Future with AI

Roman Yampolskiy

Director – CyberSecurity Lab

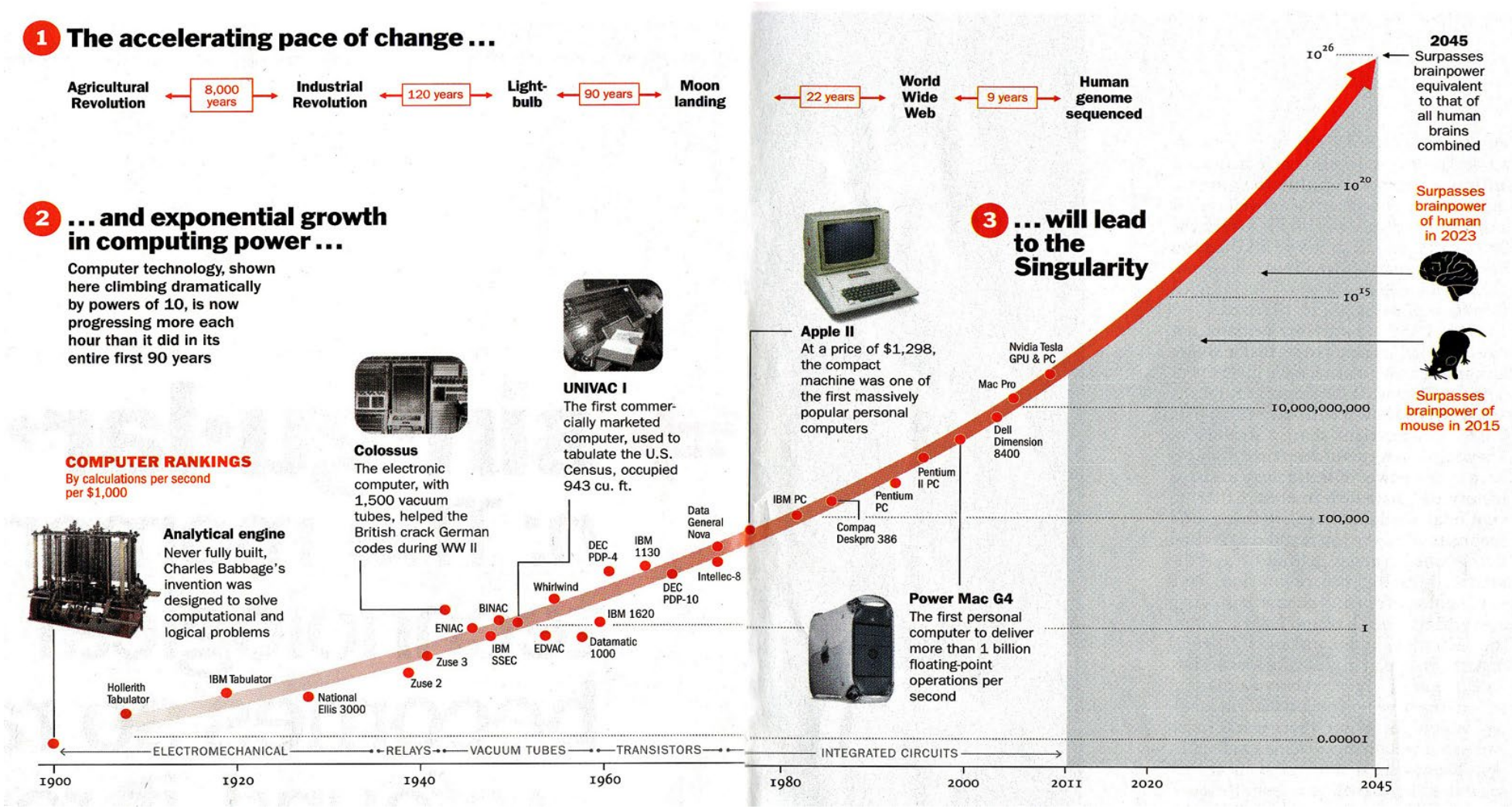
twitter **@romanyam**

Follow me on Facebook **/roman.yampolskiy**

Disclaimer

- Speaking as a Kentuckian.
- Sharing my opinion, not that of my employer.
- Bipartisan issue, the goal is to inform.
- Not a lobbyist.
- For information purposes only.

