BROOKINGS

Report

Aligned with whom? Direct and social goals for AI systems

Anton Korinek and Avital Balwit Tuesday, May 10, 2022

Editor's Note:

This is a Brookings Center on Regulation and Markets working paper.

Abstract

As artificial intelligence (AI) becomes more powerful and widespread, the AI alignment problem – how to ensure that AI systems pursue the goals that we want them to pursue – has garnered growing attention. This article distinguishes two types of alignment problems depending on whose goals we consider, and analyzes the different solutions necessitated by each. The direct alignment problem considers whether an AI system accomplishes the goals of the entity operating it. In contrast, the social alignment problem considers the effects of an AI system on larger groups or on society more broadly. In particular, it also considers whether the system imposes externalities on others. Whereas solutions to the direct alignment problem typically center around more robust implementation, social alignment problems typically arise because of conflicts between individual and group-level goals, elevating the importance of AI governance to mediate such conflicts. Addressing the social alignment problem requires both enforcing existing norms on their developers and operators and designing new norms that apply directly to AI systems.

Download the full working paper here.

The Brookings Institution is financed through the support of a diverse array of foundations, corporations, governments, individuals, as well as an endowment. A list of donors can be found in our annual reports published online here. The findings, interpretations, and conclusions in this report are solely those of its author(s) and are not influenced by any donation.