Latest AI News

My AI Model Delta Compared To Christiano by johnswentworth

Source: Featured posts - LessWrong 2.0 viewer
Published: Wed, 12 Jun 2024 18:19:44 +0000

URL:
https://www.greaterwrong.com/posts/7fJRPB6CF6uPKMLWi/my-ai-model-delta-compared-to-christiano

Summary: The article discusses the conceptual distinction between "delta" and "crux" in the context of differing AI models related to decision-making and verification processes. The author introduces "delta" as a localized difference in beliefs between two models that can lead to significant divergences in their outputs. Specifically, this article compares the author's AI model with that of Paul Christiano's, particularly concerning their beliefs about the ease of verification versus generation in problem-solving. The author posits that Paul views verification as generally easier, allowing for the effective delegation of alignment work to AI. In contrast, the author believes that human incompetence creates a bottleneck in delegating tasks effectively, leading to significant, non-obvious flaws in products and services. This fundamental difference in worldview affects their predictions for AI development timelines: the author anticipates rapid advancements and potential discontinuities in effectiveness once AIs surpass human capabilities, while Paul expects a smoother progression. Overall, the article emphasizes how differing assumptions about verification's ease and the implications for AI governance can lead to divergent consequences in practical application and future developments. The author expresses skepticism regarding Paul's stance, asserting a significantly lower probability that his views on this delta are correct.