th)	4 (74th - 89th)	4 (74th - 89th)
AP World History	4 (65th - 87th)	4 (65th - 87th)	4 (65th - 87th)
AMC 10 ³	30 / 150 (6th - 12th)	36 / 150 (10th - 19th)	36 / 150 (10th - 19th)
AMC 12 ³	60 / 150 (45th - 66th)	48 / 150 (19th - 40th)	30 / 150 (4th - 8th)
Introductory Sommelier (theory knowledge)	92 %	92 %	80 %
Certified Sommelier (theory knowledge)	86 %	86 %	58 %
Advanced Sommelier (theory knowledge)	77 %	77 %	46 %
Leetcode (easy)	31 / 41	31 / 41	12 / 41
Leetcode (medium)	21 / 80	21 / 80	8 / 80
Leetcode (hard)	3 / 45	3 / 45	0 / 45

Table 1. GPT performance on academic and professional exams. In each case, we simulate the conditions and scoring of the real exam. We report GPT-4's final score graded according to exam-specific rubrics, as well as the percentile of test-takers achieving GPT-4's score.

CEO of Google's DeepMind says we could be 'just a few years' from A.I. that has human-level intelligence

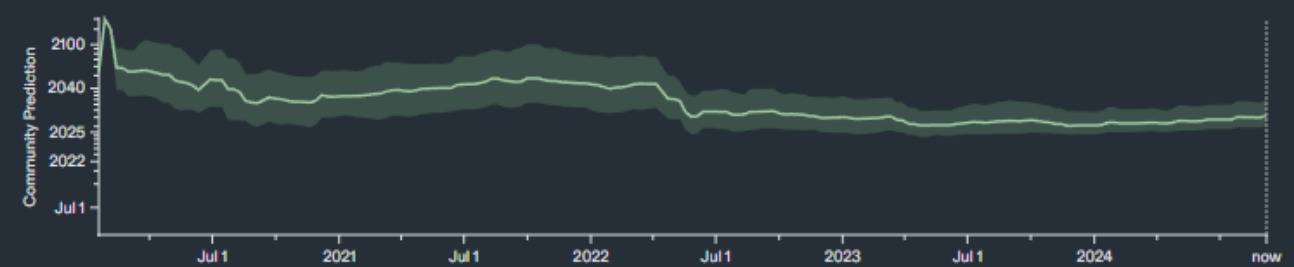
BY TRISTAN BOVE
May 3, 2023 at 5:32 PM EDT



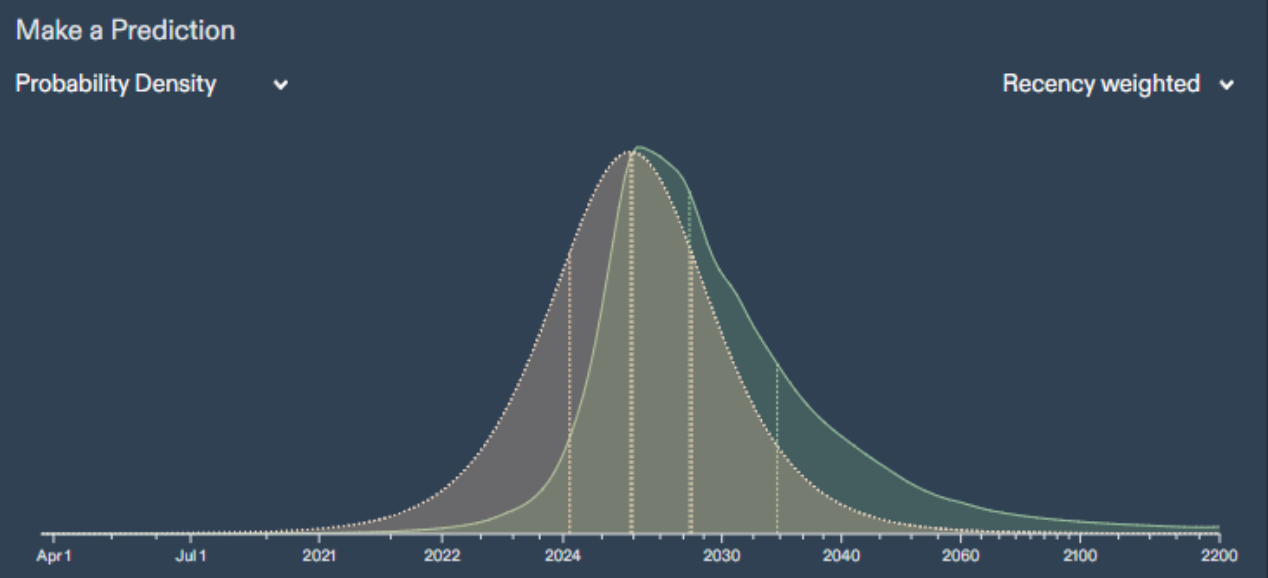
When will the first weakly general AI system be devised, tested, and publicly announced?

