

THE HONG KONG POLYTECHNIC UNIVERSITY DEPARTMENT OF MANAGEMENT AND MARKETING

Departmental Research Seminar

When AI Alignment Misaligns: Output Quality, Distribution, and Welfare on AI Platforms By



Prof. Juan FENG
Tsinghua University

Date : 7 Oct 2026 (Wed)

Time : 2:30 pm – 4 pm

Venue : HHBUG06, UG/F, PolyU Hung Hom Bay Campus

Abstract

AI alignment has emerged as a key strategy through which AI platforms make outputs better match user preferences based on user feedback. Yet user feedback varies in both quality and the preferences it reflects, making it important to understand how AI alignment reshapes the vertical quality and horizontal distribution of AI outputs and whether it benefits both platforms and users. To examine these issues, we develop a two-period game-theoretic model. In the first period, the platform chooses its initial training effort, while users provide input to the AI and give heterogeneous feedback on the resulting outputs. In the second period, the platform aligns its AI system with this feedback, after which users interact with the aligned system and choose their input effort again. We show that AI alignment affects AI outputs along two dimensions. Horizontally, the aligned output distribution may become polarized or centralized, or remain uniform, with the first two cases reducing output diversity. Vertically, alignment reduces second-period output quality when low-quality feedback has a sufficiently strong adverse effect on algorithm quality. Alignment also induces the platform to substitute distributional optimization for initial training: After adopting AI alignment, the platform always lowers its training effort to save initial training costs. This cost saving, however, does not necessarily improve profitability. AI alignment reduces platform profit when the aligned output distribution is polarized and the AI's preference-matching capability is sufficiently low, and increases platform profit otherwise. Contrary to widespread optimism, AI alignment may also harm users when the reduction in preference misfit fails to offset the loss from lower output quality. Thus, AI alignment can create platform-user conflict, and mutual benefits arise only under certain conditions. Our findings offer important implications for AI platforms and policymakers in assessing when and how AI alignment should be deployed.

Prof. Juan FENG is a Hon Hai Chair Professor at the Department of Management Science & Engineering, School of Economics & Management, Tsinghua University. Prof. Feng holds a B.A. in Economics from Renmin University of China, and a PhD in Business Administration from Pennsylvania State University, with a dual title in Operations Research. Prof. Feng's research interests are in economics of Information systems, focusing on both analytical modeling and empirical analysis. She has been working on topics such as keyword auctions, advertising and pricing, the economics of online review, block chain and data ownership, etc.

All interested are welcome.



THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學

Department of
**MANAGEMENT
& MARKETING**
管理及市場學系