Will Al Accelerate the Intra-Firm Digital Divide? Evidence From a Field Experiment on How Managers Use Explainable Al

Selina Carter
Carnegie Mellon University
shcarter@andrew.cmu.edu

Jonathan Hersh
Chapman University
hersh@chapman.edu

Abstract:

Recent advances in machine learning have created an “AI skills gap” both across and within firms. Approximately three fourths of managers report that addressing digitization skills gap is a top priority for their firm. As AI becomes increasingly embedded in firm processes, it is unknown whether this will accelerate or decelerate the digital divide between workers with AI skills and those without.

In this paper we ask whether managers trust AI to predict consequential events, what manager characteristics are associated with increasing trust in AI predictions, and whether explainable AI affects users’ trust in AI predictions. Explainable AI includes a number of recent advances in machine learning attempting to help AI models explain why they make certain decisions. Partnering with a large bank, we generated AI predictions to predict whether a loan will be late in its final disbursement. We embedded these predictions into a dashboard, surveying 685 analysts and managers before and after viewing the tool to determine what factors affect workers’ trust in AI predictions.

We further randomly assigned some managers and analysts to receive an explainable AI

treatment that adds both global and local explainability to the model prediction screens. For global predictions we focus on global feature importance measures (feature variable importance and partial dependence plots) and for local measures of importance we include Shapely waterfall plots that describe why an individual observation was classified with a given predicted loan delay.

First we find that more senior employees and self described AI novices are baseline much less likely to trust the AI predictions. However, when these “AI-reluctant” groups are presented with the explainable AI module, they are 5-7 times more likely to trust the AI predictions compared to the same groups that did not have the explainable AI module. We find that the explainable AI module overall leads to a 4.5% increase in AI trust, suggesting that AI experts are not particularly affected by the explainable AI module. These results suggest that workers within firms can converge in their AI and ML aptitude if appropriate tools are provided. These results are important for the design of machine learning systems and for the process of technological determinism within the context of firm digital transformation.