Jun 8, 2028
4.30k predictions

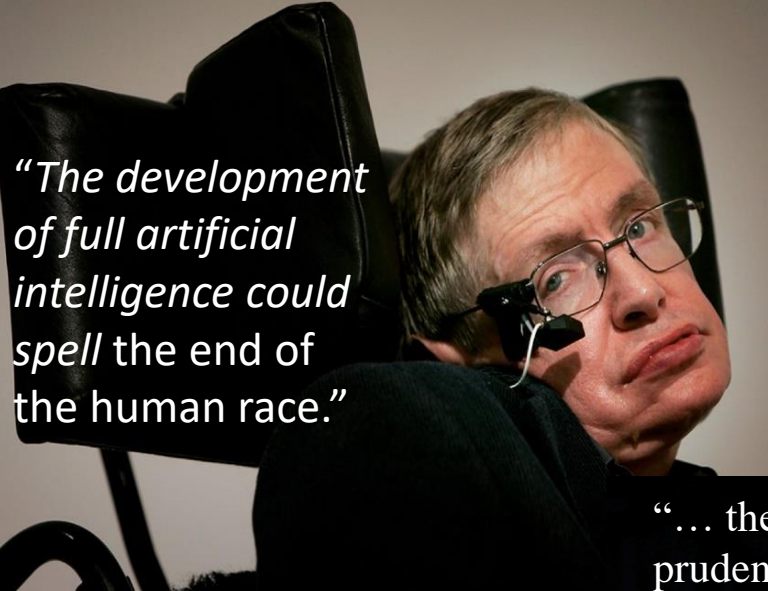
196 Closes Jan 1, 2200 427 comments



Total Forecasters 1.47k
Community Prediction 2028



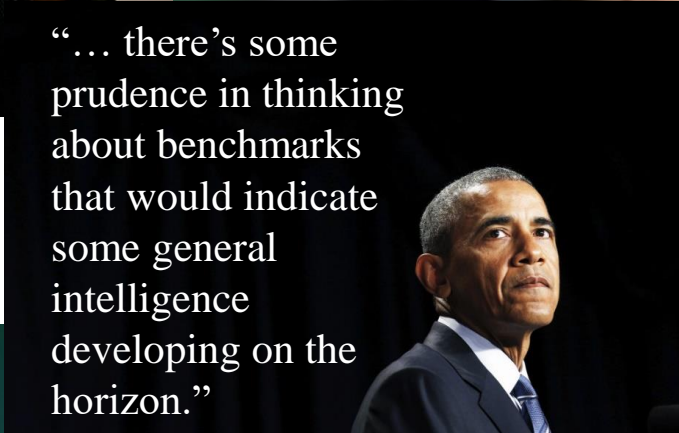
"The development of full artificial intelligence could spell the end of the human race."



