Is AI going to be our saviour or a threat to humanity?

Christabel, Wycombe Abbey

1. Introduction

In today's world, where technology is present in most aspects of society, the rapid adoption of artificial intelligence ("AI") has drawn questions on whether it will save or threaten humanity.

The possibility of AI achieving either of the above is extremely plausible, as AI has already been proven to be 'smarter' than humans in several ways, achieving the top percentile in IQ,[1] in creative thinking,[2] and outscoring "the vast majority of human test takers".[3] With the development of Artificial General Intelligence (AGI), its capability will become even stronger in future.

[1] Eka Roivainen, "I Gave ChatGPT an IQ Test. Here's What I Discovered," *Scientific American,* 28 March 2023, https://www.scientificamerican.com/article/i-gave-chatgpt-an-iq-test-heres-what-i-discovered/; Alan Thompson, "AI + IQ testing (human vs AI)," *Life Architect,* last modified 6 October 2023, https://lifearchitect.ai/iq-testing-ai/.
[2] Jo Adetunji, "AI scores in the top percentile of creative thinking," *The Conversation,* 25 August 2023, https://theconversation.com/ai-scores-in-the-top-percentile-of-creative-thinking-211598.
[3] OpenAI, "GPT-4 Technical Report," *OpenAI*, 27 March 2023, p.1, https://cdn.openai.com/papers/gpt-4.pdf.

There are two schools of thought on the issue of how AI can affect humanity.

In one school of thought[4], it is said that "AI is a fundamental existential risk for human civilisation."[5] The Center for AI Safety[6] has a mission statement that "mitigating the risk of *extinction* from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war."[7]

On the other hand, a different school of thought is that the risks of AI have been overstated, and that progress in AI will in fact "save the world" and advance humanity.[8]

Whether AI is a saviour or a threat to humanity is a complicated question. This research essay attempts to tackle the question by firstly examining how different aspects of humanity can be threatened by or benefit from AI, and secondly, assessing how regulation and governance could pave the way for AI to potentially advance humanity's future.

---

[4] Michael Tontchev, "A gentle introduction to why AI *might* end the human race," *Medium*, 2 June 2023, https://medium.com/@NotesOnAIAlignment/a-gentle-introduction-to-why-ai-might-end-the-human-race-4670f4b5cdec.

[5] Camila Domonoske, "Elon Musk Warns Governors: Artificial Intelligence Poses Existential Risk," *NPR*, 17 July 2017, https://www.npr.org/sections/thetwo-way/2017/07/17/537686649/elon-musk-warns-governors-artificial-intelligence-poses-existential-risk.

