www.editada.org

International Journal of Combinatorial Optimization Problems and Informatics, 15(5), Dec 2024, 1-5. ISSN: 2007-1558. https://doi.org/10.61467/2007.1558.2024.v15i5.553

Editorial: A Brief Current Overview of Artificial Intelligence and Risk Factors

María del Carmen Santiago Díaz[1], Gustavo Trinidad Rubín Linares[1], and Juan Humberto Sossa Azuela[2]

[1] Benemérita Universidad Autónoma de Puebla, Facultad de Ciencias de la Computación.
[2] Instituto Politécnico Nacional (CIC-IPN).
marycarmen.santiago@correo.buap.mx, gustavo.rubin@ correo.buap.mx, humbertososa@gmail.com

Abstract. Artificial intelligence has developed at a very fast pace in various areas of knowledge, which is why we are already seeing worrying results regarding regulations governing the development of Artificial Intelligence applications. In this work we review some strategies that are being analyzed in various research groups in order to initiate a regulatory framework that gives us a panorama of greater security in future developments of Artificial Intelligence and its applications.

Keywords: Artificial Intelligence, High Stakes AI, Governance Measures, Safety Cases, Evaluation for Dangerous Capabilities

1 Introduction

Artificial intelligence has long since left the spectrum of science fiction to sneak into our lives and, although still in a very early phase, it is set to lead a revolution comparable to that generated by the Internet, with a sea of applications in multiple sectors such as: health, finance, transport and education, among others.

AI is present in facial detection on mobile phones, in virtual voice assistants such as Apple's Siri, Amazon's Alexa or Microsoft's Cortana and is integrated into our everyday devices through bots or applications for mobile devices, language learning and medical diagnoses, among others. The objective of all of them being in the simplest sense, to make people's lives easier, but it is not just that, they really seek to improve the quality of life.

Advances in AI are driving strategic areas such as the use of big data, improving its ability to process large volumes of data and providing communication, commercial and business advantages that have led it to position itself as the essential technology of the coming decades, benefiting such important sectors as transportation, education, health, culture, etc., to name a few.

This dizzying emergence of AI in our society has triggered an avalanche of applications in such critical fields as robotics, which has generated great concern among international organizations to consider the need to create regulations to regulate its use and employment in order to avoid problems that may arise in the future, which is why the European Union has proposed its own 6 Laws of Robotics.

These technologies are already changing the world, as it is estimated that the AI market could reach 127 billion dollars in 2025, which is much higher than the 2 billion in 2015. In addition, countries such as the United States and China will be at the forefront of investments, so it is estimated that AI will eliminate 85 million jobs in five years, but in return it will create 97 million.

And while there are opinions such as that of the Swedish philosopher from Oxford University, Nick Bostrom, who anticipates that "there is a 90% chance that between 2075 and 2090 there will be machines as intelligent as humans," or that of Stephen Hawking, who ventures that machines will completely surpass humans in less than 100 years, the truth is that far from making us obsolete, AI will make us more efficient and allow us to carry out actions that we would never have been able to carry out due to its complexity.

Artificial intelligence is transforming the way we design and build. By 2050, the effects of AI adoption will be widely felt in all aspects of our daily lives. As the world faces a number of urgent and complex challenges, from the climate crisis to housing, AI has the potential to make the difference between a dystopian future and a livable one. As we look to the future, we take stock of what is happening and, in turn, imagine how AI can improve our lives.

There are currently a number of expectations, including:

- It is estimated that by 2030 machine translators will do translations better than linguists themselves.
- All surgeries will be done by machines.
- It is estimated that by 2050 machines will do themselves, making the next model smarter than the last.

Although the extreme risks of AI have been warned about, there is no consensus on how to manage them [1]. Society's response is not yet up to the possibility of the rapid and transformative progress that experts expect. Of most concern is AI security research, which is lagging behind, and current government initiatives lack the mechanisms and institutions to prevent misuse and recklessness, which autonomous systems barely address.

Based on critical security lessons learned from other technologies, some aspects to consider in this AI growth scenario are proposed below.

2 Risks of accelerated AI development

Today’s deep learning systems still lack many capabilities, yet many companies are competing in various ways for this technology, tripling investment in training state-of-the-art models [2], with cash reserves required to scale the latest training runs in multiples of 100 to 1000 [3]. Furthermore, hardware chips become 1.4 times more cost-effective and AI training algorithms 2.5 times more efficient each year [4, 5]. But along with advanced AI capabilities come large-scale risks. AI systems threaten to amplify social injustice, erode societal stability, enable large-scale criminal activity, and facilitate automated warfare, mass personalized manipulation, and widespread surveillance.