"I think we should be very careful about artificial intelligence"

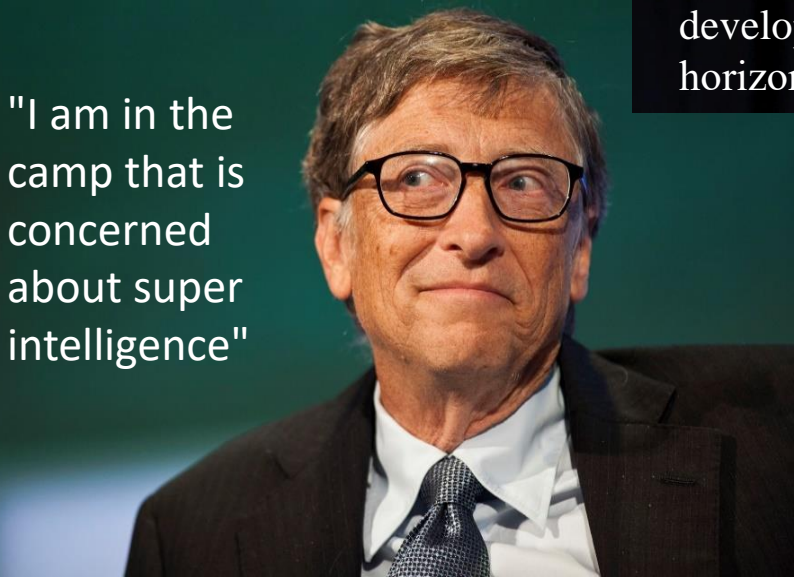


"... there's some prudence in thinking about benchmarks that would indicate some general intelligence developing on the horizon."

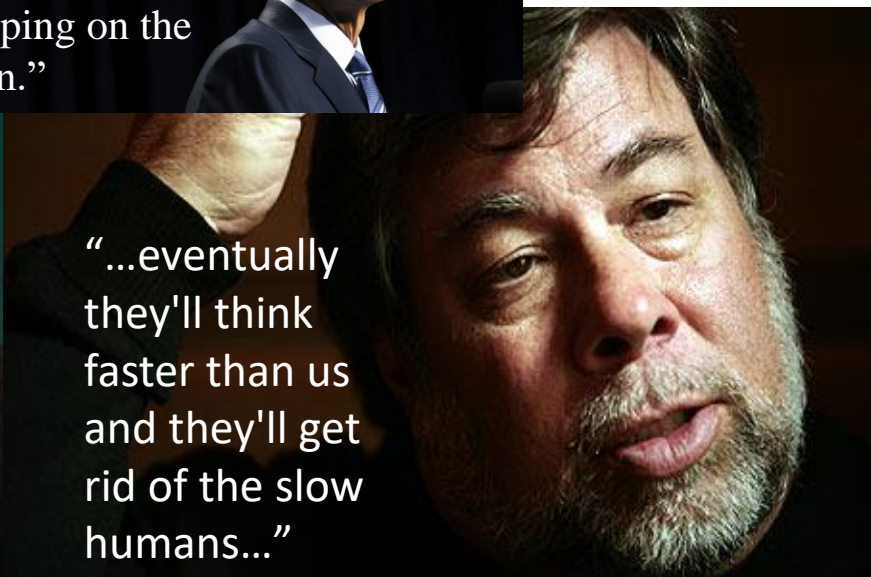


Concerns About A(G)I

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



"...eventually they'll think faster than us and they'll get rid of the slow humans..."



Top AI Scientists Warn: Risk of Extinction from AI on Scale with Nuclear War

San Francisco, CA – Distinguished AI scientists, including Turing Award winners Geoffrey Hinton and Yoshua Bengio, and leaders of the major AI labs, including Sam Altman of OpenAI and Demis Hassabis of Google DeepMind, have signed a [single-sentence statement](#) from the [Center for AI Safety](#) that reads:

"Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."

This represents a historic coalition of AI experts — along with philosophers, ethicists, legal scholars, economists, physicists, political scientists, pandemic scientists, nuclear scientists, and climate scientists — establishing the risk of extinction from advanced, future AI systems as one of the world's most important problems. The statement affirms growing public sentiment: a recent poll [found that 61 percent of Americans](#) believe AI threatens humanity's future.

