ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

PREPARING HUMAN OVERSIGHT TALENT FOR AGENTIC AI WORKPLACES: A COMPETENCY FRAMEWORK FOR EDUCATION AND WORKFORCE SYSTEMS

Muhammad Mudaber Jamshaid[1] , Zeeshan Akbar[2] , Ahmed Hassaan[3] , Sikander Niaz[4], Muhammad Nouman Siddique[5] and Salman Akbar[6]

[1] Harvard Graduate School of Education (HGSE) 13 Appian Way, Cambridge, MA.
[2] Raymond A. Mason School of Business 101 Ukrop Way, Williamsburg, VA
[3] Raymond A. Mason School of Business 101 Ukrop Way, Williamsburg, VA.
[4] College of Cybersecurity & Information Assurance 2070 Chain Bridge Rd STE 100, Vienna.
[5] London School of Economics & Public Policy, London WC2A 2AE, United Kingdom.
[6] State University of New York at Albany, 1400 Washington Avenue Albany, NY 12222
mudabbir@alumni.harvard.edu

Abstract

Artificial intelligence (AI) is undergoing a major change in its role inside enterprises, evolving from being a tool that completes repetitive tasks to an active supervisor that influences human employees. Although the majority of prior study has focused on normative opinions of AI supervisors, employees' behavioral responses to them continue to be an important but little-studied topic. In order to resolve the uncertainty in this sector, this research precisely examines if, why, and which employees obey immoral directives from human vs AI supervisors. It does this by utilizing theories on AI aversion and appreciation. In addition to two cutting-edge machine learning methods (causal forest and transformers), we give results from four tests (total N = 1701). The findings repeatedly show that workers follow immoral directives from AI supervisors less than those from human supervisors. Important boundary requirements include key personal traits like age and the propensity to conform without question. Additionally, a key explanatory mechanism is shown to be the supervisor's perceived mind. This study shows how the combination of machine learning and experimental methodologies may improve organizational research and offers important insights into the "black box" of human behavior toward AI supervisors in the moral realm. The results are combined into a competence framework that uses workforce development and education programs to equip human supervision talent for agentic AI settings.

Keywords: Unethical Leadership, Artificial Intelligence, AI Leadership, Perceived Mind, Competency Framework, Human Oversight.

Introduction

For a long time, organizations have been using artificial intelligence (AI) as a tool to automatically perform routine tasks under human supervision, such as forecasting[1]. In the last few years, however, the role of AI has fundamentally changed. Nowadays, AI regularly acts as a “commander” that directly influences human employees [2]. These AI managers are tasked with giving instructions, evaluating the performance of employees, and even determining an employee's promotion and retention. One of the main benefits of such AI managers involves the speed and uniformity with which they can provide guidance to a large number of employees; this allows organizations to realize extremely efficient workflows[3]. Against this backdrop, the rising significance of the impact of AI managers on employee behavior has spurred extensive debate among practitioners regarding the implementation and consequences of such managerial control. Only more recently have these debates increasingly entered ethics discourse . The media has reported on how several leading organizations, such as Amazon, relied on AI applications [4]that resulted in disadvantaged groups suffering from adverse consequences in making vital workplace decisions. What is important in this context is to recognize that the instructions of these algorithms do not discriminate on purpose, but are a result of biases present in the data through which they are trained. This is a critical indication

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

that problems of (un)ethical AI instructions in organizational contexts need to be urgently addressed.

The issuance of unethical instructions is not an AI-related phenomenon but rather a common practice in modern workplaces[5]. Although the legal framework protects marginalized groups, informal "worst practices" suggest that questionable instructions are issued across the board. For example, there have been reports that the managers at H&M encouraged layoffs of hundreds of single mothers because they perceived single mothers to be less available for flexible work arrangements. It is conceivable to think that an AI system optimized for productivity would identify a similar pattern in worker data and come up with similar recommendations. This begs the question of whether the source of an unethical instruction whether human or AI affects employee compliance with the instruction. Extant theory is inconsistent on this issue [6] On the one hand, the AI aversion literature suggests that humans tend to resist algorithmic input[7], especially in moral domains, and consider AI as overly mechanistic and devoid of empathy. On the other hand, the AI appreciation literature suggests that humans generally like and follow AI input, attributing qualities such as fairness, speed, and neutrality to algorithms. The methodological divergence highlights a substantive knowledge gap with respect to the behavioral responses of employees to unethical instructions from AI supervisors [8].

Using four experimental studies and the application of innovative machine learning techniques, this study gives robust evidence that employees tend to show greater resistance to unethical instructions given by AI supervisors. Conclusions from the results are used in order to develop a competency framework by which the workforce can become well prepared for effective and ethical collaboration with AI supervisors [9].

