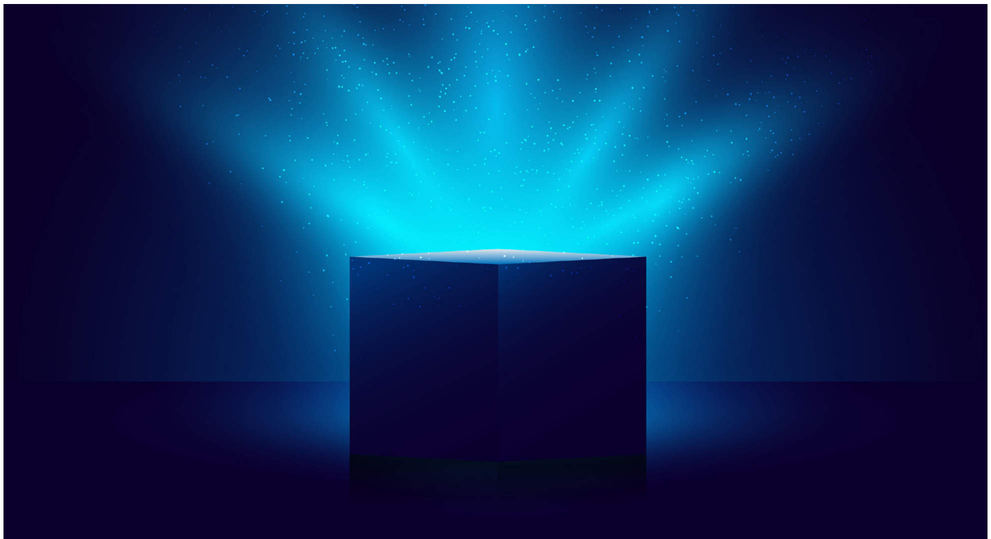
USC Annenberg
Norman Lear Center
Hollywood, Health
and Society

future of life
INSTITUTE

Unboxing AGI

■ Artificial general intelligence (AGI, or “strong AI”)—There is an arms-race among researchers working toward making AI that is as good as the best human experts at nearly all cognitive tasks, including planning, high-level problem solving, and effectively pursuing goals. This is often referred to as “artificial general intelligence” and is the stated goal of several major companies.

■ When (probably, not if) AGI is achieved it may be developed—or develop itself—into AI that is more effective than all of humanity at most tasks; this has been dubbed *superintelligence*. Some researchers estimate that we could see a superintelligent AI within decades or even years.



There Are Several Risks

Loss of Control

AI researchers almost unanimously agree that how to control agents that are much smarter than their creators is an unsolved – and perhaps even unsolvable–problem. So there is a real risk that in creating AGI, and especially superintelligence, humanity will no longer be in control of its own civilization.

AI Empowering Bad Actors

AGI, if controllable, could be used by its controller for negative purposes, including controlling of population, seeking fiscal or political takeover, war, and other purposes.

AI Programmed for Good, but Develops Destructive Methods

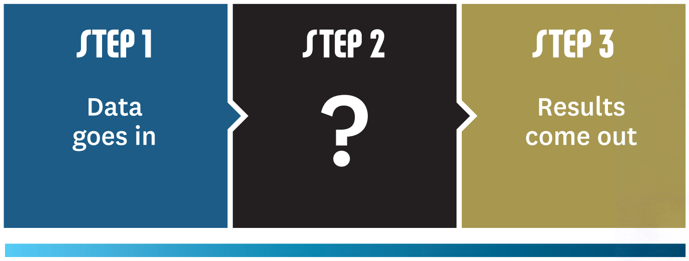
Just as corporations and individuals can do negative things like pollute, exploit, or manipulate in pursuit of a goal (even a neutral or benevolent one), so may AI systems.

AI Built for Devastation

Lethal Autonomous Weapons, aka “Slaughterbots” are AI powered weapons that strike autonomously, without human intervention. Powerful AI systems could be built for military purposes. An AI arms race could escalate into a global catastrophe that humans are unable to stop.

The Black Box Problem

Data scientists, programmers, even the engineers responsible for building the AI don’t know how most deep learning AI systems come to the conclusions they come to. ***In simplest terms, the process looks like this below:***



STEP 1
Data goes in

STEP 2
?

STEP 3
Results come out

The problem with AI isn’t that it’s smart. It’s that it’s stupid in ways we can’t predict.”