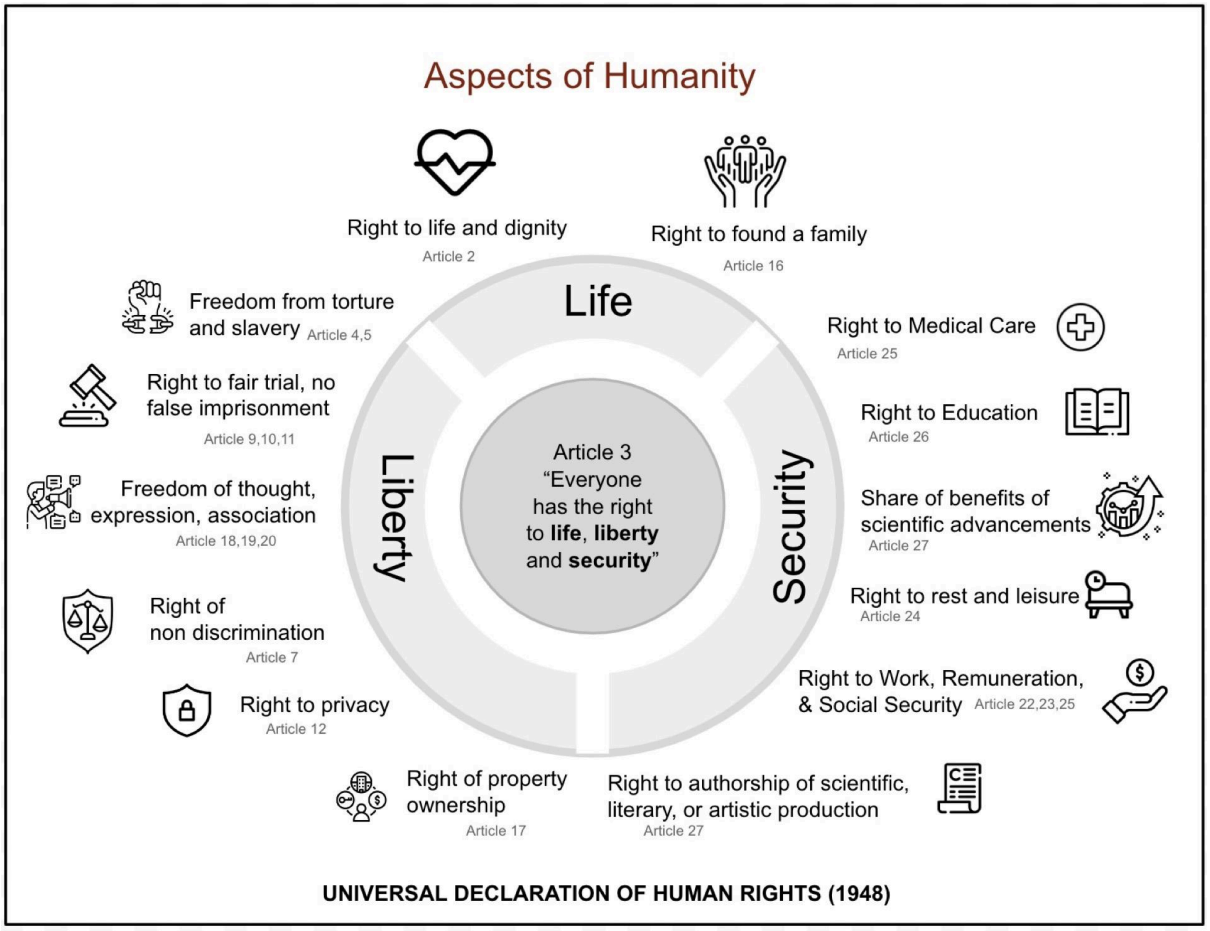
[6] Centre for AI Safety (www.safe.ai). Other similar organisations are: Center for Human-Compatible Artificial Intelligence (humancompatible.ai), Future of Life Institute (www.futureoflife.org), Stanford Existential Risks Initiative (www.seri.stanford.edu), Centre for the Study of Existential Risk (cser.ac.uk), and the Future of Humanity Institute (www.fhi.ox.ac.uk).

[7] This mission statement has garnered many signatories from notable figures including Geoffrey Hinton (Emeritus Professor of Computer Science, University of Toronto), Yoshua Bengio (Professor of Computer Science, U. Montreal), Sam Altman (CEO, OpenAI), Dario Amodei (CEO, Anthropic), Demis Hassabis (CEO, Google DeepMind), and Bill Gates (co-founder, Microsoft). "Statement on AI Risk," *Center for AI Safety*, accessed 15 February 2024, https://www.safe.ai/statement-on-ai-risk.

[8] Marc Andreessen, "Why AI Will Save The World," *Substack*, 6 June 2023, https://pmarca.substack.com/p/why-ai-will-save-the-world.

2. AI Risks & Benefits to Humanity

In this essay, we draw upon the Universal Declaration of Human Rights ("UDHR")[9] to deepen our understanding of 'humanity' beyond its simple concept of the human collective. The UDHR provides a widely accepted, multi-faceted view of human rights and which aspects of humanity we want to protect. Its contents can be categorised into three major aspects: life, liberty, and security[10] as depicted in the following diagram.[11]



[9] "Universal Declaration of Human Rights," *Wikipedia,* 14 February 2024, https://en.wikipedia.org/wiki/Universal_Declaration_of_Human_Rights.