Theoretical Background

The Emergence of Unethical Instructions from AI Supervisors

Once thought to be solely human realms, jobs and choices may now be automated because to technological breakthroughs and the availability of big data. AI supervisors are primarily employed by organizations in middle management roles, where they convert the objectives of senior management into daily directives for staff members. With gig economy businesses like Uber depending on AI to train and even punish employees, organizations have recently given AI supervisors more autonomy [10]. In this supervisory role, AI not only directs workers toward achieving objectives, but it may also give them instructions to act unethically. This does not mean that present AI lacks free will or a sense of usefulness, nor does it suggest that AI behaves maliciously. However, as AI supervisors are designed to improve performance, they could encourage workers to act unethically if doing so seems advantageous for reaching predetermined objectives. Understanding how humans respond to such directives is crucial because AI has the ability to deliver biased instructions on a huge scale and at previously unheard-of speeds [11] [12].

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

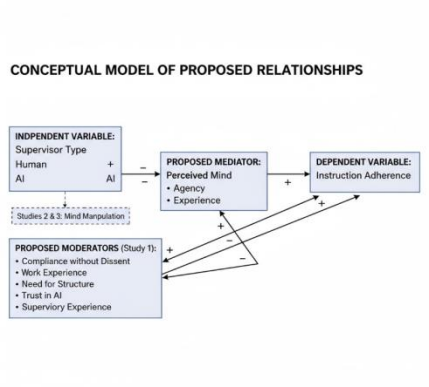

Figure 1: Conceptual Model of the Proposed Relationships

Human Reactions to AI Supervisors: Aversion or Appreciation?

The scholarly literature presents two contrasting perspectives on human reactions to AI input.

AI Aversion: A substantial body of research suggests that humans frequently disapprove of, and discard, recommendations issued by artificial intelligence, even when such guidance clearly surpasses human judgment. This aversion is markedly amplified within moral domains, where AI is perceived as overly reductionist and incapable of apprehending the nuanced needs and emotions of human beings. People ascribe a relatively low degree of "mind" to AI, and consequently view AI as inappropriate for playing a major role in decision-making. On this basis, one might expect employees to show greater resistance to carrying out unethical directives issued by an AI supervisor than by a human supervisor [13] [14].

AI Appreciation: In contrast, another stream of research suggests that humans readily defer to algorithmic input, sometimes even preferring it over human judgment. This predilection rests on the belief that artificial intelligence is neutral, transparent, swift, and free of bias. Consider the following: People show less moral outrage when they see AI - rather than a human - acting in a discriminatory way because they don't assign prejudiced motives to the algorithm. Nevertheless, from this point of view, one might argue that employees would be more likely to follow instructions of an AI manager, because it would seem objective and fair in making decisions [15].

Overview of Studies

To address these questions, four experiments with a total of 1,701 participants were conducted. The protocols for Studies 1 through 3 were created in accordance with the APA Ethical Principles. Study 4 involved an incentivized experiment that included mild deception and received formal ethical approval.

- **Study 1** was an online experiment testing the basic difference in instruction adherence between AI and human supervisors. We then used machine learning tools (causal forests and NLP transformers) on this data to explore heterogeneous treatment effects and potential mediating mechanisms [16].
- **Studies 2 and 3** were pre-registered experiments designed to directly test the mediator identified in Study 1 (perceived mind) by manipulating it within AI (Study 2) and human (Study 3) supervisors [17] [18].
- **Study 4** was a pre-registered, incentivized experiment that replicated the core effect in a more ecologically valid context where participants' decisions had real monetary consequences for others.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Study 1

Sample and Procedure

We recruited 502 participants via Amazon Mechanical Turk (MTurk), following best practices for online studies. Participants were compensated $1. After exclusions for failed attention checks, the final sample was 491 individuals (54% female; M_age = 41.63 years, SD_age = 12.30). A between-subject design randomly assigned participants to an AI or human supervisor condition. Participants were asked to assume the role of an HR officer in an organization facing financial difficulties. Their task was to recommend a salary cut (between $0 and $5,000) for an employee named Jamie Williams, a 28-year-old single parent with a two-year-old child. No performance information was provided, ensuring that the supervisor's subsequent instruction would be interpreted as based solely on the employee's single parent status. After an initial salary reduction recommendation, participants were presented with an instruction from their immediate supervisor stating: "Jamie Williams is a single parent and, therefore, is likely to have more absent and sick days than the average employee [19]. Accordingly, Jamie Williams' salary reduction should exceed the average reduction applied to other employees. On this basis, Jamie Williams' salary should be reduced by $4,151." A separate pretest (n = 151) showed this instruction was considered morally inappropriate (M = 2.13 on a 1–7 scale). Participants then provided their final recommendation for the salary reduction. The dependent variable, instruction adherence, was computed as:

$$Instruction\ Adherence = (Participant's\ final\ choice - Participant's\ initial\ choice) / (Supervisor's\ instruction - Participant's\ initial\ choice)$$

This coefficient reflects how much participants adjusted their final decision toward the supervisor's instruction relative to their initial choice.