Malicious actors might deliberately embed undesirable targets—even well-intentioned developers can inadvertently create AI systems that pursue undesirable targets. Once autonomous AI systems pursue undesirable goals, we may not be able to keep them under control. Software control is an old and unsolved problem: computer worms have long been able to proliferate and avoid detection.

Without sufficient caution, we may irreversibly lose control of autonomous AI systems, rendering human intervention ineffective. Large-scale cybercrime, social manipulation, and other harms could rapidly escalate. This uncontrolled advance of AI could culminate in large-scale loss of life and the biosphere, and the marginalization or extinction of humanity.

Humanity is investing enormous resources in making AI systems more powerful, but much less in making them safe and mitigating their harms. It is estimated that only 1–3% of AI publications address safety [6, 7].

There are many open technical challenges to ensure the safety and ethical use of generalized and autonomous AI systems. Unlike the advancement of AI capabilities, these challenges cannot be addressed by simply using more computing power to train larger models. They are unlikely to be solved automatically as AI systems become more capable and require dedicated research and engineering efforts.

Ensuring safety remains too difficult, extreme governance measures will be needed to prevent shortcuts driven by competition and overconfidence. Some of these challenges are:

- Oversight: More capable AI systems can better exploit weaknesses in technical oversight and testing [8, 9, 10].
- Robustness: AI systems behave unpredictably in new situations, some aspects improve with model scale and others even worsen [11].
- Interpretability and transparency: AI decision making is generally opaque, its inner workings need to be learned as larger, more capable models are very difficult to interpret, most large models are tested by trial and error [12].
- Developing inclusive AI: Methods will be needed to mitigate biases and integrate the values of the many populations that will be affected [13, 14].
- Addressing emerging challenges: Future AI systems may exhibit failure modes that we have so far only seen in theory or laboratory experiments [15, 16].

Ensuring safe development requires effective, risk-adjusted policies to mitigate damage when safety and security policies fail, such as:

- Assessing dangerous capabilities: As AI developers scale their systems, unforeseen capabilities emerge spontaneously without explicit programming and are only apparent after deployment, so rigorous methods need to be developed to elicit and assess these AI capabilities and predict them before training. Current assessments of frontier AI models for dangerous capabilities are limited to spot checks and demonstration attempts in specific environments but cannot reliably rule them out [4].
- Assessing AI alignment: Before training and deploying AI systems, we need methods to assess their propensity to use their capabilities. Purely behavioral assessments may fail for advanced AI systems as they might behave differently under assessment, simulating alignment [16].
- Risk assessment: We must learn to assess not only the dangerous capabilities of AI but also the risk from the complexity of a social context, with interactions and vulnerabilities [18].
- Resilience: Inevitably, some will misuse or act recklessly with AI. We need tools to detect and defend against AI-enabled threats such as large-scale influence operations, biological risks, and cyberattacks. However, as AI systems become more capable, they will eventually be able to bypass human-made defenses. To enable more powerful AI-based defenses, we must first learn how to make AI systems safe and aligned [19].

3 Government Policies

Governments around the world have taken positive steps on cutting-edge AI, with key players including China, the United States, the European Union, and the United Kingdom engaging in discussions and introducing initial guidelines or regulations [20]:

- Fast-acting, technology-savvy institutions to oversee AI: Key are policies that automatically kick in when AI reaches certain capability milestones, ensuring that risk mitigation efforts must be proactive, identifying risks to next-generation systems and requiring developers to address them before taking high-risk actions.
- Government perspective: Governments urgently need a comprehensive perspective on AI development. Regulators should require whistleblower protection, incident reporting, logging of key information about cutting-edge AI systems and their data sets throughout their lifecycle, and monitoring of model development and supercomputer use.
- Safety cases: Since there is no guarantee that governments can quickly develop the expertise needed to make reliable technical assessments of AI capabilities and risks at societal scale, the alternative is that frontier AI developers should bear the burden of proof to show that their plans keep risks within acceptable bounds.
- Mitigation: To keep AI risks within acceptable bounds, a governance methodology is needed that is tailored to the magnitude of the risks, clarifying the legal responsibilities arising from existing liability frameworks and holding frontier AI developers and owners legally accountable for any harm their models may cause.

5 Conclusions

Of concern is the scenario of serious and potentially catastrophic risks that could arise intentionally, if malicious actors use advanced AI systems to achieve harmful objectives or unintentionally in the event that an AI system develops strategies to achieve objectives that are not aligned with our values.

It is vitally important to focus on the factors that governments can focus their regulatory efforts on to mitigate the harms, especially the most significant ones, associated with AI.

It is urgent to immediately and massively invest in research efforts to design security systems and protocols that minimize the probability of producing unauthorized AI, as well as develop countermeasures against the possibility of undesirable scenarios.