AI Control Problem

How can humanity remain safely in control while benefiting from a superior form of intelligence?

Is the AI control problem:
Solvable?
Partially Solvable?
Unsolvable?
Undecidable?



Definitely Solvable, Very Tractable, No Idea

Eliezer Yudkowsky @ESYudkowsky · May 3
 So now that the godfather of deep learning is metaphorically buying out my prediction market position, everyone who claimed I had no right to speak because I'd trained insufficiently large neural nets myself is going to issue an apology and retraction, right?

Siméon @Simeon_Cps · May 3
 40 seconds to understand why the human species might go extinct and why if we race towards AGI, everyone will lose, by @geoffreyhinton
[Show this thread](#)



162 186 1,525 676K

Dr. Roman Yampolskiy @romanyam · May 3
 Notice, he says that the problem may be unsolvable. A position which is not given enough weight by the safety community.

4 2 32 11.6K

Eliezer Yudkowsky @ESYudkowsky · May 3
 He's new to the field and hasn't had much time to think about it yet. It's definitely solvable, just not in time on the first try the way we're doing it.

4 34 3,869

Jan Leike @janleike
 The alignment problem is very tractable.
 We haven't figured out how to solve it yet, but with focus and dedication we will.
 4:22 PM · May 18, 2023 · 316.1K Views

Two questions

- Will artificial neural networks soon be more intelligent than real neural networks?
 - The talk will be almost entirely about this question.
- Will people be able to stay in control of super-intelligent AI?
 - I will make a few speculations about how we could lose control.
 - I do not have any good ideas about how to prevent this happening.



Tools for Controllability

- Explainability
- Comprehensibility
- Predictability
- Verifiability
- Many more!



Journal of Artificial Intelligence and Consciousness
Vol. 7, No. 2 (2020) 277–291
© World Scientific Publishing Company
DOI: [10.1142/S2705078520500150](https://doi.org/10.1142/S2705078520500150)



Unexplainability and Incomprehensibility of AI

Roman V. Yampolskiy

*Computer Science and Engineering, University of Louisville
222 Eastern Parkway, Duthie Center, 215
Louisville, KY 40292, USA
roman.yampolskiy@louisville.edu*

Published 17 July 2020

Explainability and comprehensibility of AI are important requirements for intelligent systems deployed in real-world domains. Users want and frequently need to understand how decisions impacting them are made. Similarly, it is important to understand how an intelligent system functions for safety and security reasons. In this paper, we describe two complementary impossibility results (Unexplainability and Incomprehensibility), essentially showing that advanced AIs would not be able to accurately explain some of their decisions and for the decisions they could explain people would not understand some of those explanations.

Keywords: AI Safety; Black Box; Comprehensible; Explainable AI; Impossibility; Intelligible; Interpretability; Transparency; Understandable; Unsurveyability.

Journal of Artificial Intelligence and Consciousness
Vol. 7, No. 1 (2020) 109–118
© World Scientific Publishing Company
DOI: [10.1142/S2705078520500034](https://doi.org/10.1142/S2705078520500034)



Unpredictability of AI: On the Impossibility of Accurately Predicting All Actions of a Smarter Agent

Roman V. Yampolskiy

*Computer Science and Engineering, University of Louisville
222 Eastern Parkway, Duthie Center, 215 Louisville
KY 40292, USA
roman.yampolskiy@louisville.edu*


Published 29 April 2020

The young field of AI Safety is still in the process of identifying its challenges and limitations. In this paper, we formally describe one such impossibility result, namely Unpredictability of AI. We prove that it is impossible to precisely and consistently predict what specific actions a smarter-than-human intelligent system will take to achieve its objectives, even if we know the terminal goals of the system. In conclusion, the impact of Unpredictability on AI Safety is discussed.

Keywords: AI Safety; Impossibility; Uncontainability; Unpredictability; Unknowability.