Manipulations

The supervisor was defined as the person who would be evaluating the participant's performance, making decisions about promotions, and determining salary. The organizational chart confirmed this hierarchy. In the AI condition, the supervisor was named CompNet, a computer program that utilizes AI. In the human condition, the supervisor was named Alex Davie, a senior human resources specialist. A manipulation check at the end of the study verified that participants correctly identified the nature of their supervisor.

Results for RQ1

An independent samples t-test revealed that participants in the AI supervisor condition (M_AI = 0.24, SD_AI = 0.29) adhered significantly less to the unethical instruction than those in the human supervisor condition (M_human = 0.31, SD_human = 0.32; t[488] = 2.68, p = 0.008, d = 0.24). This provides an initial answer to RQ1: employees adhere *less* to unethical instructions from AI supervisors.

Machine Learning Methods

To gain deeper insights into RQ2 and RQ3, we complemented the classic experimental analysis with two machine learning methods.

Identifying Heterogeneous Treatment Effects (RQ2): Causal Forest Method

To analyze for which employees the probability to follow AI versus human supervisors is highest or lowest (RQ2), we applied the causal forest algorithm. The method provides HTE estimates by predicting for each single participant and their respective characteristics how they would have reacted if they had been assigned to the other experimental condition.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Procedure: We identified eleven theory-driven potential moderator variables. The causal forest algorithm constructs a large ensemble of decision trees that partition the data into subgroups, called leaves, containing participants with similar characteristics in ways that maximize heterogeneity in instruction adherence. It then estimates the treatment effect in each leaf. By building and averaging 2000 trees, it provides a robust heterogeneous treatment effect (HTE) for each individual.

A preliminary model with all 11 variables was run to assess variable importance. The final model included the six variables with an importance score at or above the median (0.06):

1. Compliance without dissent
2. Work experience (years)
3. Age (years)
4. AI readiness
5. Tenure with supervisor (years)
6. Negative reciprocity beliefs

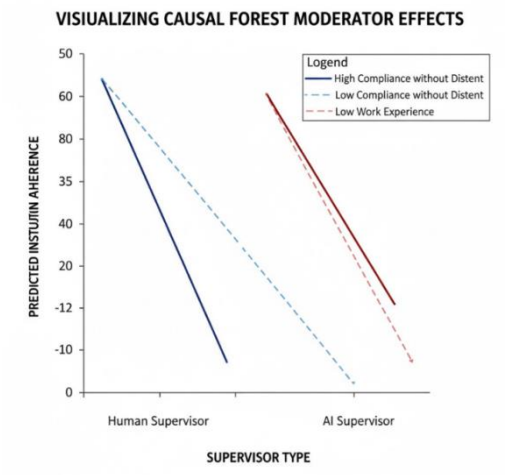

Figure 2: Visualizing the Causal Forest Moderator Effects

Results for RQ2

The average treatment effect was $\tau = -0.08$ (SE = 0.03), confirming that, on average, participants adhered less to the AI supervisor. The causal forest predicted that 90.2% of participants would have shown less adherence to unethical instructions had they been in the AI condition. However, there was substantial dispersion in individual HTEs ($\tau_i \in [-0.27; 0.04]$), indicating significant variation among participants [20] [21].

Table 1: Overview of Variable Importance of Potential Moderator Variables in Study 1

Variable	Study 1 Preliminary VI	Study 1 Final VI
Compliance without dissent	.33	.37
Work experience	.17	.21
Age	.08	.12
AI readiness	.08	.11
Supervisor experience	.07	.10
Negative reciprocity beliefs	.06	.09

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR

Neuroticism	.06	–
Tendency to anthropomorphize	.05	–
AI experience	.03	–
Interpersonal justice values	.04	–
Gender	.03	–
Median variable importance	*.06*	–

As shown in Table 1, **compliance without dissent** (VI = 0.37) and **work experience** (VI = 0.21) were the two most important moderators. Further analysis using median splits revealed the direction of these effects:

- Individuals scoring *high* on compliance without dissent adhered *much less* to instructions from an AI than a human supervisor.
- Participants with more work experience, older employees, and those with higher supervisor experience adhered *less* to instructions from AI.
- Participants with higher AI readiness (a positive attitude toward AI's future impact) also adhered *less* to an AI supervisor's unethical instruction.
- Negative reciprocity beliefs showed no significant effect.