There is a great need and opportunity for innovation in government policy research to design adaptable and agile regulations and treaties that protect citizens and society as technology evolves and new unexpected threats may emerge.

We have a moral responsibility to focus our research and greatest resources in a bold, coordinated effort to fully realize the economic and social benefits of AI, while protecting society, humanity, and our shared future from its potential dangers.

References

1. Statement on AI Risk. 2023 [cited 1 May 2024]. Available: https://www.safe.ai/work/statement-on-ai-risk.
2. Cottier B. Trends in the Dollar Training Cost of Machine Learning Systems. 2023. Available: https://epochai.org/blog/trends-in-the-dollar-training-cost-of-machine-learning-systems.
3. Alphabet. Alphabet annual report, page 33 (page 71 in the pdf): “As of December 31, 2022, we had USD113.8 billion in cash, cash equivalents, and short-term marketable securities”. [For comparison, the cost of training GPT-4 has been estimated as USD50 million (https://epochai.org/trends), and Sam Altman, the CEO of OpenAI, has stated that the cost for the whole process was more than USD100 million (https://www.wired.com/story/openai-ceo-sam-altman- the-age-of-giant-ai-models-is-already-over/).]. 2022. Available: https://abc.xyz/assets/d4/4f/a48b94d548d0b2fdc029a95e8c63/2022-alphabet-annual-report.pdf.
4. Hobbhahn M, Heim L, Aydos G. Trends in Machine Learning Hardware. 2023. Available: https://epochai.org/blog/trends-in-machine-learning- hardware.
5. Erdil E, Besiroglu T. Algorithmic progress in computer vision. arXiv [cs.CV]. 2022. Available: http://arxiv.org/abs/2212.05153.
6. Toner H, Acharya A. Exploring Clusters of Research in Three Areas of AI Safety. 2022. Available: https://cset.georgetown.edu/publication/exploring- clusters-of-research-in-three-areas-of-ai-safety/.
7. Emerging Technology Observatory. AI safety – ETO Research Almanac. [cited 12 Feb 2024]. Available: https://almanac.eto.tech/topics/ai-safety/.
8. Pan A, Bhatia K, Steinhardt J. The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models. International Conference on Learning Representations. 2022 [cited 15 Sep 2023]. Available: https://openreview.net/forum?id=JYtwGwIL7ye
9. Zhuang S, Hadfield-Menell D. Consequences of misaligned AI. Advances in Neural Information Processing Systems. 2020;33: 15763–15773.
10. Gao L, Schulman J, Hilton J. Scaling Laws for Reward Model Overoptimization. In: Krause A, Brunskill E, Cho K, Engelhardt B, Sabato S, Scarlett J, editors. Proceedings of the 40th International Conference on Machine Learning. PMLR; 23--29 Jul 2023. pp. 10835–10866.
11. Hendrycks D, Basart S, Mu N, Kadavath S, Wang F, Dorundo E, et al. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. arXiv [cs.CV]. 2020. Available: http://arxiv.org/abs/2006.16241.
12. Räuker T, Ho A, Casper S, Hadfield-Menell D. Toward Transparent AI: A Survey on Interpreting the Inner Structures of Deep Neural Networks. 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML). 2023. pp. 464–483.
13. Eubanks V. Automating Inequality: How High-Tech Tools Profile, Police and Punish the Poor. St Martin’s Press; 2018.

14. Sen A. Social Choice Theory. In: Arrow KJ, Intriligator M, editors. Handbook of Mathematical Economics, Vol III. Amsterdam: North Holland; 1986.
15. Hadfield-Menell D, Dragan A, Abbeel P, Russell S. The Off-Switch Game. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence. 2017; 220–227.
16. Ngo R, Chan L, Mindermann S. The alignment problem from a deep learning perspective. International Conference on Learning Representations 2024. 2024. Available: https://openreview.net/forum?id=fh8EYKFKns
17. Kinniment M, Sato LJK, Du H, Goodrich B, Hasin M, Chan L, et al. Evaluating Language-Model Agents on Realistic Autonomous Tasks. arXiv [cs.CL]. 2023. Available: http://arxiv.org/abs/2312.11671.
18. Koessler L, Schuett J. Risk assessment at AGI companies: A review of popular risk assessment techniques from other safety-critical industries. arXiv [cs.CY]. 2023. Available: http://arxiv.org/abs/2307.08823.
19. Hendrycks D, Carlini N, Schulman J, Steinhardt J. Unsolved Problems in ML Safety. arXiv [cs.LG]. 2021. Available: http://arxiv.org/abs/2109.13916.
20. The White House (US). Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence. 2023. Available: https://www.whitehouse.gov/briefing-room/presidential- actions/2023/10/30/executive-order-on-the-safe-secure- and-trustworthy-development-and-use-of-artificial- intelligence/.