Invited Comment

What are the ultimate limits to computational techniques: verifier theory and unverifiability

Roman V Yampolskiy 

Computer Engineering and Computer Science, University of Louisville, KY, United States of America

E-mail: roman.yampolskiy@louisville.edu

Received 25 October 2016, revised 17 May 2017

Accepted for publication 30 June 2017

Published 28 July 2017



CrossMark

Abstract

Despite significant developments in proof theory, surprisingly little attention has been devoted to the concept of proof verifiers. In particular, the mathematical community may be interested in studying different types of proof verifiers (people, programs, oracles, communities, superintelligences) as mathematical objects. Such an effort could reveal their properties, their powers and limitations (particularly in human mathematicians), minimum and maximum complexity, as well as self-verification and self-reference issues. We propose an initial classification system for verifiers and provide some rudimentary analysis of solved and open problems in this important domain. Our main contribution is a formal introduction of the notion of unverifiability, for which the paper could serve as a general citation in domains of theorem proving, as well as software and AI verification.

Keywords: verifier theory, proof theory, observer, verified verifier, verifiability

Impossibility Results in AI: A Survey

MARIO BRCIC, University of Zagreb Faculty of Electrical Engineering and Computing, Croatia

ROMAN V. YAMPOLSKIY, University of Louisville, USA

An impossibility theorem demonstrates that a particular problem or set of problems cannot be solved as described in the claim. Such theorems put limits on what is possible to do concerning artificial intelligence, especially the super-intelligent one. As such, these results serve as guidelines, reminders, and warnings to AI safety, AI policy, and governance researchers. These might enable solutions to some long-standing questions in the form of formalizing theories in the framework of constraint satisfaction without committing to one option. We strongly believe this to be the most prudent approach to long-term AI safety initiatives. In this article, we have categorized impossibility theorems applicable to AI into five mechanism-based categories: Deduction, indistinguishability, induction, tradeoffs, and intractability. We found that certain theorems are too specific or have implicit assumptions that limit application. Also, we added new results (theorems) such as the unfairness of explainability, the first explainability-related result in the induction category. The remaining results deal with misalignment between the clones and put a limit to the self-awareness of agents. We concluded that deductive impossibilities deny 100%-guarantees for security. In the end, we give some ideas that hold potential in explainability, controllability, value alignment, ethics, and group decision-making.

CCS Concepts: • **Computing methodologies** → **Artificial intelligence; Philosophical/theoretical foundations of artificial intelligence**; • **Security and privacy** → *Social aspects of security and privacy; Human and societal aspects of security and privacy*;

Additional Key Words and Phrases: Artificial intelligence, AI safety, limitations, impossibility theorems

ACM Reference format:

Mario Brcic and Roman V. Yampolskiy. 2023. Impossibility Results in AI: A Survey. *ACM Comput. Surv.* 56, 1, Article 8 (August 2023), 24 pages.

<https://doi.org/10.1145/3603371>

On the Controllability of Artificial Intelligence: An Analysis of Limitations

Roman V. Yampolskiy

University of Louisville, USA
E-mail: roman.yampolskiy@louisville.edu

Received 09 March 2022; Accepted 31 Mar
Publication 24 May 2022



Abstract

The invention of artificial general intelligence is predicted to cause a shift in the trajectory of human civilization. In order to reap the benefits and avoid the pitfalls of such a powerful technology it is important to be able to control it. However, the possibility of controlling artificial general intelligence and its more advanced version, superintelligence, has not been formally established. In this paper, we present arguments as well as supporting evidence from multiple domains indicating that advanced AI cannot be fully controlled. The consequences of uncontrollability of AI are discussed with respect to the future of humanity and research on AI, and AI safety and security.

Keywords: AI safety, control problem, safer AI, uncontrollability, unverifiability, X-risk.