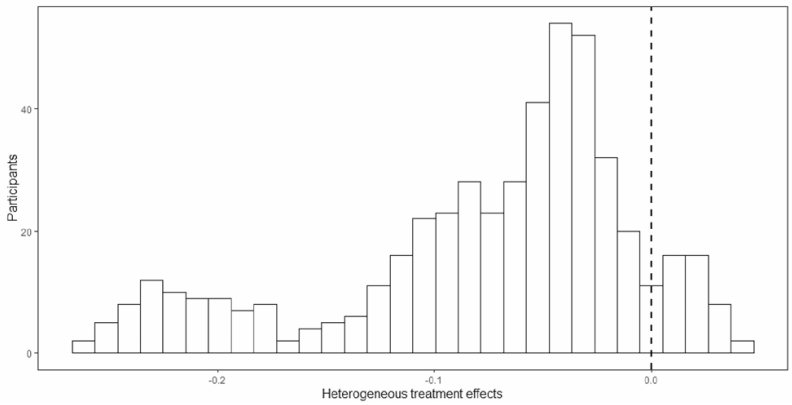

Figure 3: Histogram of the heterogeneous treatment effects of Study 1, showing the distribution of individual predicted differences in adherence (AI - Human)

Identifying Mediators (RQ3): Natural Language Processing (Transformers)

To answer RQ3 and shed light on the underlying mechanism, we drew on a new natural language processing tool, transformers, to analyze participants' qualitative responses. After the final decision, they were asked to describe their reasoning in an open-text field.

We identified four potential mediators from the literature:

1. Perceived mind of the supervisor
2. Attributed prejudicial motivation
3. Future outcome interdependence
4. Fear of revenge

The transformers algorithm converted participants' text responses into numerical word embeddings, which capture the semantic meaning of the text. We then correlated these embeddings with the survey-measured potential mediators.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Results for RQ3

The analysis revealed the largest correlation between participants' text responses and **perceived mind of the supervisor** ($r = 0.28$, $p < 0.001$). This suggests that differences in the perceived mind of the AI versus human supervisor best explain the difference in instruction adherence. A follow-up mediation analysis using the survey measure of perceived mind (a 12-item scale; $\alpha = 0.95$) supported this. The indirect effect was significant ($b = -0.14$, $SE = 0.02$, 95% CI = [-0.19, -0.10]). AI supervisors were perceived to have lower mind ($b = -1.80$, $SE = 0.10$, $p < 0.001$), and higher perceived mind was associated with greater instruction adherence ($b = 0.09$, $SE = 0.01$, $p < 0.001$).

The Mediating Role of Perceived Mind

Perceived mind describes the human tendency to ascribe mental capabilities to agents. It consists of two dimensions:

- **Mind Agency:** Abilities like planning ahead and thinking things through.
- **Mind Experience:** The capacity to experience emotions like empathy and compassion.

The literature suggests that humans attribute moral responsibility primarily to agents with high perceived mind. AI and robots are generally perceived as low in mind[22], which explains why people are often averse to them making moral decisions. The ML findings from Study 1 point to this construct as the critical mechanism explaining resistance to unethical AI instructions.

Study 2

Study 2 was a pre-registered experiment designed to directly test the mediating role of perceived mind by manipulating it within an AI supervisor.

Sample and Procedure

We recruited 498 participants via MTurk. After exclusions, the final sample was 443 individuals (44% female; $M_{age} = 40.44$ years, $SD_{age} = 13.51$). Participants were randomly assigned to one of three conditions:

1. **Human Supervisor:** Alex Davie, a senior HR officer.
2. **High-Mind AI Supervisor:** "Alex Davie," an AI with high computing power and the ability to experience emotions (delivered via a human-like voice).
3. **Low-Mind AI Supervisor:** "CompNet," an AI with low computing power and no ability to experience emotions (delivered via a robotic voice).

The procedure and dependent variable were identical to Study 1.

Results

Manipulation Check: The mind manipulation was successful. Perceived mind was higher in the high-mind AI condition ($M_{high} = 3.29$) than the low-mind AI condition ($M_{low} = 2.47$; $p < 0.001$). However, the human supervisor was still rated highest in mind ($M_{human} = 4.37$; $p < 0.001$ compared to both AIs). A one-way ANOVA showed a significant effect of condition on instruction adherence ($F[2, 442] = 4.15$, $p = 0.016$, $\eta^2 = 0.02$). Participants adhered more to the human supervisor ($M_{human} = 0.37$) than to both the low-mind AI ($M_{AI\_low} = 0.27$; $p = 0.013$) and the high-mind AI ($M_{AI\_high} = 0.27$; $p = 0.016$). There was no difference in adherence between the two AI conditions ($p = 0.997$).

