Question 2

2. Racial Bias in Medical Algorithms

The use of AI in healthcare has faced challenges with racial bias, notably in a 2019 case where a U.S. healthcare algorithm discriminated against Black patients by using biased spending data, leading to reduced care compared to white patients [1, 2].

Addressing ethical challenges requires adherence to fairness principles like transparency and accountability [3]. Transparency enables audits to detect biases [4, 5], although potentially reducing model accuracy [6], whilst accountability allows for the ongoing improvement of algorithms following bias identification [7]. A method for enhancing fairness is through legislation, such as the EU AI Act [8], which mandates developers to reveal their algorithms' variables, data sources, and inclusion rationale [9].

Obermeyer et al. [10], who identified the biased algorithm in US healthcare, present a key case study on applying fairness principles. By adjusting the training labels of the algorithm, they significantly reduced its bias. This highlights the potential mitigation of ethical issues in healthcare algorithms through transparency and accountability.

3. AI system safety and existential risks in warfare

The risk of super-intelligent AI diverging from human welfare and making catastrophic decisions poses a significant challenge [11, 12]. Recent developments in Large Language Models (LLMs) have led many researchers to believe 'High-level machine intelligence' will be achieved in the next century [13, 14].

A specific example of an existential risk is AI's application in military contexts [15], where AI could independently decide to maximise human casualties [16]. The current use of Loitering Attack Munitions (LAMs), automated missiles that activate upon target acquisition [17, 18], underscores ethical concerns regarding AI's role in lethal decisions [19, 20]. If an advanced AI was used in more destructive military applications, the potential consequences could be catastrophic [21].

Addressing these existential threats requires promoting transparency, fostering international cooperation, and enacting global legislation to align AI development with human safety and interests [22, 23]. Additionally, creating controlled AI shutdown mechanisms is vital [24], though Russell [25] warns that a super-intelligent AI may be capable of overriding these safety measures. Ultimately, a unified global strategy is required to prioritise human safety in the advancement of AI technologies.

References

[1] S. Jemielity. (2019) Health care prediction algorithm biased against black patients, study finds. [Online]. Available: https://news.uchicago.edu/story/health-care-prediction-algorithm-biased-against-black-patients-study-finds. [Accessed 27 03 2024].

[2] H. Ledford, "Millions affected by racial bias in health-care algorithm," *Nature*, vol. 574, pp. 608–609, Oct 2019.

[3] T. Lawry, *AI in Health*, 1st ed. Boca Raton: CRC Press, Taylor & Francis Group, 2020.

[4] B. Séroussi, K. F. Hollis, and L. F. Soualmia, "Transparency of health informatics processes as the condition of healthcare professionals' and patients' trust and adoption: the rise of ethical requirements," *Yearbook of Medical Informatics*, vol. 29, no. 01, pp. 007–010, 2020.

[5] P. D. Winter and A. Carusi, "(de)troubling transparency: artificial intelligence (ai) for clinical applications," *Medical Humanities*, vol. 49, no. 1, pp. 17–26, 2023.

[6] M. Kearns and A. Roth, *The ethical algorithm*, 1st ed. New York: Oxford University Press, 2020.

[7] J. Donovan, J. Matthews, R. Caplan, and L. Hanson, "Algorithmic accountability: A primer," 2018.

[8] E. Parliament. (2023) Eu ai act: first regulation on artificial intelligence. [Online]. Available: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence. [Accessed 27 03 2024].

[9] L. Edwards, "The eu ai act: a summary of its significance and scope," *Artificial Intelligence (the EU AI Act)*, vol. 1, 2021.

[10] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, pp. 447–453, Oct 2019.

[11] J. Barrat, *Our Final Invention*, 1st ed. New York: Thomas Dunne Books, 2013.

[12] T. Ord, *The Precipice: Existential Risk and the Future of Humanity*, 1st ed. New York: Hachette Books, 2020.

[13] K. Grace, J. Salvatier, A. Dafoe, B. Zhang, and O. Evans, "When will ai exceed human performance? evidence from ai experts," *Journal of Artificial Intelligence Research*, vol. 62, pp. 729–754, Jul. 2018.

[14] P. Welsby and B. M. Y. Cheung, "Chatgpt," *Postgraduate Medical Journal*, vol. 99, no. 1176, pp. 1047–1048, 2023.

[15] R. V. Yampolskiy, "Taxonomy of pathways to dangerous artificial intelligence," 2016.

[16] M. L. Cummings, "Artificial intelligence and the future of warfare," 2017.

[17] K. Atherton. (2021) Brookings: Loitering munitions preview the autonomous future of warfare. [Online]. Available: https://www.brookings.edu/articles/loitering-munitions-preview-the-autonomous-future-of-warfare. [Accessed 30 03 2024].

[18] I. Bode and T. Watts, "Loitering munitions and unpredictability: Autonomy in weapon systems and challenges to human control," 2023.

[19] R. J. Emery, "Algorithms, ai, and ethics of war," *Peace Review: A Journal of Social Justice*, vol. 33, pp. 205–212, 2021.

[20] P. Scharre, *Army of None: Autonomous Weapons and the Future of War*. New York: W. W. Norton & Company, 2018.

[21] M. Tegmark, *Life 3.0: Being Human in the Age of Artificial Intelligence*, 1st ed. Toronto: Alfred A. Knopf, 2017.

[22] P. Cihon, "Standards for ai governance: International standards to enable global coordination in ai research & development," 2019.

[23] D. Leslie, "Understanding artificial intelligence ethics and safety," *CoRR*, vol. abs/1906.05684, 2019.

[24] A. Critch and D. Krueger, "Ai research considerations for human existential safety," 2020.

[25] S. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*, 1st ed. Viking.