Employees Adhere More to Unethical Instructions from Human Than AI Supervisors

Lanz et al. (2024)

Context

AI systems increasingly act as supervisors, issuing instructions that shape workplace decisions like hiring, pay, and performance. Existing research is split: AI aversion suggests employees resist algorithmic input in moral domains, while AI appreciation finds people often prefer it. This study tests employee adherence to unethical instructions from AI versus human supervisors across four experiments (N=1,701), incl. an incentivised setting. It identifies perceived mind (the capacity to think and feel) as key mediating mechanism and pinpoints employee characteristics shaping resistance.

Key Insights

Employees push back harder on unethical AI orders

- Across three vignette studies, adherence to unethical instructions was consistently lower under AI supervisors (24-27%) than human supervisors (31-42%)
- An incentivised replication with real money on the line confirmed the pattern (29% AI vs 40% human)
- AI aversion outweighs AI appreciation in moral domains, though one-in-three adherence remains material at AI's deployment scale

Perceived mind drives the resistance

- Employees grant AI less capacity to think, plan, and feel, which weakens the moral authority humans typically grant agents capable of moral reasoning
- When researchers stripped mind perception from a human supervisor, adherence dropped to AI levels, showing mind perception itself drives the gap rather than the AI label
- Even a "high-mind" AI was still rated below any human on mind, suggesting anthropomorphic design (giving AIs human-like features like a name, voice, tone) has a ceiling and cannot fully close the gap

Resistance varies by employee characteristics

- Older and more experienced employees showed stronger resistance to AI orders relative to human orders, while the gap was smaller for younger and less tenured employees
- Employees high in "compliance without dissent" resisted AI orders more, because their deference is bound to human authority specifically rather than authority in general
- Pro-AI employees resisted AI orders more than AI-sceptics, likely because the unethical instruction violated their positive expectations of the technology

Implications

- Where in your operations would a 29% compliance rate with biased AI produce outcomes you could not defend publicly?
- How might your AI's perceived authority shape if employees treat outputs as advice or instructions?
- How do you build critical scrutiny of AI outputs in junior staff, given resistance against unethical AI concentrates in experienced employees?
- Where might incentives in your firm override ethical pushback employees would otherwise apply?

Lanz, L., Briker, R. & Gerpott, F.H. Employees Adhere More to Unethical Instructions from Human Than AI Supervisors: Complementing Experimental Evidence with Machine Learning. J Bus Ethics 189, 625–646 (2024). https://doi.org/10.1007/s10551-023-05393-1