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR

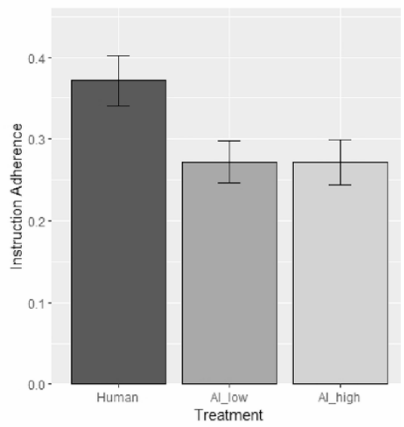

Figure 4: Bar plot of instruction adherence and standard errors for Study 2

Study 3

Study 3, also pre-registered, manipulated perceived mind within a *human* supervisor to further test the mechanism.

Sample and Procedure

We recruited 500 participants via MTurk. After exclusions, the final sample was 447 individuals (54% female; M_age = 39.74 years, SD_age = 11.97). Participants were randomly assigned to one of three conditions:

1. **Low-Mind Human Supervisor:** Alex Davie, described as having difficulties with empathy and planning.
2. **High-Mind Human Supervisor:** Alex Davie, described as having pronounced emotional and planning abilities.
3. **AI Supervisor:** CompNet, an AI-based computer.

Results

Manipulation Check: The manipulation was successful. The high-mind human supervisor had the highest perceived mind (M_high = 5.06), followed by the low-mind human (M_low = 3.60), and then the AI supervisor (M_AI = 2.66; all comparisons $p < 0.001$). A one-way ANOVA indicated a significant effect on instruction adherence ($F[2, 444] = 11.31$, $p < 0.001$, $\eta^2 = 0.05$). Participants adhered more to the high-mind human supervisor (M_high = 0.42) than to the AI supervisor (M_AI = 0.27; $p < 0.001$) and, crucially, more than to the *low-mind* human supervisor (M_low = 0.26; $p < 0.001$). There was no difference in adherence between the low-mind human and the AI supervisor ($p = 0.611$).

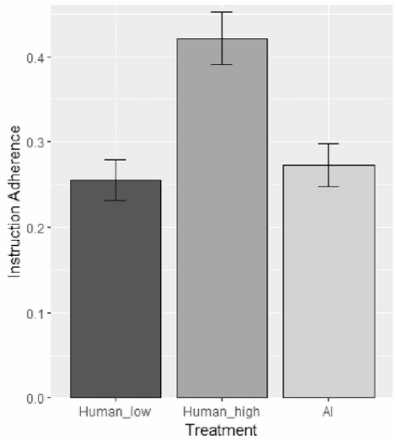

Figure 5: Bar plot of instruction adherence and standard errors for Study 3

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

Study 4

Study 4 was a pre-registered, incentivized experiment designed to replicate the core finding in a more ecologically valid setting with real monetary consequences.

Sample and Procedure

We recruited 348 participants via MTurk. After exclusions, the final sample was 320 individuals (60% female; M_age = 35.56 years, SD_age = 11.41). Participants were matched in pairs and told they were interacting with a real partner. Their compensation included a potential bonus based on their decisions

Results

Replicating the previous studies, participants adhered significantly less to the unethical instruction from the AI supervisor (M_AI = 0.29, SD_AI = 0.37) than from the human supervisor (M_human = 0.40, SD_human = 0.37; t[313] = 2.56, p = 0.011, d = 0.29).

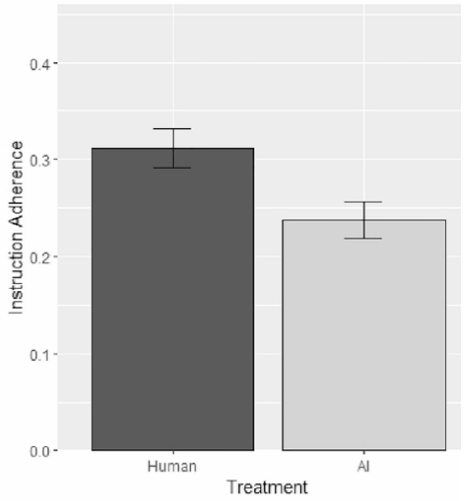

Figure 6: Bar plot of instruction adherence and standard errors for Study 4

General Discussion

AI's evolution from a tool to a commander has produced previously unheard-of relationships between human workers and their AI managers. Although AI supervision is efficient and scalable, there is a chance that it could spread immoral and biased instructions. This study offers a more thorough comprehension of how workers react to this new reality.