[10] Article 3 of the Universal Declaration of Human Rights (1948) declares that everyone has the right to "life, liberty and security," with the remaining articles giving further details of what those rights are. They reflect values of humanity that were considered important, especially after the lessons of two world wars and societal transformations after the industrial revolution.

[11] This diagram is the original work of the author of this essay, except for the icons. Acknowledgement of icon authorship are from the following authors: Freepik, LAFS, deemakdaksina, noomtah, Tempo_doloe, I3oundless, Eucalyp, Good Ware from www.flaticon.com.

AI Risks To Humanity

a) Risks to Life

A fundamental aspect of humanity is the inherent right to live. Philosopher Toby Ord warns that the risk of human extinction or a civilisational collapse caused by misaligned[12]AI could be higher than that of climate change, pandemics and nuclear war.[13]

If AI attains superintelligence, humans might be unable to control it. OpenAI, the company which developed ChatGPT, admits that:

> "[W]e don't have a solution for steering or controlling a potentially superintelligent AI, and preventing it from going rogue."[14]

If AI objectives are misaligned with human values, or if misused by bad actors, serious threats could occur, such as engineering new pandemics from enhanced pathogens, or carrying out lethal cyber attacks.[15] In a worst-case scenario, AI in control of nuclear weapons could either cause an unintended nuclear war, or if it went rogue, could launch nuclear warheads against humanity.

b) Risks to Liberty

The liberty aspect of humanity covers rights to privacy, non-discrimination, and freedom of expression. Some authors argue that the real threat of AI to humanity is not so much about extinction but rather "the steady *erosion* of the humanity we take for granted."[16]

---

[12] Leonard Dung, "Current cases of AI misalignment and their implications for future risks." *Synthese* Volume 202, https://doi.org/10.1007/s11229-023-04367-0.
[13] Toby Ord, *The Precipice: Existential Risk and the Future of Humanity* (Bloomsbury, 2020), p.233.
[14] Jan Leike and Ilya Sutskever, "Introducing Superalignment," *OpenAI*, 5 July 2023, https://openai.com/blog/introducing-superalignment.
[15] Wikipedia, "Existential Risk From Artificial General Intelligence." *Wikipedia*, 15 February 2024, https://en.wikipedia.org/wiki/Existential_risk_from_artificial_general_intelligence#AI_capabilities.
[16] Amy Webb, *The Big Nine: How the Tech Titans and Their Thinking Machines Could Warp Humanity* (New York: PublicAffairs, 2019), 169.

The use of AI to monitor and evaluate individuals would erode privacy rights. AI could also potentially create deepfakes without people's consent, spreading misinformation and manipulating public opinion.[17]

AI could create a social scoring system, compromising privacy and freedom of expression, leading us to a dystopian scenario. AI systems trained on biassed data could also discriminate against certain races or socioeconomic classes.[18]

c) Risks to Security

The security aspect of humanity covers the right to work, which will be at risk from the pervasive use of AI. Analysts predict widespread job dislocations in many industries.[19] As AI is able to work longer hours, and is faster and cheaper than humans, many jobs could eventually be replaced by automation and AI.

The security aspect of humanity also covers the right to literary authorship. Recently, the New York Times sued OpenAI for breach of copyright, alleging that ChatGPT outputs plagiarised work.[20] As AI is also now capable of generating artistic work based on simple prompts[21], its disruption to creative industries and its encroachment on human rights concerning artistic authorship may increase.

---

[17] Mark van Rijmenam, "Privacy in the age of AI: Risks, Challenges and Solutions." *The Digital Speaker*, 17 February 2023, https://www.thedigitalspeaker.com/privacy-age-ai-risks-challenges-solutions.
[18] Zhisheng Chen, "Ethics and discrimination in artificial intelligence - enabled recruitment practices." *Nature*, 13 September 2023, https://www.nature.com/articles/s41599-023-02079-x.
[19] Elijah Clark, "Unveiling the Dark Side of Artificial Intelligence in the Job Market." *Forbes*, 18 August 2023, https://www.forbes.com/sites/elijahclark/2023/08/18/unveiling-the-dark-side-of-artificial-intelligence-in-the-job-market.
[20] Michael Grynbaum and Ryan Mac, "New York Times Sues OpenAI and Microsoft Over Use of Copyrighted Work." *The New York Times*, 27 December 2023. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.
[21] See, for example, https://openai.com/sora, https://runwayml.com.