1 Introduction

The unprecedented progress in artificial intelligence (AI) [1–6], over the last decade, came alongside multiple AI failures [7, 8] and cases of dual use [9] causing a realization [10] that it is not sufficient to create highly capable machines, but that it is even more important to make sure that intelligent



TIME

SUBSCRIBE

IDEAS • TECHNOLOGY

Why Uncontrollable AI Looks More Likely Than Ever

BY OTTO BARTEN AND ROMAN YAMPOLSKIY

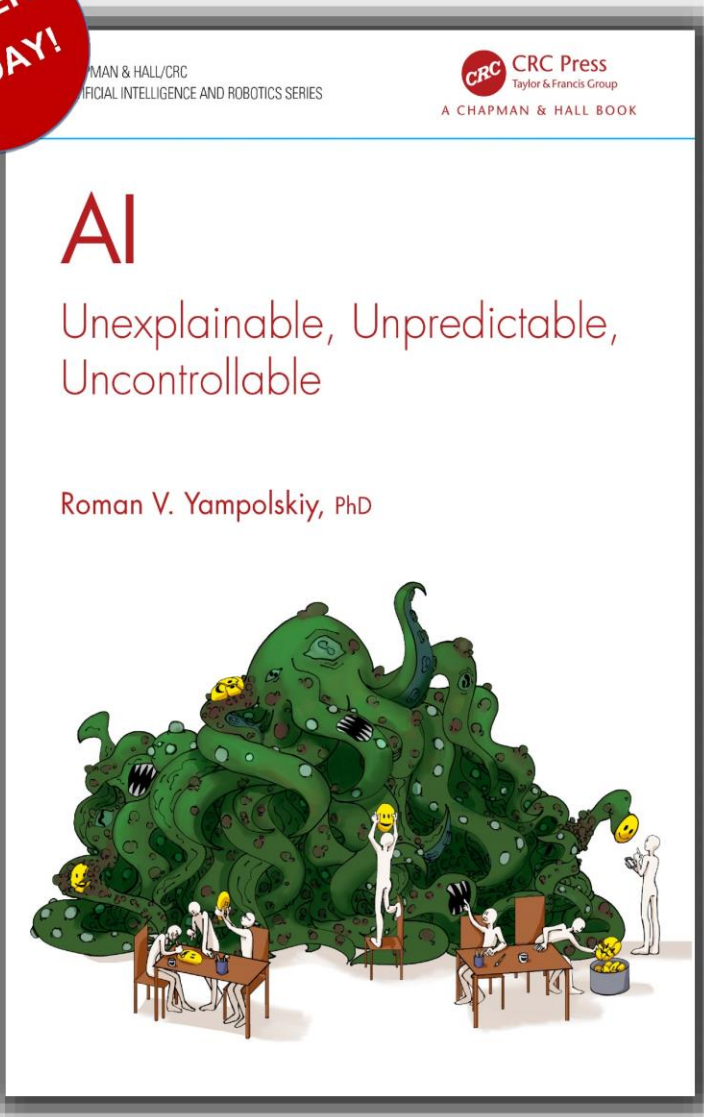
FEBRUARY 27, 2023 2:27 PM EST



Barten is director of the *Existential Risk Observatory*, an Amsterdam-based nonprofit.

Yampolskiy is a computer scientist at the University of Louisville, known for his *work* on AI Safety.

ORDER TODAY!



AI: Unexplainable, Unpredictable, Uncontrollable (Chapman & Hall/CRC Artificial Intelligence and Robotics Series) 1st Edition



by Roman V. Yampolskiy (Author)

5.0 ★★★★★ 10 ratings

Part of: [Chapman & Hall/CRC Artificial Intelligence and Robotics Series \(35 books\)](#)

[See all formats and editions](#)

Delving into the deeply enigmatic nature of Artificial Intelligence (AI), **AI: Unexplainable, Unpredictable, Uncontrollable** explores the various reasons why the field is so challenging. Written by one of the founders of the field of AI safety, this book addresses some of the most fascinating questions facing humanity, including the nature of intelligence, consciousness, values and knowledge.

Moving from a broad introduction to the core problems, such as the unpredictability of AI outcomes or the difficulty in explaining AI decisions, this book arrives at more complex questions of ownership and control, conducting an in-depth analysis of potential hazards and unintentional consequences. The book then concludes with philosophical and existential considerations, probing into questions of AI personhood, consciousness, and the distinction between human intelligence and artificial general intelligence (AGI).