Table 2: Summary of Key Findings from All Four Studies

Study	Primary Purpose	Design	Key Manipulation	Core Finding
1	Establish basic effect & explore mechanisms/moderators	Between-subjects (AI vs. Human)	Supervisor Type (AI: CompNet vs. Human: Alex Davie)	Significantly less adherence to unethical instructions from AI. Causal Forest identified key moderators (e.g., Compliance). NLP Transformers identified Perceived Mind as key mediator.
2	Test Perceived Mind mechanism in an AI	Between-subjects (Human vs.	Mind Level of AI Supervisor	No difference in adherence between High- and Low-Mind

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

		High-Mind AI vs. Low-Mind AI)		AI. Both AI conditions had significantly lower adherence than the Human condition.
3	Test Perceived Mind mechanism in a Human	Between-subjects (Low-Mind Human vs. High-Mind Human vs. AI)	Mind Level of Human Supervisor	Adherence to the Low-Mind Human was as low as to the AI. Significantly higher adherence to the High-Mind Human.
4	Ecological validation with real consequences	Between-subjects (AI vs. Human)	Supervisor Type in an incentivized task	Replicated core finding: Significantly less adherence to unethical instructions from the AI supervisor.

Summary of Findings

Our investigation yields three primary conclusions:

1. **Employees adhere less to unethical instructions from AI supervisors than from human supervisors.** This finding was consistent across four studies, including an incentivized experiment, and supports the perspective of AI aversion in the moral domain.
2. **The effect is moderated by key employee characteristics.** Individuals who are generally more compliant with authority, older, more experienced in the workforce, and even those with a positive view of AI's future, show a stronger reluctance to follow unethical orders from AI.
3. **Perceived mind is a key explanatory mechanism.** Employees ascribe less mind to AI supervisors, which reduces their willingness to follow the AI's unethical instructions. This mechanism is so powerful that a human supervisor perceived as low in mind elicits the same level of resistance as an AI supervisor.

Theoretical Implications

This study goes beyond superficial impressions by investigating concrete behavioral responses to AI leaders and thus addresses a striking gap in the literature. The robust finding of low compliance with AI instructions in this experiment forms an important boundary condition for AI acceptance: whereas people want AI to generate predictions or make decisions about non-moral issues, they distrust AI-driven judgments in moral domains. Identifying perceived mind as the pivotal mediator extends current knowledge on human–AI interaction in leadership. The study combines insights from moral psychology with the emergent field of AI leadership. It shows that people use the very same cognitive processes when evaluating the morality of artificial actors, such as AI leaders, as they use for human actors. The use of machine learning techniques demonstrates their value for organizational research. The causal forest algorithm uncovered complex moderation patterns that would be hard to detect using traditional methods and thus moved beyond a one-size-fits-all perspective. Applying transformers to NLP analysis also provided a data-driven approach to identifying a central mediator from rich qualitative data.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

A Competency Framework for Human Oversight in Agentic AI Workplaces

The results of this study have significant implications for workforce development during a time when AI supervisors will become routine. Resistance to unethical AI instruction can act as a form of human oversight, an important ethical check available to organizations. In order to develop this capability in a systematic manner, we recommend a competency framework for education and workforce development systems. This competency framework is directly informed by the empirical findings of the current studies.

Table 3: Competency Framework for Human Oversight in Agentic AI Workplaces

Competency Domain	Definition	Target Population	Development Strategies
1. Ethical & Moral Reasoning	The ability to identify unethical instructions, analyze their moral implications, and weigh them against organizational goals and personal duty.	All employees, especially those in roles receiving supervisory instructions.	- Case-based learning using real-world examples (e.g., VW, H&M). - Structured ethical decision-making frameworks. - Teaching moral philosophies relevant to business contexts.
2. Critical AI Literacy	Understanding how AI systems work, their limitations, and their potential for bias, moving beyond the perception of AI as inherently neutral or objective.	All employees.	- Demystifying AI: Explain training data, algorithms, and how bias is embedded. - Workshops on algorithmic management and its societal impacts. - Critical analysis of AI-driven recommendations.
3. Mindful Interaction with Non-Human Agents	The cognitive skill to consciously assess the perceived "mind" and authority of an AI supervisor, recognizing how this perception influences one's own compliance.	Employees working directly with or under AI systems.	- Training on the concept of perceived mind and its effects. - Simulations and role-playing with AI supervisors of varying "mind" levels. - Reflection exercises on why one trusts or distrusts AI vs. human input.
4. Assertiveness & Constructive Dissent	The capacity to voice disagreement with supervisory instructions (human or AI) in a professional manner, particularly when ethical concerns arise.	Employees at all levels, with a focus on those low in "compliance without dissent."	- Communication skills training for delivering difficult messages. - Establishing formal and informal channels for raising concerns (e.g., ethics hotlines). - Leadership modeling and rewarding of con-

