Experimental Evidence That AI-Managed Workers Tolerate Lower Pay Without Demotivation

Mengchen Dong[1]*, Levin Brinkmann[1], Omar Sherif[2], Shihan Wang[3], Xinyu Zhang[3], Jean-François Bonnefon[4], and Iyad Rahwan[1]

[1] Center for Humans and Machines, Max Planck Institute for Human Development

[2] Department of Information & Computing Sciences, Utrecht University

[3] Technische Universität Berlin

[4] Toulouse School of Economics, Centre National de la Recherche Scientifique (TSM-R), University of Toulouse Capitole

dong@mpib-berlin.mpg.de, brinkmann@mpib-berlin.mpg.de, omar.sherif@tu-berlin.de, s.wang2@uu.nl, xinyu.rain@outlook.com, jean-francois.bonnefon@iast.fr, rahwan@mpib-berlin.mpg.de

Abstract

Experimental evidence on worker responses to AI management remains mixed, partly due to limitations in experimental fidelity. We address these limitations with a customized workplace in the Minecraft platform, enabling high-resolution behavioral tracking of autonomous task execution, and ensuring that participants approach the task with well-formed expectations about their own competence. Workers (N = 382) completed repeated production tasks under either human, AI, or hybrid management. An AI manager trained on human-defined evaluation principles systematically assigned lower performance ratings and reduced wages by 40%, without adverse effects on worker motivation and sense of fairness. These effects were driven by a muted emotional response to AI evaluation, compared to evaluation by a human. The very features that make AI appear impartial may also facilitate silent exploitation, by suppressing the social reactions that normally constrain extractive practices in human-managed work.