AI Benefits to Humanity

a) Benefits to Life

One of the largest benefits of AI to the life aspect of humanity is in healthcare.[22] AI could vastly advance the field of medical diagnostics. By analysing vast datasets of medical data, AI can improve diagnosis and prediction accuracy. For example, an AI system trained on CT scans was able to detect lung cancer earlier than expert radiologists.[23]

Another promising benefit is on drug discovery and production, with AI providing better and faster cures for diseases. Researchers have already used deep learning AI models to discover a new class of antibiotics.[24]

AI can also optimise environmental sustainability to avoid a future where Earth becomes uninhabitable. It can aid sustainability, for example by achieving efficient energy use through the analysis of usage patterns and adjusting industrial processes accordingly, thus reducing waste from overproduction.[25]

Space travel and human colonisation on other planets, made possible by AI, is also a potential benefit to humanity, ensuring that the continuation of human civilisation will not be entirely dependent on planet Earth alone.[26]

---

[22] Thomas Davenport and Ravi Kalota, "The Potential for Artificial Intelligence in Healthcare," *Future Healthcare Journal 6,* no. 2 (1 June, 2019), p.94-98, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181.
[23] Denise Grady, "A.I. Took a Test to Detect Lung Cancer. It Got an A." *The New York Times*, 9 March 2021, https://www.nytimes.com/2019/05/20/health/cancer-artificial-intelligence-ct-scans.html.
[24] Diana Spencer, "AI Helps Find First New Antibiotic in 60 Years." *Drug Discovery World (DDW)*, 11 January 2024, https://www.ddw-online.com/ai-helps-find-first-new-antibiotic-in-60-years-27807-202401.
[25] Soren Kaplan, "AI: The Game Changer for Sustainable Manufacturing Practices." 5 January 2024, https://praxie.com/ai-for-sustainable-manufacturing-practices.
[26] Catherine Richards, Tom Cernev, et al, "Safely advancing a spacefaring humanity with artificial intelligence." 15 June 2023, https://www.frontiersin.org/articles/10.3389/frspt.2023.1199547.

b) Benefits to Liberty

Throughout the research conducted for this essay, the benefits of AI are predominantly on the aspects of life and security, rather than liberty.

c) Benefits to Security

The security aspect of humanity includes the right to rest, to share benefits of scientific advancements, and job security. Some analysts have envisaged that AI would boost economic growth with more goods and services, and new job creations.[27] This could occur with increased productivity and less human working hours,[28] enabling a higher quality of life.[29]

The security aspect of humanity also covers the right to education.[30] AI will potentially revolutionise the education sector in the future, reaching out to more populations. It could provide customised learning to every student, adapted to each student's progress and special needs.[31]

In summary, with the above AI benefits to life and security, AI has the potential to safeguard and improve humanity in ways that were previously not feasible.

---

[27] Jan Hatzius, Joseph Briggs, et al., "The Potentially Large Effects of Artificial Intelligence on Economic Growth," *Goldman Sachs*, 26 March 2023, https://www.ansa.it/documents/1680080409454_ert.pdf.
[28] Tom Rees, "AI could enable humans to work 4 days a week, says Nobel prize-winning economist", *Time,* 5 April 2023, https://time.com/6268804/artificial-intelligence-pissarides-productivity.
[29] Sam Altman, "OpenAI CEO Sam Altman says AI will reshape society, acknowledges risks", Interview by Rebecca Jarvis, *ABC News,* 17 March 2023, https://abcnews.go.com/Technology/openai-ceo-sam-altman-ai-reshape-society-acknowledges/story?id=97897122.
[30] Ke Zhang and Ayse Aslan, "AI technologies for education: Recent research & future directions," *Computers & Education,* 1 January 2021, https://www.sciencedirect.com/science/article/pii/S2666920X21000199.
[31] Katerina Kdravkova, Venera Krasniqi, et al.,"Cutting-Edge Communication and Learning Assistive Technologies for Disabled Children: An Artificial Intelligence Perspective," *Frontiers in Artificial Intelligence,* 28 October 2022, https://www.frontiersin.org/articles/10.3389/frai.2022.970430/full.