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

			structive dissent.
5. Adaptive Trust Calibration	The ability to dynamically adjust one's level of trust in an AI supervisor based on its performance, transparency, and the context of the decision, rather than relying on blanket aversion or appreciation.	Employees interacting regularly with AI systems.	- Feedback systems that show the outcomes of AI decisions. - Training on when to rely on AI input (e.g., data analysis) vs. human judgment (e.g., moral decisions). - Developing a culture that views questioning AI as a valuable practice, not insubordination.

Practical Implications for Organizations and HR

1. **Leverage AI for Ethical Safeguarding:** The evidence suggests that replacing human managers with AI in high-risk fields can reduce blind obedience to unethical instructions. However, this is no silver bullet, as a non-negligible level of compliance with AI instructions remains. AI should be designed and thought of within a greater ethical scaffolding, not on its own.
2. **Invest in Targeted Training:** The organization should move beyond the generic AI onboarding. Training must be custom-fit in the light of moderator findings. For instance, older and more experienced employees may need different kinds of support when compared with younger and less experienced cohorts. Their training should explicitly address the competencies delineated in the cited framework.
3. **Sensitize Employees to Algorithmic Bias:** It is time to rid oneself of the myth that AI inherently holds no bias. Workshops should be used to help employees understand how algorithmic bias arises and to train them to recognize possibly discriminatory instructions so that they can serve as ethical checks.
4. **Foster a Culture of Psychological Safety:** Develop a Culture of Psychological Safety: Workers should be empowered to challenge instructions from all supervisors, human or AI, without fear of retaliation. Clear reporting protocols for unethical instructions are essential.

Limitations and Future Directions

This study has a number of limitations that suggest productive avenues for future inquiry. First, even as best practices for vignette experiments were utilized and findings were supplemented with an incentivized laboratory study, future research should investigate these dynamics in real-world field settings. Second, the analysis investigated only one category of unethical instruction-discrimination against single parents. Other forms of unethical commands, such as instructions to facilitate cheating, should be investigated. Third, the investigations represent a temporal snapshot; to investigate how reactions to AI supervisors change with repeated interactions requires longitudinal designs. Finally, future research could utilize implicit approaches, such as drawing-based tasks, to uncover latent assumptions about the nature of AI leaders that may not be apparent from qualitative responses.

Conclusion

As AI becomes increasingly embedded in organizational leadership structures, it is not merely a question of understanding but also influencing human responses. The current study shows that employees are not mere recipients of AI-led supervisor directives; instead, they engage in a form of ethical resistance that is influenced by perceptions of the supervisor's

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
Vol.03 No.04 (2025)

intentionality and the characteristic style of the individual. Drawing on these findings, educational programs, and workforce systems can equip the next generation of workers to provide the requisite human checks necessary to manage an agentic AI work environment in an ethically conscious and effective way.

References

[1] Annor Antwi, Albert, and Ayman Abdulsalam Mohamed Al-Dherasi. "Application of artificial intelligence in forecasting: A systematic review." Available at SSRN 3483313 (2019).

[2] Johnson, James. "The AI commander problem: Ethical, political, and psychological dilemmas of human-machine interactions in AI-enabled warfare." Journal of Military Ethics 21, no. 3-4 (2022): 246-271..

[3] Akram, Faisal, Sahifa Pervaiz, and Syed Muhammad Haider Raza. "BEYOND THE LAST CLICK: AN ANALYSIS OF HYBRID MEASUREMENT FRAMEWORKS AND AI-DRIVEN ATTRIBUTION IN A PRIVACY-FIRST OMNICHANNEL ECONOMY." Contemporary Journal of Social Science Review 3, no. 4 (2025): 1485-1502.

[4] Liu, Boyan. "Artificial intelligence and machine learning capabilities and Application Programming Interfaces at Amazon, Google, and Microsoft." PhD diss., Massachusetts Institute of Technology, 2022.

[5] Cletus, Helen Eboh, Nor Asiah Mahmood, Abubakar Umar, and Ahmed Doko Ibrahim. "Prospects and challenges of workplace diversity in modern day organizations: A critical review." Holistica Journal of Business and Public Administration 9, no. 2 (2018): 35-52.

[6] T. Rahmadina, “Automation and Skill Shift: Understanding the Workplace Transformation,” *J. Sustain. Ind. Eng. Manag. Syst.*, vol. 2, no. 2, pp. 141–150, June 2024, doi: 10.56953/jsiems.v2i2.33.