Past



- Raymond Kurzweil https://content.time.com/time/interactive/0,31813,2048601,00.html
- 2023-2045+

Present

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.

https://arxiv.org/pdf/2303.08774.pdf

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.

Future

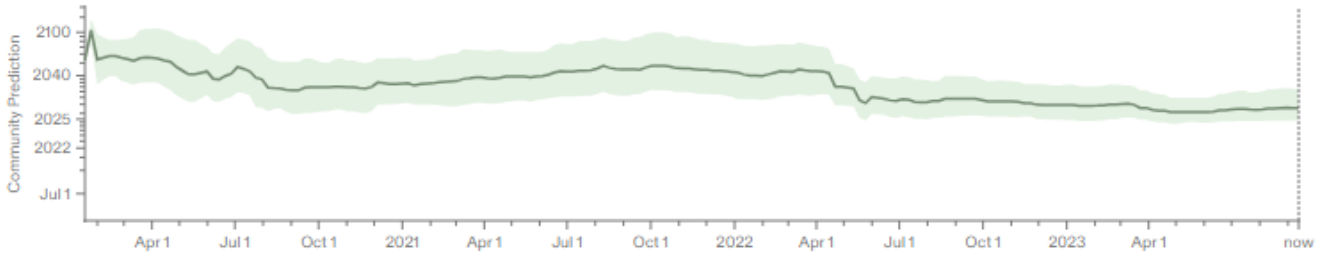
CEO of Google's DeepMind says we could be 'just a few years' from A.I. that has human-level intelligence

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?

Feb 9, 2027
2.97k predictions

161 | Closes Jan 1, 2200 | 296 comments



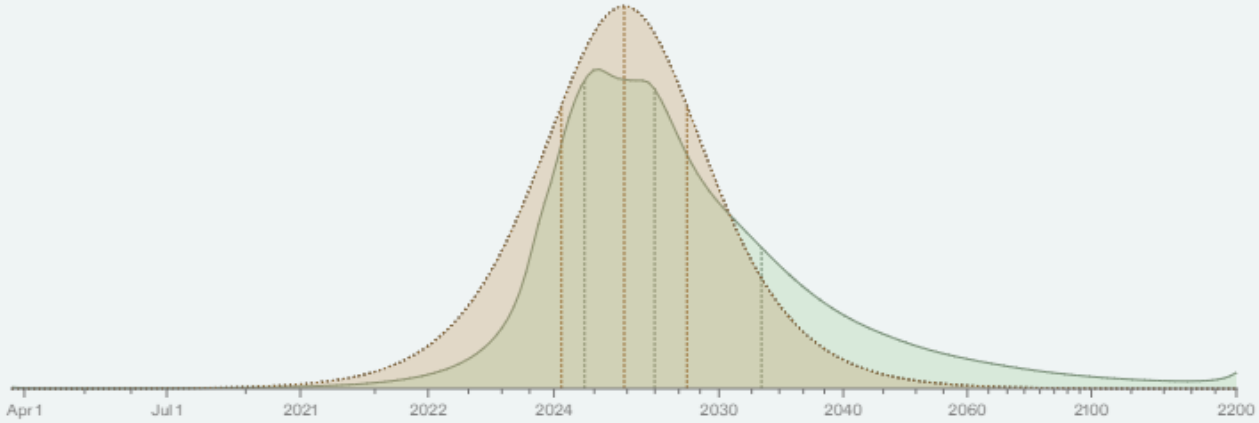
Total Forecasters
927

Community Prediction
Feb 9, 2027

Make a Prediction

Probability Density

Recency weighted



https://www.metaculus.com/questions/3479/date-weakly-general-ai-is-publicly-known/

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?