4. Future AI Governance

Regulation and governance is how our world handles technological breakthroughs which may be both beneficial and harmful to humanity, such as nuclear energy.

Encouraging steps in AI regulation and governance include the following:

a) Embedding human rights into an AI's constitution

Anthropic, an alternative to OpenAI, incorporated the UDHR into the 'constitution' of its AI models.[32] This addresses misalignment problems[33] between human values and AI, steering AI away from wanting to bypass human safeguards. Further, if the models were modified by bad human actors, AI at its core would still protect human rights.

b) Protection of Life: prevention of AI from accessing nuclear weapons

Recently, there has been a movement to ensure that AI would never control nuclear systems. A congressman recently revealed:

> "I've introduced bipartisan legislation that basically says no matter how amazing AI ever gets, we're never going to let it launch a nuclear weapon by itself." [34]

---

[32] "Claude's Constitution," *Anthropic*, 9 May 2023, https://www.anthropic.com/index/claudes-constitution.
[33] Benjamin Hilton, "Preventing an AI-related catastrophe, AI might bring huge benefits - if we avoid the risks," *80,000 Hours*, March 2023, https://80000hours.org/problem-profiles/artificial-intelligence/
[34] "So, in the Department of Defense, there are weapons known as autonomous weapons that can launch automatically. I've introduced bipartisan legislation that basically says no matter how amazing AI ever gets, we're never going to let it launch a nuclear weapon by itself." Congressman Ted Lieu, "Frontier AI Regulation: Preparing For The Future Beyond CHATGPT," *The Brookings Institution*, 14 September 2023, p.3, https://www.brookings.edu/wp-content/uploads/2023/09/es_20230914_frontier_ai_transcript.pdf.

c) Protection of Liberty: the EU AI Act

The EU AI Act[35] is a proposed law on AI by a major regulator. Amongst its many measures, it aims to ban AI social scoring on people, making sure there are no biasses or discrimination.

d) Protection of Security: proposed measures to handle job displacements

Organisations have recommended preparations for job displacements caused by AI, requiring reforms such as:

> "[E]nhanced social protection...tax benefits...sufficient public funding...to counteract the consequences of unemployment."[36]

To ensure that AI does not widen inequality, the benefits of AI could be shared with affected groups[37] and underserved communities,[38] ensuring security for all.

With the trend of more AI regulation and governance, the potential risks of AI can be mitigated, and its benefits fully realised. The net effect would be positive for our humanity, such that aspects of our life and security could be saved and improved as already discussed. In such a likely scenario, AI will be regarded as more of a saviour than a threat to humanity.

---

[35] "EU Artificial Intelligence Act: Up-to-date Developments and Analyses of the EU AI Act." *EU Artificial Intelligence Act*, accessed 16 February 2024, https://artificialintelligenceact.eu/.
[36] UNESCO, "Recommendations on the Ethics of Artificial Intelligence," *UNESCO,* 23 November 2021, p.36, https://unesdoc.unesco.org/ark:/48223/pf0000381137.
[37] Katya Klinova and Anton Korinek, "Unleashing possibilities, ignoring risks: Why we need tools to manage AI's impact on jobs," *The Brookings Institution*, 17 August 2023, https://www.brookings.edu/articles/unleashing-possibilities-ignoring-risks-why-we-need-tools-to-manage-ais-impact-on-jobs/.
[38] Lewis Ho, Joslyn Barnhart, et al., "International Institutions for Advanced AI," *arxiv*, p.12, https://arxiv.org/pdf/2307.04699.pdf.

