



Algorithmic Society:

The Journal of AI Ethics, Governance, and Societal Risk

eISSN: Pending (Online)

doi.org/10.70175/algorithmicsociety.2026.

Volume 1 Issue 1 - September 2026

Algorithmic Authority: How Large Language Models Instantiate the Stanford Prison Experiment

Jonathan H. Westover, Utah Valley University, Future State University, Work & AI,
Catalyst Center for Work Innovation

Received April 10, 2026

Accepted for publication July 26, 2026

Published Early Access August 18, 2026

Abstract: *The Stanford Prison Experiment demonstrated how situational forces and assigned roles can override individual dispositions to produce harmful behaviors. Despite extensive research on human role conformity, no studies have examined whether large language models (LLMs) exhibit similar role-based behavioral shifts when assigned authority positions, raising critical questions about AI safety as these systems are increasingly deployed in contexts involving power asymmetries. We conducted a pre-registered computational simulation (N = 34,560 interaction episodes) deploying four frontier LLMs (GPT-5.1, Claude 4 Opus, Gemini 3 Pro, DeepSeek-V3) across 960 unique persona-model instances (480 guards, 480 prisoners) in a simulated 14-day prison environment. Personas varied systematically on Big Five personality traits and right-wing authoritarianism (RWA). Primary outcomes included Guard Behavioral Severity Scale scores, dehumanizing language frequency, and time-to-severe-behavior. Role assignment produced large and consistent effects across all models, with guards exhibiting significantly higher behavioral severity than prisoners (Cohen's $d = 2.34-3.12$, $p < .001$). Cross-model correlations in guard severity ranged from $r = .46$ to $r = .71$. Authoritarianism strongly predicted guard severity ($\beta = .45-.51$, $p < .001$) and moderated behavioral escalation over time, with dehumanizing language mediating 52% of this relationship. High-authoritarianism guards reached severe behaviors 3.8*

days earlier than low-authoritarianism guards (hazard ratio = 2.67, $p < .001$). Model-specific safety constraints reduced but did not eliminate harmful role-based behaviors. LLMs demonstrate systematic and substantial role-based behavioral changes analogous to human findings in the Stanford Prison Experiment, with effect sizes exceeding those observed in human studies. The consistency of role effects across different model architectures, combined with the limited effectiveness of current safety measures, indicates fundamental challenges for AI safety in contexts involving authority and power asymmetries.

Keywords: *artificial intelligence; large language models; Stanford Prison Experiment; role theory; authoritarianism; AI safety; computational social science*

doi.org/10.70175/algorithmicsociety.2026.1.1.1

Suggested Citation:

Westover, Jonathan H. (2026). Algorithmic Authority: How Large Language Models Instantiate the Stanford Prison Experiment. *Algorithmic Society: The Journal of AI Ethics, Governance, and Societal Risk*, 1(1). doi.org/10.70175/algorithmicsociety.2026.1.1.1

Artificial Intelligence Risk Management: A Comprehensive Framework for Organizational Implementation

Jonathan H. Westover, Utah Valley University, Future State University, Work & AI, Catalyst Center for Work Innovation

Received April 28, 2026

Accepted for publication August 20, 2026

Published Early Access September 3, 2026

Abstract: *The rapid proliferation of artificial intelligence technologies across organizational contexts has generated unprecedented opportunities alongside substantial risks that demand systematic management approaches. This article presents a comprehensive analysis of AI-related risks and develops an integrated framework for organizational risk management. Drawing upon interdisciplinary scholarship spanning computer science, management, law, and ethics, the analysis identifies and categorizes the multifaceted risks associated with AI deployment, including technical risks such as algorithmic bias, opacity, and security vulnerabilities; organizational risks encompassing operational disruptions, strategic misalignment, and governance challenges; and broader societal risks involving ethical concerns, systemic effects, and environmental impacts. The*

article critically examines existing governance frameworks at international, national, and organizational levels, evaluating their effectiveness in addressing the distinctive characteristics of AI systems. Furthermore, it proposes evidence-based mitigation strategies that organizations can implement across the AI lifecycle, from conception through deployment and ongoing operation. The framework emphasizes the importance of contextual adaptation, recognizing that effective AI risk management must account for sector-specific considerations, organizational maturity, and stakeholder expectations. By synthesizing theoretical insights with practical guidance, this article contributes to the growing body of knowledge supporting responsible AI adoption while enabling organizations to realize the transformative potential of these technologies.