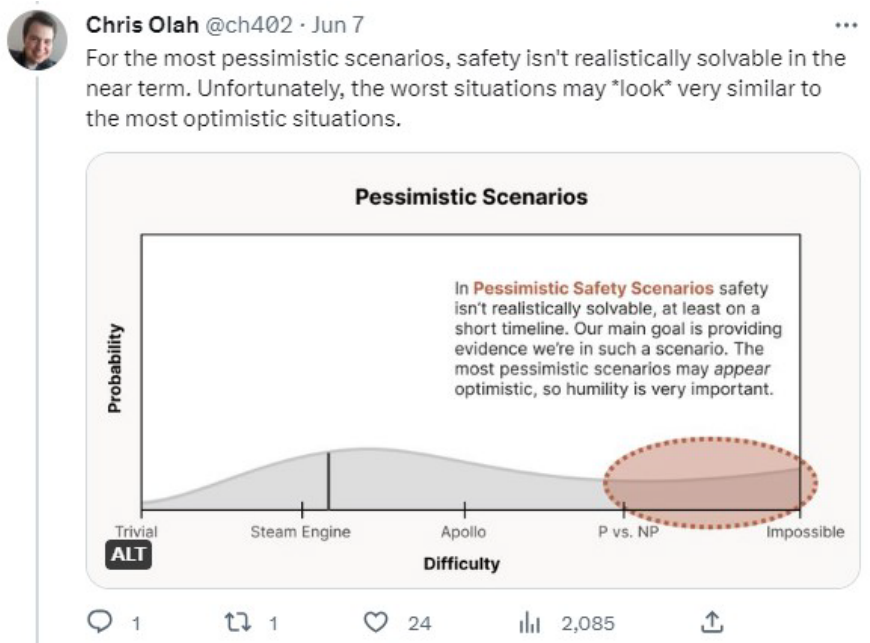
"Pessimistic Scenarios"

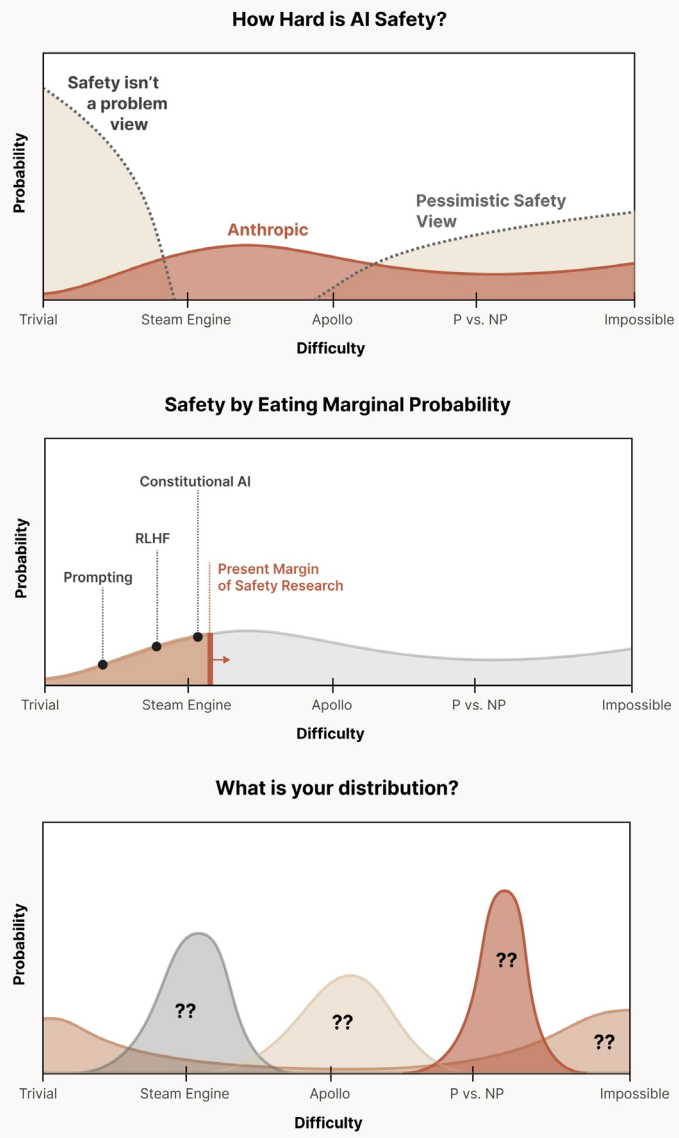
ANTHROP\C

Pessimistic scenarios: AI safety is an essentially unsolvable problem – it's simply an empirical fact that we cannot control or dictate values to a system that's broadly more intellectually capable than ourselves – and so we must not develop or deploy very advanced AI systems. It's worth noting that the most pessimistic scenarios might look like optimistic scenarios up until very powerful AI systems are created. Taking pessimistic scenarios seriously requires humility and caution in evaluating evidence that systems are safe.