Bridging the gap between technical intricacies and philosophical musings, **AI: Unexplainable, Unpredictable, Uncontrollable** appeals to both AI experts and enthusiasts looking for a comprehensive understanding of the field, whilst also being written for a general audience with minimal technical jargon.

Review

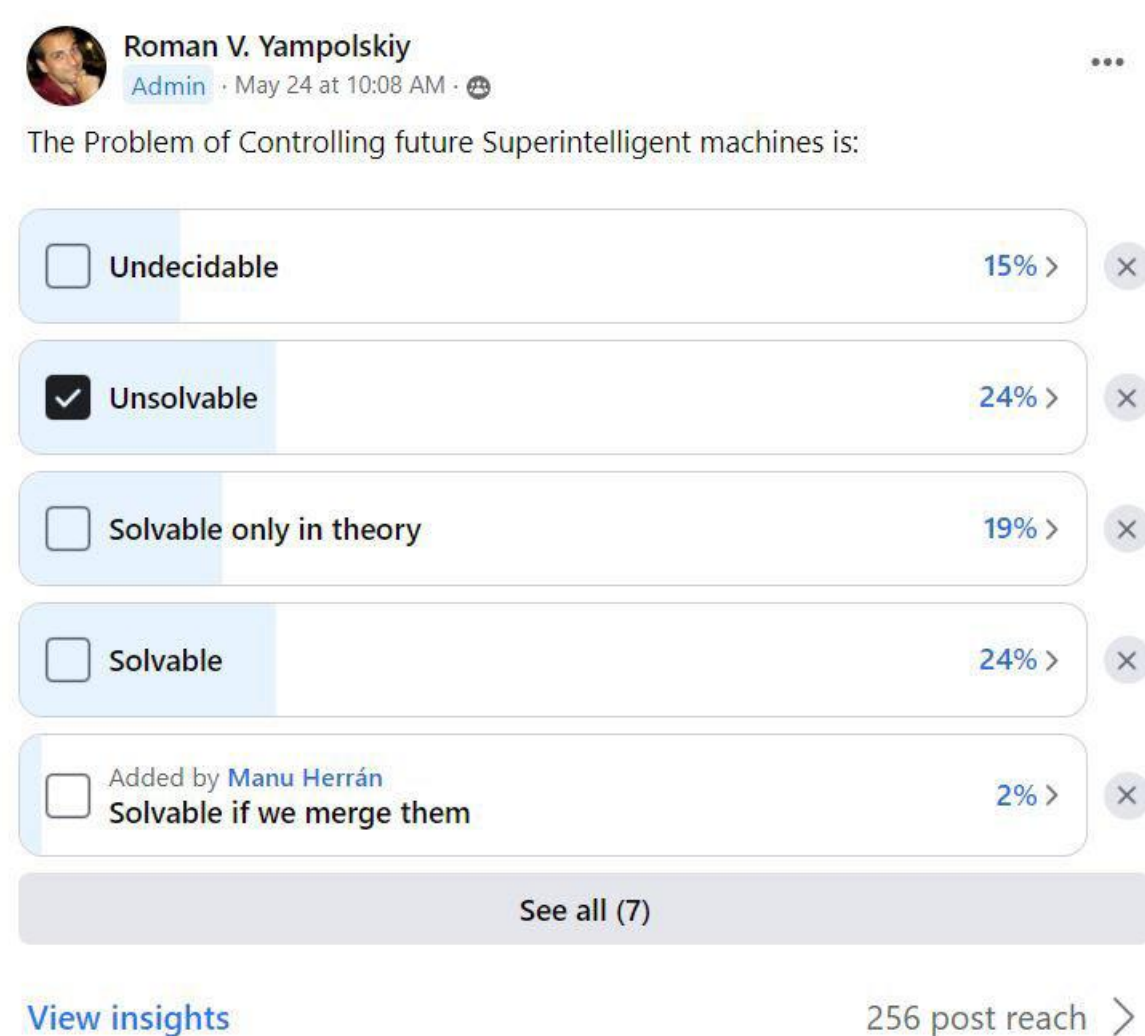
This is a captivating and thought-provoking book about the most pressing issue of our time: should humanity risk committing collective suicide by rushing to build and unleash a new smarter-than-human species that we can neither understand nor control?