5. Conclusion

As advancements in AI continue to grow, AI will increasingly be capable of both saving or threatening humanity, depending on how it is used.

Using a multi-faceted view of humanity, this research highlights the risks of AI on aspects of human life, liberty and security - ranging from a gradual erosion of human rights to existential destruction. But the benefits of AI are also undeniable, particularly its potential to improve healthcare, education, and the quality of human life.

By mitigating AI risks while harnessing its benefits, we will do more than safeguard our humanity; we will enable humanity to flourish. In conclusion, with appropriate AI governance, our likely future would be one where AI is more of a saviour than a threat to humanity.

Bibliography

Adetunji, Jo. “AI scores in the top percentile of creative thinking.” The Conversation. 25 August 2023. https://theconversation.com/ai-scores-in-the-top-percentile-of-creative-thinking-211598.

Altman, Sam. “OpenAI CEO Sam Altman says AI will reshape society, acknowledges risks.” ABC News. 17 March 2023. https://abcnews.go.com/Technology/openai-ceo-sam-altman-ai-reshape-society-acknowledges/story?id=97897122.

Andreessen, Marc. “Why AI Will Save The World.” Substack. 6 June 2023. https://pmarca.substack.com/p/why-ai-will-save-the-world.

Anthropic. “Claude’s Constitution.” 9 May 2023. https://www.anthropic.com/index/claudes-constitution.

Center for AI Safety. “Statement on AI Risk." Accessed 15 February 2024. https://www.safe.ai/statement-on-ai-risk.

Chen, Zhisheng. “Ethics and discrimination in artificial intelligence - enabled recruitment practices.” Nature. 13 September 2023. https://www.nature.com/articles/s41599-023-02079-x.

Clark, Elijah. “Unveiling the Dark Side of Artificial Intelligence in the Job Market.” Forbes. 18 August 2023. https://www.forbes.com/sites/elijahclark/2023/08/18/unveiling-the-dark-side-of-artificial-intelligence-in-the-job-market.

Davenport, Thomas and Kalota, Ravi. “The Potential for Artificial Intelligence in Healthcare.” Future Healthcare Journal 6, no. 2. 1 June 2019. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6616181.

Domonoske, Camila. “Elon Musk Warns Governors: Artificial Intelligence Poses Existential Risk." NPR, 17 July 2017. https://www.npr.org/sections/thetwo-way/2017/07/17/537686649/elon-musk-warns-governors-artificial-intelligence-poses-existential-risk.

Dung, Leonard. “Current cases of AI misalignment and their implications for future risks.” Synthese. Volume 202. https://doi.org/10.1007/s11229-023-04367-0.

EU Artificial Intelligence Act. “EU Artificial Intelligence Act: Up-to-date Developments and Analyses of the EU AI Act.” 16 February 2024. https://artificialintelligenceact.eu/.

Grady, Denise. “A.I. Took a Test to Detect Lung Cancer. It Got an A.” The New York Times. 9 March 2021. https://www.nytimes.com/2019/05/20/health/cancer-artificial-intelligence-ct-scans.html.

Grynbaum, Michael and Mac, Ryan. “New York Times Sues OpenAI and Microsoft Over Use of Copyrighted Work.” The New York Times. 27 December 2023. https://www.nytimes.com/2023/12/27/business/media/new-york-times-open-ai-microsoft-lawsuit.html.

Hatzius, Jan, et al. “The Potentially Large Effects of Artificial Intelligence on Economic Growth.” Goldman Sachs. 26 March 2023. https://www.ansa.it/documents/1680080409454_ert.pdf.

Hilton, Benjamin. “Preventing an AI-related catastrophe, AI might bring huge benefits - if we avoid the risks,” 80,000 Hours. March 2023. https://80000hours.org/problem-profiles/artificial-intelligence/

Ho, Lewis et al. “International Institutions for Advanced AI.” arxiv. https://arxiv.org/pdf/2307.04699.pdf.