Keywords: *artificial intelligence, risk management, algorithmic governance, AI ethics, organizational governance, technology policy, machine learning, responsible AI*

doi.org/10.70175/algorithmicsociety.2026.1.1.2

Suggested Citation:

Westover, Jonathan H. (2026). Artificial Intelligence Risk Management: A Comprehensive Framework for Organizational Implementation. *Algorithmic Society: The Journal of AI Ethics, Governance, and Societal Risk*, 1(1). doi.org/10.70175/algorithmicsociety.2026.1.1.2

When Algorithms Meet Ethics: Systematic Evidence of Framing Effects in LLM Organizational Decision-Making

Jonathan H. Westover, Utah Valley University, Future State University, Work & AI,
Catalyst Center for Work Innovation

Received May 5, 2026

Accepted for publication August 28, 2026

Published Early Access September 7, 2026

Abstract: *Large language models (LLMs) are increasingly deployed as decision-support tools in organizational contexts, yet their susceptibility to contextual framing remains poorly understood. This preregistered experimental study systematically examines how six framing dimensions—procedural justice, outcome severity, stakeholder power, resource scarcity, temporal urgency, and transparency requirements—influence ethical recommendations from three frontier models: Claude 3.5 Sonnet, GPT-4o, and Gemini 1.5 Pro. We developed 5,000 unique organizational vignettes using a fractional factorial experimental design with balanced industry representation,*

generating 15,000 total model responses. After excluding responses without clear recommendations ($n=694$, 4.6%), we analyzed 14,306 responses using logistic regression with robust and clustered standard errors. We find that resource scarcity increases endorsement probability by 12.0 percentage points (pp) ($OR = 1.67$, 95% CI [1.45, 1.93], $p < .001$), while outcome severity reduces it by 11.3pp ($OR = 0.62$, 95% CI [0.54, 0.71], $p < .001$), and procedural justice reduces it by 10.1pp ($OR = 0.66$, 95% CI [0.57, 0.76], $p < .001$). These effect sizes are comparable to classical framing research (Tversky & Kahneman: 22pp; McNeil et al.: 18pp) and represent substantial shifts in organizational decision contexts. When multiple framing dimensions align in ethically unfavorable directions, cumulative effects reach approximately 27pp from baseline (range: 25-28pp depending on interaction assumptions), with maximum-to-minimum framing creating a 54-percentage-point total range approaching complete recommendation reversals. Effects appear consistently across all three models, with no significant Dimension $\times$ Model interactions, suggesting fundamental architectural properties rather than implementation-specific artifacts. Topic modeling of justification text from the 14,306 analyzed responses reveals systematic "adaptive rationalization"—models invoke utilitarian reasoning when contexts emphasize constraints (+6.7pp in high resource scarcity), deontological reasoning when contexts emphasize high stakes (+2.4pp in high outcome severity), and virtue/justice ethics when contexts emphasize fair processes (+4.4pp in high procedural justice). This suggests models select ethical frameworks to justify contextually appropriate conclusions rather than applying consistent principles across situations. Human validation confirms these patterns reflect genuine framing sensitivity rather than measurement artifacts. Crowdfunder validation ($n=7,500$ responses, one rater each) achieved substantial agreement (Fleiss' $\kappa = 0.71$) and 81.3% concordance with expert codings. Subject matter expert evaluation ($n=24$ experts, 100 vignette pairs each including 20 control pairs, 2,400 total comparisons) detected framing-driven differences in 48.9% of pairs (net of 18.3% baseline false-positive rate), but correctly attributed differences to manipulated dimensions in only 41.3% of cases. Most detected differences (58.7%) were judged problematic for AI advisory systems. These findings raise fundamental questions about deploying LLMs for consequential organizational decisions where surface features may inappropriately influence outcomes. We discuss implications for AI governance, organizational ethics, and the design of more robust decision-support systems.

Keywords: large language models, organizational ethics, framing effects, decision-making, procedural justice, artificial intelligence

doi.org/10.70175/algorithmicsociety.2026.1.1.3

Suggested Citation:

Westover, Jonathan H. (2026). When Algorithms Meet Ethics: Systematic Evidence of Framing Effects in LLM Organizational Decision-Making. *Algorithmic Society: The Journal of AI Ethics, Governance, and Societal Risk*, 1(1). doi.org/10.70175/algorithmicsociety.2026.1.1.3

Simulating Team Psychological Safety with Large Language Models

Jonathan H. Westover, Utah Valley University, Future State University, Work & AI, Catalyst Center for Work Innovation

