MERCATUS CENTER
George Mason University

RESEARCH SUMMARY

On Defining “Killer AI”

Significant public attention has focused on artificial intelligence (AI) and its potential for positive societal changes. But could AI also be exhibiting tendencies that could eventually result in human injury or harm, thus turning an initially benign system into a “killer”? In “On Defining Killer AI,” Nathan Summers and Sergio Coronado introduce an initial framework that can help assess the potential for harm from AI and develop policies and approaches that can mitigate this danger.

CAUSE FOR OPTIMISM

AI systems have been used to better diagnose medical conditions, more efficiently develop pharmaceutical drugs, and increase productivity across a wide range of domains. With the rise of generative models, humans now have more time to focus on higher-level tasks for which AI systems are currently not well suited. The developments currently underway could increase human productivity to an extent previously unforeseen and could likely demonstrate social advances on the scale of the agricultural and industrial revolutions.

But the advances in AI do not come without challenges. Such systems can lead to adverse consequences, harm, or death, even when not explicitly designed to do so. Such developments raise the specter of a “Killer AI.”

WHAT IS A KILLER AI?

The authors define a Killer AI as a system employing artificial intelligence techniques that, either by design or by circumstance, directly lead to physical harm or death.

- This definition distinguishes between physical AI systems and virtual AI systems, with the intent of specifying how the latter can also be directly responsible for harm and death.
- AI systems that were previously incapable of being considered directly responsible for harm can now therefore be assessed more rigorously.
- To that end, the authors propose a framework that weighs the wellbeing of many people over the wellbeing of one or few people and on the basis that injuries not resulting in death can be recovered from more easily than those that do and are thus less severe.

ENSURING THE SAFETY OF THIS EMERGENT TECHNOLOGY

There has been no previous attempt to define the Killer AI phenomenon, even as the technology underlying it continue to become increasingly prevalent in our daily lives. The definition should help encourage further examination of AI and its risks—more comprehensive legislation, more appropriate regulation, more nuanced ethical discussion, and more research on the potential ubiquity of AI systems and how to minimize any harm they cause to the greatest extent possible. The authors seek to emphasize the urgency for further research on this subject.

The ideas presented in this document do not represent official positions of the Mercatus Center or George Mason University.