If we're in a pessimistic scenario... Anthropic's role will be to provide as much evidence as possible that AI safety techniques cannot prevent serious or catastrophic safety risks from advanced AI, and to sound the alarm so that the world's institutions can channel collective effort towards preventing the development of dangerous AIs. If we're in a "near-pessimistic" scenario, this could instead involve channeling our collective efforts towards AI safety research and halting AI progress in the meantime. Indications that we are in a pessimistic or near-pessimistic scenario may be sudden and hard to spot. We should therefore always act under the assumption that we still may be in such a scenario unless we have sufficient evidence that we are not.

[Core Views on AI Safety: When, Why, What, and How. https://www.anthropic.com/index/core-views-on-ai-safety, March 23, 2023.]





[https://twitter.com/ch402/status/1666482929772666880]

Tools for Controllability

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

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

World Scientific
www.worldscientific.com

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; Unserveyability.

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

World Scientific
www.worldscientific.com

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.

IOP Publishing | Royal Swedish Academy of Sciences Physica Scripta

Phys. Scr. 92 (2017) 093001 (8pp) https://doi.org/10.1088/1402-4896/aa7ca8

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

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 of Unambiguous Communication as a Source of Failure in AI Systems

William J. Howe[1] , **Roman V. Yampolskiy**[2]

[1]Johns Hopkins University

[2]University of Louisville

whowe1@jhu.edu, roman.yampolskiy@louisville.edu

Abstract

Ambiguity is pervasive at multiple levels of linguistic analysis effectively making unambiguous communication impossible. As a consequence, natural language processing systems without true natural language understanding can be easily "fooled" by ambiguity, but crucially, AI also may use ambiguity to fool its users. Ambiguity impedes communication among humans, and thus also has the potential to be a source of failure in AI systems.

2 Phonology

Computational phonology is a core component of speech-based NLP systems. The ultimate goal of automatic speech recognition is to take an acoustic waveform as input and decode it into a string of words as text [Jurafsky, 2000]. The field which for several years was dominated by the Gaussian Mixture Model - Hidden Markov Model (GMM-HMM) framework has now made significant advancements using deep neural network (DNN) architectures to enable technologies like Siri, Alexa, and Google Assistant [Yu and Deng, 2016]. In particular recurrent neural networks which can-

[William J. Howe, Roman V. Yampolskiy. Impossibility of Unambiguous Communication as a Source of Failure in AI Systems. IJCAI-21 Workshop on Artificial Intelligence Safety (AISafety2021). Montreal, Canada. August 19-20, 2021.]

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

Journal of Cyber Security and Mobility, Vol. 11_3, 321–404.
doi: 10.13052/jcsm2245-1439.1132
© 2022 River Publishers