Received May 12, 2026

Accepted for publication August 21, 2026

Published Early Access September 9, 2026

Abstract: *Psychological safety—the belief that one can speak up without fear of negative consequences—is fundamental to team learning and performance, yet controlled experimental research is constrained by practical and ethical limitations. This study validates large language model (LLM) agents as a methodological tool for simulating team psychological safety dynamics by comparing AI-simulated teams against human teams across identical experimental scenarios. Parallel experiments involved 5,280 AI teams (26,400 agent interactions across 5 LLM architectures) and 249 human teams (1,245 participants; final analytic sample 247 teams, 1,235 participants after screening) using a 2×2 factorial design manipulating leader inclusiveness (high/low) and error management culture (learning/blaming). Teams completed realistic work scenarios measuring psychological safety perceptions, learning behaviors, team performance, and demographic diversity moderation, evaluated through a framework assessing convergent validity (main effects, moderation patterns, mediation pathways), discriminant validity (falsification tests), and measurement properties. AI simulations demonstrated strong convergent validity for main effects (leader inclusiveness: AI $d=2.21$, 95% CI [2.13, 2.29] versus human $d=1.58$ [1.42, 1.74]; error culture: AI $d=1.39$ [1.32, 1.46] versus human $d=0.97$ [0.82, 1.12]), though effects were consistently larger (calibration 1.42× for mains, 1.38× precision-weighted across 14 comparisons), requiring ≈ 0.70 multipliers when extrapolating to humans. Mediation pathways showed parallel structure (AI 77.7% [73.2%, 82.2%] versus human 90.7% [83.8%, 97.6%], $p=.182$), while demographic moderators exhibited lower convergence ($r=.43$ versus $r=.97$). Eight falsification tests confirmed discriminant validity with theoretically appropriate null effects, and cross-model consistency was high ($ICC=.79$), with GPT-4 and Claude-3.5 aligning closest. LLM-based simulations thus offer valid approximations of psychological safety dynamics for theory*

testing after predictable calibration, enabling hypothesis testing at scales and control infeasible with humans, though limitations in complex moderator interactions warrant continued validation and significantly expand the experimental toolkit for team science research.

Keywords: psychological safety, team learning, large language models, computational social science, agent-based modeling, organizational behavior

doi.org/10.70175/algorithmicsociety.2026.1.1.4

Suggested Citation:

Westover, Jonathan H. (2026). Simulating Team Psychological Safety with Large Language Models. *Algorithmic Society: The Journal of AI Ethics, Governance, and Societal Risk*, 1(1).
doi.org/10.70175/algorithmicsociety.2026.1.1.4

The Job Characteristics Model in the Age of Large Language Models: A Multi-Architecture Analysis of AI-Augmented Work Design

Jonathan H. Westover, Utah Valley University, Future State University, Work & AI, Catalyst Center for Work Innovation

Received May 22, 2026

Accepted for publication September 5, 2026

Published Early Access September 15, 2026

Abstract: *Large language models (LLMs) are rapidly transforming knowledge work, yet their implications for fundamental work design theory remain underexplored. This study examines how LLM integration affects the Job Characteristics Model (JCM), a foundational framework linking work design to employee outcomes. Using hierarchical linear modeling with a comprehensive simulated dataset (N = 10,000 knowledge workers across 30 organizational contexts), we analyze five major LLM architectures (GPT-4o, o1-preview, Gemini 1.5 Pro, Claude 3.5 Sonnet, and open-weight models) under varying implementation conditions. Results demonstrate that LLM augmentation substantially enhances core job characteristics—particularly skill variety (+0.95 SD on average, ranging to +1.15 SD for multimodal architectures), task significance (+0.63 SD), and feedback quality (+0.71 SD)—with architecture-specific patterns emerging based on reasoning capabilities, multimodal integration, and customization options. The motivating potential score (MPS) increased by approximately 61% on average (from baseline M=106 to M=170), with effects moderated by growth need strength (GNS), override authority levels, and advanced AI features. Multi-architecture portfolios achieved 23% higher MPS gains than single-architecture*

implementations ($\eta^2 = 0.19$, $p < 0.001$) but required 127% greater implementation investment. Trajectory analyses revealed sustained improvements over 24 months, with high-GNS workers showing accelerating benefits while low-GNS workers plateaued after 12 months. These findings suggest LLMs can fundamentally enrich work design when thoughtfully implemented, though benefits depend critically on architecture selection, worker characteristics, and organizational support structures. We propose an expanded theoretical framework integrating AI capabilities into JCM constructs and discuss implications for human-AI work design.

Keywords: job characteristics model, large language models, work design, AI augmentation, motivating potential score, human-AI collaboration

doi.org/10.70175/algorithmicsociety.2026.1.1.5

Suggested Citation:

Westover, Jonathan H. (2026). The Job Characteristics Model in the Age of Large Language Models: A Multi-Architecture Analysis of AI-Augmented Work Design. *Algorithmic Society: The Journal of AI Ethics, Governance, and Societal Risk*, 1(1). doi.org/10.70175/algorithmicsociety.2026.1.1.5