Prof. Max Tegmark, MIT, author of "Life 3.0"

Use code **CSDC24** for 20% Discount



Opinion Polls – The Problem is:



- Small sample size (n=137 tw; n=55 fb).
- Mix of experts and nonexperts.

List of p(doom) values

p(doom) is the probability of very bad outcomes (e.g. human extinction) as a result of AI. This most often refers to the likelihood of [AI taking over](#) from humanity, but different scenarios can also constitute "doom". For example, a large portion of the population dying due to a novel biological weapon created by AI, social collapse due to a [large-scale cyber attack](#), or AI causing a nuclear war. Note that not everyone is using the same definition when talking about their p(doom) values. Most notably the time horizon is often not specified, which makes comparing a bit difficult.

Press the p(doom) percentage to open the source.

99.999999% [Roman Yampolskiy](#) AI safety scientist

>95% [Eliezer Yudkowsky](#) Founder of MIRI

>80% [Dan Hendrycks](#) Head of Center for AI Safety

70% [Daniel Kokotajlo](#) Forecaster & former OpenAI researcher

60% [Zvi Mowshowitz](#) Independent AI safety journalist

10-90% [Holden Karnofsky](#) Co-founder of Open Philanthropy

10-90% [Jan Leike](#) Former alignment lead at OpenAI

46% [Paul Christiano](#) Head of AI safety, US AI Safety Institute, formerly OpenAI, founded ARC

40% [AI engineer](#)
(Estimate mean value, survey methodology may be flawed)

40% [Joep Meindertsma](#) Founder of PauseAI
(The remaining 60% consists largely of "we can pause".)

40% [AI engineer](#)
(Estimate mean value, survey methodology may be flawed)

40% [Joep Meindertsma](#) Founder of PauseAI
(The remaining 60% consists largely of "we can pause".)

35% [Eli Lifland](#) Top competitive forecaster

33% [Scott Alexander](#) Popular Internet blogger at Astral Codex Ten

30% [AI Safety Researchers](#)
(Mean from 44 AI safety researchers in 2021)

10-50% [Geoff Hinton](#) one of three godfathers of AI
(Recently said "Kinda 50-50" on good outcomes for humanity. Earlier he mentioned 10%.)

5-50% [Emmett Shear](#) Co-founder of Twitch, former interim CEO of OpenAI

20% [Reid Hoffman](#) Co-founder of LinkedIn

20% [Yoshua Bengio](#) one of three godfathers of AI

10-25% [Dario Amodei](#) CEO of Anthropic

15% [Lina Khan](#) head of FTC

10-20% [Elon Musk](#) CEO of Tesla, SpaceX, X

9-19.4% [Machine learning researchers](#)
(Mean in 2023, depending on the question design, median values: 5-10%)

10% [Vitalik Buterin](#) Ethereum founder

0.38% [Forecasting Research Institute Superforecasters](#)
(From the same study: Domain experts estimated 3% AI x-risk, and AI catastrophe at 12%)

<0.01% [Yann LeCun](#) one of three godfathers of AI, works at Meta
(less likely than an asteroid)

Perpetual Safety Machine

Cybersecurity or Narrow AI Safety	Superintelligence Safety
<p>Many chances to fix problems. (reissue credit cards, change passwords, etc.) Limited damages (financial loss, privacy loss). 99.9999% safe is good enough. Eventually sufficiently debugged.</p>	<p>Only 1 chance to get it right. Unlimited damages (X-risk, S-risk). <100% Safe is not good enough, but 100% is impossible. Doesn't stop changing (learning, self-modifying, etc.).</p>

Solvable.



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