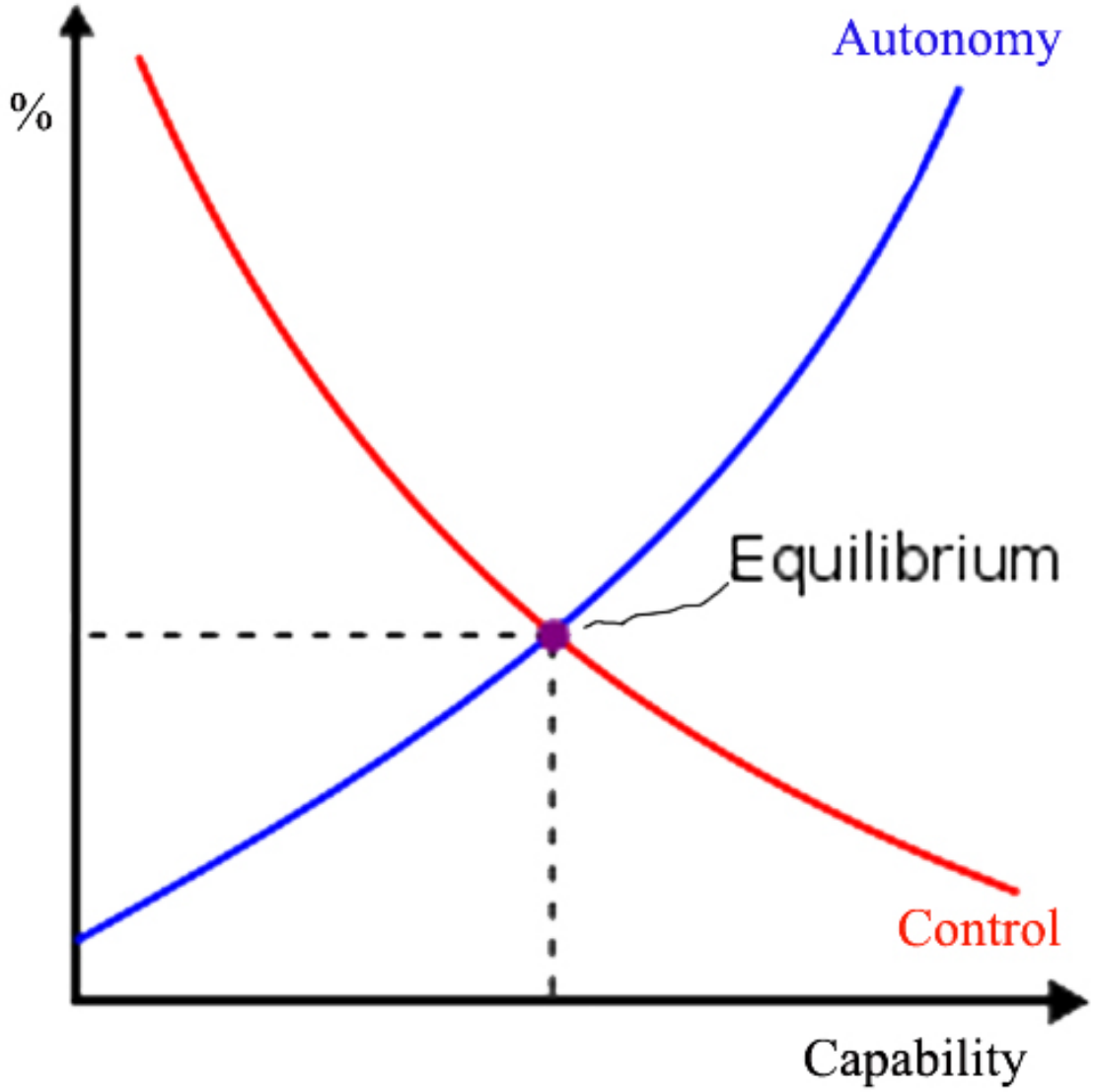


Figure 1: Control and Autonomy curves as Capabilities of the system increase.

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 paper, 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.