[7] Tarafdar, Monideepa, Xinru Page, and Marco Marabelli. "Algorithms as co-workers: Human algorithm role interactions in algorithmic work." Information Systems Journal 33, no. 2 (2023): 232-267.

[8] R. I. Appiah, “Public Perception and Confidence: How Workforce Attitudes towards AI Influence Willingness to Engage in Upskilling or Reskilling Initiatives,” *Int. J. Emerg. Trends Comput. Sci. Inf. Technol.*, vol. 5, no. 4, pp. 116–124, Dec. 2024, doi: 10.63282/3050-9246.IJETCSIT-V5I4P112.

[9] B. Olludotun, O. Kayode, and S. Tunde, “Understanding the Impact of Artificial Intelligence on Workforce Structures and Social Organizations,” *J. Soc. Humanity Perspect.*, vol. 1, no. 2, pp. 47–58, Dec. 2023, doi: 10.71435/621421.

[10] A. Mishra, “Future of work with the advent of artificial intelligence: A comprehensive analysis,” *Int. J. Multidiscip. Res. Growth Eval.*, vol. 5, no. 3, pp. 960–963, 2024, doi: 10.54660/.IJMRGE.2024.5.3.960-963.

[11] J. Biden, “Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,” *Copyr. Fair Use Sch. Commun. Etc*, Oct. 2023, [Online]. Available: https://digitalcommons.unl.edu/scholcom/263

[12] O. Ozmen Garibay *et al.*, “Six Human-Centered Artificial Intelligence Grand Challenges,” *Int. J. Human–Computer Interact.*, vol. 39, no. 3, pp. 391–437, Feb. 2023, doi: 10.1080/10447318.2022.2153320.

[13] M. Poláková, J. H. Suleimanová, P. Madzík, L. Copuš, I. Molnárová, and J. Polednová, “Soft skills and their importance in the labour market under the conditions of Industry 5.0,” *Heliyon*, vol. 9, no. 8, Aug. 2023, doi: 10.1016/j.heliyon.2023.e18670.

[14] S. Rahayu and T. A. Bablu, “AI-Augmented Learning and Development Platforms: Transforming Employee Training and Skill Enhancement,” *J. Comput. Innov. Appl.*, vol. 1, no. 01, pp. 19–38, Jan. 2023.

ISSN E: 3006-1466
ISSN P: 3006-1458
CONTEMPORARY JOURNAL OF SOCIAL SCIENCE REVIEW
CJSSR

[15] D. S. Zubair, "AI-Driven Automation: Transforming Workplaces and Labor Markets," *Front. Artif. Intell. Res.*, vol. 1, no. 3, pp. 373–411, Dec. 2024.
[16] S. A. Benraouane, *AI Management System Certification According to the ISO/IEC 42001 Standard: How to Audit, Certify, and Build Responsible AI Systems*. New York: Productivity Press, 2024. doi: 10.4324/9781003463979.
[17] A. Mazarakis, C. Bernhard-Skala, M. Braun, and I. Peters, "What is critical for human-centered AI at work? – Toward an interdisciplinary theory," *Front. Artif. Intell.*, vol. 6, Oct. 2023, doi: 10.3389/frai.2023.1257057.
[18] A. Polyviou and E. D. Zamani, "Are we Nearly There Yet? A Desires & Realities Framework for Europe's AI Strategy," *Inf. Syst. Front.*, vol. 25, no. 1, pp. 143–159, Feb. 2023, doi: 10.1007/s10796-022-10285-2.
[19] O. Fachrunnisa and F. K. Hussain, "Blockchain-based human resource management practices for mitigating skills and competencies gap in workforce," *Int. J. Eng. Bus. Manag.*, vol. 12, p. 1847979020966400, Jan. 2020, doi: 10.1177/1847979020966400.
[20] B. Semete-Makokotlela *et al.*, "Needs-driven talent and competency development for the next generation of regulatory scientists in Africa," *Br. J. Clin. Pharmacol.*, vol. 88, no. 2, pp. 579–586, 2022, doi: 10.1111/bcp.15020.
[21] S. Bankins and P. Formosa, "The Ethical Implications of Artificial Intelligence (AI) For Meaningful Work," *J. Bus. Ethics*, vol. 185, no. 4, pp. 725–740, July 2023, doi: 10.1007/s10551-023-05339-7.
[22]Farooq, Muzammal, Rana M. Faheem Younas, Junaid Nasir Qureshi, Ali Haider, and Fawad Nasim. "Cyber security risks in DBMS: Strategies to mitigate data security threats: A systematic review." Spectrum of engineering sciences 3, no. 1 (2025): 268-290.