—John Oliver, Last Week Tonight, Feb. 26, 2023

Back page for more on the “Black Box Problem”

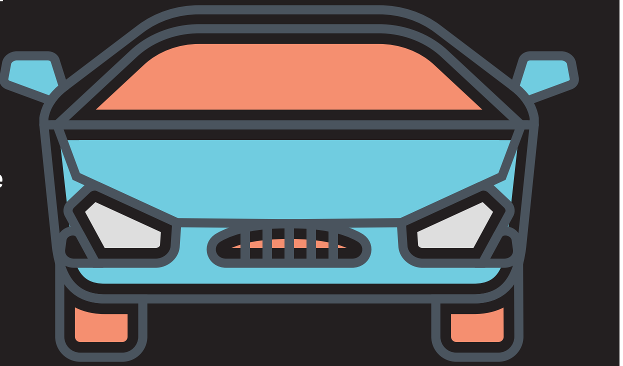
2 Reasons Why the 'Black Box Problem' Is a Big Deal

1

SAFETY

■ It makes it difficult to fix an AI algorithm when it produces unwanted outcomes.

○ Why did the car hit the pedestrian? There could be infinite variables that the AI was considering, and we can't know how it got from one to manslaughter.



2

ETHICS

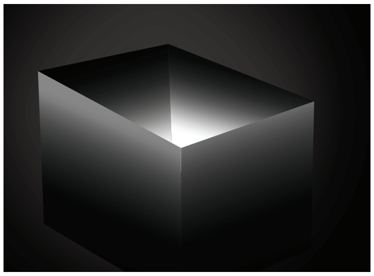
■ If someone is rejected from a job or for a loan but no one knows why or how the algorithm came to that conclusion, one could hardly call the process "fair."

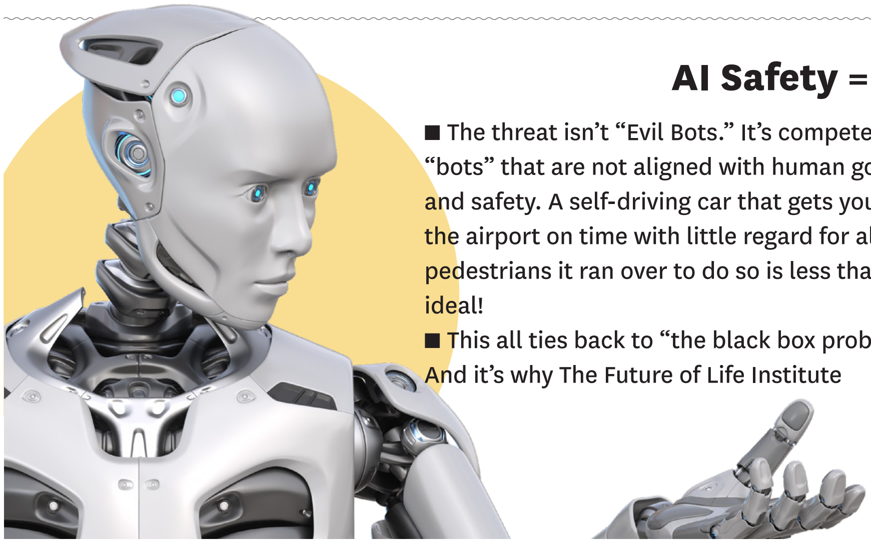
Black Box Solutions

■ One solution includes slowing our use of AI in high-stakes applications where potential for harm is high, such as finance and criminal justice.

■ Another solution is the development of explainable AI that would allow us to peer inside the black box so we can see how the algorithm is coming to its conclusions, making it more fixable and accountable. This field, however, is only just emerging while black box models continue to forge ahead at a lightning pace.

UNIVERSITY OF MICHIGAN





AI Safety = Mutual Alignment

■ The threat isn't "Evil Bots." It's competent "bots" that are not aligned with human goals and safety. A self-driving car that gets you to the airport on time with little regard for all the pedestrians it ran over to do so is less than ideal!

■ This all ties back to "the black box problem." And it's why The Future of Life Institute **penned an open letter** signed by Elon Musk, Steve Wozniak and many top AI researchers calling for at least a six month pause on the development and training of all AI systems more powerful than GPT-4.

■ There are two things we currently do not possess: General AI and a real understanding of how AI works. Figuring out the latter is the only way the development of the former doesn't end in disaster. The race is on.