AI Safety and Alignment

Table of contents

Introduction

What is AI Safety and Alignment?

In this lesson, we learn about AI safety and alignment, focusing on how to ensure advanced AI systems behave reliably and in line with human values. We examine technical and governance approaches to preventing unintended or harmful outcomes.

Bibliography