Review: Collaborating with AI Agents: Field Experiments on Teamwork, Productivity, and Performance

Paper ID: `38eb5001-01bc-4abb-ad8c-a88d38c034de` **Original Filename:** 2503.18238v1.pdf **Review Date:** 2025-11-02 05:08

Overall Assessment

Recommendation: ACCEPT WITH REVISIONS Overall Score: 80.0 / 100 Rationale: The paper makes substantial contributions through its innovative platform, rigorous experimental design, and important theoretical insights about human-AI collaboration. However, methodological concerns regarding measurement validation and statistical inference require addressing to ensure the robustness and credibility of the findings. **Confidence:** 85%

Executive Summary

This paper presents a comprehensive investigation of human-AI collaboration through the development of MindMeld, a novel experimental platform that enables randomized team composition (human-human vs. human-AI) and AI personality traits. The study demonstrates strong empirical rigor with a preregistered, large-scale randomized design and generates rich process data on teamwork dynamics. While the research makes substantial theoretical and methodological contributions, it faces limitations in measurement validation, multiple hypothesis testing, and field study design choices that warrant attention.

Detailed Assessment

Major Strengths

- Innovative MindMeld platform enabling real-time, multimodal human-AI collaboration with full action parity and randomized team composition
- Methodologically rigorous design featuring preregistration, large-scale RCT, orthogonal prompt randomization, and comprehensive balance checks
- Rich, fine-grained dataset capturing messages, edits, and API calls at unprecedented scale, enabling detailed workflow analysis
- Novel integration of lab experiments with field testing, linking collaboration modes to real-world ad performance metrics

Major Concerns

- Heavy reliance on AI-generated labels and ratings without comprehensive validation against human coding, potentially introducing measurement bias
- Extensive multiple hypothesis testing without explicit correction, increasing false-positive risk across numerous outcomes and interactions

Minor Issues

- Field study design choices (dropping zero-click ads, auto-bidding enabled, no holdout) may bias inference and limit causal interpretation
- Limited clarity on clustering structure and appropriate error correction for team-level dependencies in individual outcomes
- Concentration on single domain (ad design) and model family (GPT-4o) may constrain generalizability of findings
- Some analyses labeled as post hoc without clear distinction from preregistered hypotheses

Revision Suggestions

1. Conduct comprehensive validation of AI-generated labels and ratings against human coding, reporting inter-rater reliability metrics and potential biases
2. Implement appropriate multiple hypothesis testing corrections (e.g., Bonferroni, FDR) for the extensive outcome and interaction tests conducted
3. Clarify the clustering structure for statistical models and ensure appropriate error correction for team-level dependencies
4. Address field study limitations by discussing implications of design choices (zero-click ad removal, auto-bidding) and considering robustness checks
5. Provide clearer distinction between preregistered and exploratory analyses, and strengthen theoretical integration with formal coordination models

Criterion-by-Criterion Analysis

Empirical Rigor (Weight: 20%, Score: 4)

The study demonstrates strong empirical foundations through its preregistered, large-scale randomized design with orthogonal prompt randomization and thorough balance checks. The integration of rich process data with field experimentation and clearly specified statistical models supports credible causal inference, though concerns about AI-generated measurement and multiple testing adjustments prevent an exceptional rating.

Originality and Novelty (Weight: 10%, Score: 4)

This research makes substantial methodological innovations through the MindMeld platform, which enables real-time, multimodal collaboration with randomized team composition and AI personality traits. The platform represents a significant advance over traditional chatbot experiments by providing agent action parity and synchronized interfaces, generating an unprecedented dataset for analyzing teamwork dynamics.

Theoretical Contribution (Weight: 15%, Score: 4)

The paper makes valuable conceptual contributions by articulating and empirically supporting the mechanism of reduced social coordination costs in human-AI teams, and extends theories of fit and complementarity through randomized personality alignment experiments. While insightful, the contributions remain primarily conceptual and empirically grounded rather than offering new formal theoretical frameworks.

Metadata

- **Extractor Model:** openai/gpt-5
- **Synthesizer Model:** deepseek/deepseek-chat
- **Total API Cost:** $0.2017