Abstract
As artificial intelligence progresses rapidly towards human-level machine intelligence, policymakers urgently need to prioritise aligning advanced AI systems with human values. Major investments are driving progress in deep learning and reinforcement learning, techniques that can achieve human-level intelligence within years according to some estimates. However, techniques for value alignment and ensuring the safe development of AI have received far less attention and resources. This policy brief argues that policymakers, especially in Africa and the Global South, must focus on issues like value alignment and ethics to ensure the equitable development of increasingly autonomous AI.

Close collaboration between policymakers, researchers and communities will be required to navigate the challenges posed by advanced AI, mitigate risks from its development and use, and optimise the societal benefits of progress in this pivotal technology. Governments need to fund research on AI risk and ethics, and work with experts to establish laws and policies that encourage promising work on value alignment and safety. Regulation should steer companies towards managing risks and ensuring accountability for their AI systems. Overall, AI safety and ethics should be a top priority today so societies can have optimistic visions of a future with advanced AI.

As AI continues its rapid progress, policymakers have an obligation to understand and prepare for the challenges posed by human and superhuman AI before these systems become pervasive or difficult to curb. We can never fully predict the future, but we know enough about the issues at hand to take them seriously and act now. By addressing this challenge, policymakers can help ensure that human judgement, ethics and compassion remain central to how technology shapes the future. The time to act is now. Overall, the equitable and beneficial development of advanced AI will require recognising it as one of the most pressing policy issues of our time. With urgent action and close collaboration, policymakers can help build a better future with AI.

Core Thesis
Policymakers in Africa and the Global South must urgently prioritise research and policies for aligning advanced AI with human values, or risk exacerbating existing inequalities as AI systems become more autonomous and ubiquitous.

To realise the rapid advances in artificial intelligence and its potential to match and exceed human intelligence in the coming decades, advanced AI systems, in this context artificial general intelligence (AGI), must be tuned. It should be one of our top priorities. We cannot wait and ponder how to ensure that such powerful technology is grounded in and guided by human values. For context, currently we rely on Reinforcement Learning from Human Feedback (RLHF). Alignment of these systems is a very complex problem that will require decades of interdisciplinary research to resolve, let alone fully resolve. At the same time, the competitive interests of military and technology companies are driving rapid advances in AI. Researchers focused on managing the risks of advanced AI need far more support to steer

progress in a responsible and ethical direction. Management and monitoring alone are probably not enough. We need to understand how to balance the motivation and value of the AI system itself. The challenges posed by superhuman AI may seem far away, but if not taken seriously, the potential consequences can be devastating. Now is the time to act, as we can lay the groundwork for artificial general intelligence that will benefit all of humanity.

Recently, we have seen OpenAI, a leading AI company publish a blog on *Superalignment* with the aim of getting ahead of the curve by setting up a team of research scientists and engineering and even setting aside 20% of the company's compute power in the establishment of the project. Ilya Sutskever, chief scientist of OpenAI and earliest pioneers in deep learning, who is also going to lead the team, estimates that we will have these advanced AI systems in about four years. By that timing, it means the implementation around the whole topic has been squeezed and requires haste. To quote them,

"*While superintelligence seems far off now, we believe it could arrive this decade.*

Here we focus on superintelligence rather than AGI to stress a much higher capability level. We have a lot of uncertainty over the speed of development of the technology over the next few years, so we choose to aim for the more difficult target to align a much more capable system."

This will require massive effort from both open source and closed [which includes the likes of OpenAI] because likewise, both have been really making great effort in the advancement of AI. Although, these big tech have computational power, financial resources etc. what they lack is community which open source provides.

Few issues require more immediate attention for lawmakers and policymakers than ensuring the responsible development of advanced artificial intelligence. In Africa alone, for instance, there has been a growing community that has been focusing on natural language understanding through Google ASR and Cohere for AI Open Science initiative to try and improve language understanding in a way communities don't feel left behind. As advances in AI accelerate, narrow systems are matching and surpassing human performance on more complex tasks every day taking in considerations of achievements of different continents and learning how they have achieved so far. Extrapolating these trends to a future in which machine intelligence far surpasses the human mind in speed, scale, and power is not too difficult.

While this prospect sparks the imagination, it also poses real risks and challenges that we must now take seriously if we want to influence the development of such powerful technologies for the benefit of mankind. Most of all, this involves the issue of alignment. Even if humans become far more autonomous and intelligent than we are, how can we design AI systems that respect human values, ethics, and priorities? This is a very complex question, requiring input from philosophers, ethicists, psychologists, social scientists and, of course, AI

researchers themselves. This project requires a level of funding and attention commensurate with its scope and importance.

Regulation of advanced AI should be proactive, not reactive, to support progress and steer it in responsible directions. Laws and policies surrounding AI should be carefully formulated to encourage promising research to solve difficult problems such as value alignment, and not impede that promising research with unrealistic constraints or deterrents. I have to. Bodies like the data protection commission in Kenya but specifically for AI following suits of EU AI Act laws could prove very useful in the near future. Government funding agencies must prioritise research focused on managing risk and ensuring the fair, safe, and ethical development of increasingly sophisticated AI. And we need laws to hold companies that develop and deploy AI accountable for undue risk and harm, and to protect citizens and consumers.

Policy makers have an obligation to understand and prepare for the challenges posed by human and superhuman AI before these systems become pervasive or difficult to curb. We can never fully predict the future, but we know enough about the issues at hand to take them seriously and act now. Advances in AI are one of the most pressing issues of our time and require close collaboration between lawmakers and researchers to ensure the safe, reliable and beneficial development of AI. By addressing this challenge, policymakers can help ensure that human judgement, ethics and compassion remain central to how technology shapes the future. The time to act is now.

Overall, AI safety and ethics should be a top priority of policymakers, academics, and industry leaders today so that we may have more optimistic visions of what the future with advanced AI could hold.

References

OpenAI. (2023, July 5 23). Introducing Superalignment. OpenAI Blog. https://openai.com/blog/superalignment/