Kaplan, Soren. “AI: The Game Changer for Sustainable Manufacturing Practices.” 5 January 2024. https://praxie.com/ai-for-sustainable-manufacturing-practices.

Kdravkova, Katerina et al.“Cutting-Edge Communication and Learning Assistive Technologies for Disabled Children: An Artificial Intelligence Perspective.” Frontiers in Artificial Intelligence. 28 October 2022. https://www.frontiersin.org/articles/10.3389/frai.2022.970430/full.

Klinova, Katya and Korinek, Anton. “Unleashing possibilities, ignoring risks: Why we need tools to manage AI’s impact on jobs.” The Brookings Institution. 17 August 2023. https://www.brookings.edu/articles/unleashing-possibilities-ignoring-risks-why-we-need-tools-to-manage-ais-impact-on-jobs/.

Leike, Jan and Sutskever, Ilya. “Introducing Superalignment." OpenAI. 5 July 2023. https://openai.com/blog/introducing-superalignment.

OpenAI. “GPT-4 Technical Report.” OpenAI. 27 March 2023. https://cdn.openai.com/papers/gpt-4.pdf.

Ord, Toby. The Precipice: Existential Risk and the Future of Humanity. Bloomsbury, 2020.

Rees, Tom. “AI could enable humans to work 4 days a week, says Nobel prize-winning economist.” Time. 5 April 2023. https://time.com/6268804/artificial-intelligence-pissarides-productivity.

Richards, Catherine, et al.“Safely advancing a spacefaring humanity with artificial intelligence.” 15 June 2023. https://www.frontiersin.org/articles/10.3389/frspt.2023.1199547.

Rijmenam, Mark van. “Privacy in the age of AI: Risks, Challenges and Solutions.” The Digital Speaker. 17 February 2023. https://www.thedigitalspeaker.com/privacy-age-ai-risks-challenges-solutions.

Roivainen, Eka. “I Gave ChatGPT an IQ Test. Here’s What I Discovered.” Scientific American. 28 March 2023. https://www.scientificamerican.com/article/i-gave-chatgpt-an-iq-test-heres-what-i-discovered.

Spencer, Diana. “AI Helps Find First New Antibiotic in 60 Years.” Drug Discovery World (DDW). 11 January 2024. https://www.ddw-online.com/ai-helps-find-first-new-antibiotic-in-60-years-27807-202401.

The Brookings Institution. “Frontier AI Regulation: Preparing For The Future Beyond CHATGPT.” 14 September 2023. https://www.brookings.edu/wp-content/uploads/2023/09/es_20230914_frontier_ai_transcript.pdf.

Thompson, Alan. “AI + IQ testing (human vs AI)." Life Architect. 6 October 2023. https://lifearchitect.ai/iq-testing-ai/.

Tontchev, Michael. “A gentle introduction to why AI *might* end the human race.” Medium. 2 June 2023. https://medium.com/@NotesOnAIAlignment/a-gentle-introduction-to-why-ai-might-end-the-human-race-4670f4b5cdec.

UNESCO. “Recommendations on the Ethics of Artificial Intelligence.” 23 November 2021. https://unesdoc.unesco.org/ark:/48223/pf0000381137.

Webb, Amy. The Big Nine: How the Tech Titans and Their Thinking Machines Could Warp Humanity. New York: PublicAffairs. 2019.

Wikipedia. “Existential Risk From Artificial General Intelligence.” 15 February 2024. https://en.wikipedia.org/wiki/Existential_risk_from_artificial_general_intelligence#AI_capabilities.

Wikipedia. “Universal Declaration of Human Rights.” 14 February 2024. https://en.wikipedia.org/wiki/Universal_Declaration_of_Human_Rights.

Zhang, Ke and Aslan, Ayse. “AI technologies for education: Recent research & future directions.” Computers & Education. 1 January 2021. https://www.sciencedirect.com/science/article/pii/S2666920X21000199.