SURVEY FREE ACCESS

Impossibility Results in AI: A Survey

Just Accepted

Authors: Mario Brcic, Roman V. Yampolskiy Authors Info & Claims

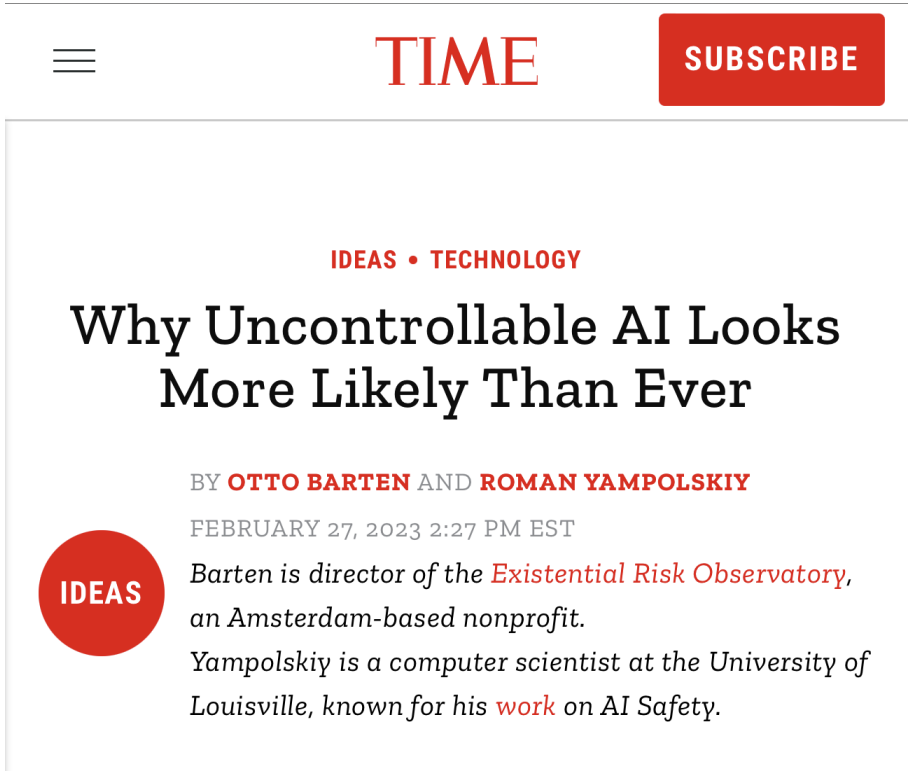
ACM Computing Surveys • https://doi.org/10.1145/3603371

Published: 02 June 2023 Publication History

Check for updates

Uncontrollable

- *"[I]t seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. ... At some stage therefore we should have to expect the machines to take control."* - Alan Turing
- *"[C]reation ... of entities with greater than human intelligence ... will be a throwing away of all the previous rules, ... an exponential runaway beyond any hope of control."* - Vernor Vinge
- *"Whereas the short-term impact of AI depends on who controls it, the long-term impact depends on whether it can be controlled at all."* - Stephen Hawking
- *"One thing is for sure: We will not control it."* - Elon Musk
- *"[T]here is no purely technical strategy that is workable in this area, because greater intelligence will always find a way to circumvent measures that are the product of a lesser intelligence."* - Ray Kurzweil.



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

IDEAS

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.

UNIVERSITY OF LOUISVILLE

The End!

Roman.Yampolskiy@louisville.edu

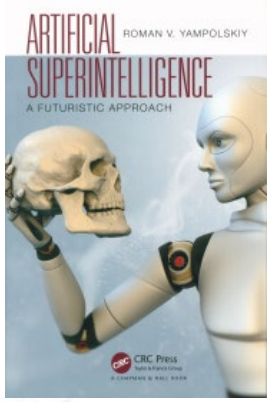
Director, CyberSecurity Lab

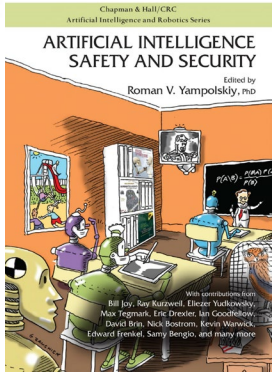
Computer Engineering and Computer Science

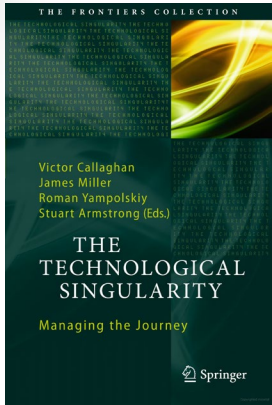
University of Louisville - cecs.louisville.edu/ry

