**Racial Bias in Medical Algorithms**

In 2019, a widely used US healthcare algorithm was found to discriminate by prioritising hospital services based on historical spending records, resulting in the allocation of relatively less future funding and fewer referrals for black patients [1, 2]. Through the application of a series of test data sets, Obermeyer *et al.* [3] identified this inadvertent bias and the team was able to mitigate against it by adjusting the model’s training labels.

The fact that this third-party assessment and adjustment were possible, demonstrates how exposing a model’s internal operation can aid bias identification and removal [4, 5]. Ensuring greater transparency of AI models is becoming the subject of legislation, for example the 2023 EU AI Act seeks to enforce transparency principles by requiring developers to disclose an algorithm’s variables, data sources, and selection logic [9, 10]. While ensuring that organizations building AI systems are held accountable for the processes used in their development may lead to algorithmic changes that reduce bias [7, 8], care needs to be taken that the removal of bias does not significantly affect the performance of the model in its application domain [6].

**AI system safety and existential risks in warfare**

Recent developments in Large Language Models have led many researchers to believe that AI systems capable of directly acting in the real world based on decisions they have taken autonomously will become available later this century [12, 13]. With such advancements comes the risk that AI systems whose decision-making does not prioritise human welfare may pose a threat to life [11].

A specific example of a military AI system posing an existential risk [14, 15], is one that decides maximising human casualties would be the best strategy to achieve a high-level battlefield objective [16]. Recent deployments of Loitering Attack Munitions, automated missiles that activate on target acquisition [17, 18], have raised ethical concerns over the use of AI in situations where human beings are potential targets [19]. If such advanced AI was given control of powerful military weapons and applied more widely, the ramifications for the human race's survival could be profound [20].

Addressing these existential threats requires international cooperation to guarantee the transparency of AI algorithms [21, 22]. The Strategic Arms Reduction Treaty serves as a historical example; during the Cold War, transparency provided a deterrent through the verification of adversary’s weapons systems [23]. An additional future safeguard could be to mandate the inclusion of a human-controlled override in all military AI systems [24], although Russell [25] warns that a super-intelligent AI may become capable of removing such safety measures. Ultimately, a global strategy that prioritises human wellbeing in all areas of AI usage will be essential.