twitter **@romanyam**

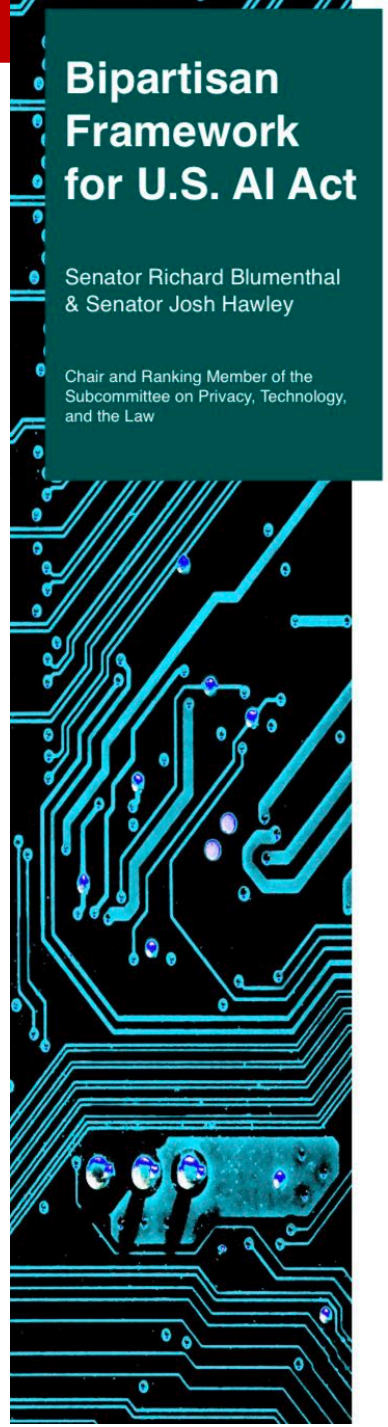
Follow me on Facebook **/Roman.Yampolskiy**







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



- **Establish a Licensing Regime Administered by an Independent Oversight Body:** Companies developing sophisticated general-purpose A.I. models (e.g., GPT-4) or models used in high-risk situations (e.g., facial recognition) should be required to register with an independent oversight body. Licensing requirements should include the registration of information about AI models and be conditioned on developers maintaining risk management, pre-deployment testing, data governance, and adverse incident reporting programs. The oversight body should have the authority to conduct audits of companies seeking licenses and cooperate with other enforcers, including considering vesting concurrent enforcement authority in state Attorneys General. The entity should also monitor and report on technological developments and economic impacts of A.I., such as effects on employment. Personnel must be subject to strong conflict of interest rules to mitigate capture and revolving door concerns.

- **Ensure Legal Accountability for Harms:** Congress should ensure that A.I. companies can be held liable through oversight body enforcement and private rights of action when their models and systems breach privacy, violate civil rights, or otherwise cause cognizable harms. Where existing laws are insufficient to address new harms created by A.I., Congress should ensure that enforcers and victims can take companies and perpetrators to court, including clarifying that Section 230 does not apply to A.I. In particular, Congress must take steps to directly prohibit harms that are already emerging from A.I., such as non-consensual explicit deepfake imagery of real people, production of child sexual abuse material from generative A.I., and election interference.

- **Defend National Security and International Competition:** Congress should utilize export controls, sanctions, and other legal restrictions to limit the transfer of advanced A.I. models, hardware and related equipment, and other technologies to China, Russia, and other adversary nations, as well as countries engaged in gross human rights violations.

- **Promote Transparency:** Congress should promote responsibility, due diligence, and consumer redress by requiring transparency from the companies developing and deploying A.I. systems.
  - Developers should be required to disclose essential information about the training data, limitations, accuracy, and safety of A.I. models to users and companies deploying systems, including through simple, comprehendible disclosures and to provide independent researchers access to data necessary to evaluate A.I. model performance.
  - Users should have a right to an affirmative notice that they are interacting with an A.I. model or system.
  - A.I. system providers should be required to watermark or otherwise provide technical disclosures of A.I.-generated deepfakes.
  - The new oversight body should establish a public database and reporting so that consumers and researchers have easy access to A.I. model and system information, including when significant adverse incidents occur or failures in A.I. cause harms.

- **Protect Consumers and Kids:** Companies deploying A.I. in high-risk or consequential situations should be required to implement safety brakes, including giving notice when A.I. is being used to make decisions, particularly adverse decisions, and have the right to a human review. Consumers should have control over how their personal data is used in A.I. systems and strict limits should be imposed on generative A.I. involving kids.