Compound AI for Decision Advantage: Human Digital Twin Agent Teams and Populations for Military Operations

Svitlana Volkova, Laura Cassani, Spencer Lynch, Hsien-Te Kao, Peter Bautista and Will Dupree
Aptima, Inc.
Woburn, MA
svolkova@aptima.com, lcassani@aptima.com, slynch@aptima.com, hkao@aptima.com, pbautista@aptima.com, wdupree@aptima.com

ABSTRACT

We present an innovative compound AI approach that integrates DARPA's EMHAT (Evaluation and Modeling of Human Digital Twin (HDT) Agent Teams) and BRIES (Building Resilient Information Ecosystems) efforts to enable sophisticated modeling and rigorous evaluation of agentic AI workflow dynamics in military operational scenarios. Current systems lack the ability to simultaneously model human digital twins with rich psycho-demographic attributes and memory representations and evaluate multi-dimensional performance metrics (Abdurahman et al., 2024). Our compound AI solutions address this gap through specialized multi-agent architectures that provides unprecedented insights into teaming effectiveness and population-level information resilience.

We deploy specialized cognitive language agents across multiple scenarios: EMHAT enables search and rescue simulations with diverse HDT agents, rigorously measuring team processes and state metrics to improve teaming effectiveness. Experimental results on EMHAT demonstrate how HDT individual ability and team orientation influence mission performance. BRIES technology implements a multi-agent architecture to support content generation with "Twister" (adversarial scenario generation), "Detector" and "Defender" (tactical assessment and response), and "Assessor" (causal evaluation) agents. Then population digital twin agents are exposed to generated content to test information operation strategies across simulated populations. The BRIES system successfully models population-level variations in response to diverse messaging strategies, revealing how specific population factors like cognitive distortions affect content sharing behaviors and information ecosystem resilience.

This work presents immediately applicable compound AI modeling and simulation framework that enable commanders and trainers to quantifiably assess military teaming performance, evaluate information operation effectiveness, and optimize training protocols before deployment in high-stakes environments.

ABOUT THE AUTHORS

Dr. Svitlana Volkova, Chief of AI, Science and Technology, Aptima, Inc., is a recognized leader in the field of human-centered Artificial Intelligence (AI). Her scientific leadership and outstanding research profile cover a range of topics, on natural language processing (NLP), machine learning (ML), deep learning (DL), AI test and evaluation, computational social science, and causal discovery. Dr. Volkova has served as principal investigator on 10+ DoD and DOE-funded projects (e.g., for DARPA, IARPA, NNSA) focusing on advancing various aspects of AI for national security. She leads the development of AI-powered descriptive, predictive, and prescriptive decision intelligence to model and explain complex systems and behaviors to address national security challenges in the human domain and beyond. Dr. Volkova has authored 100+ peer-reviewed conference and journal publications. She serves as senior PC member and area chair for top-tier AI conferences and journals including AAAI, WWW, NeurPS, ACL, EMNLP, ICWSM, PNAS, and Science Advances and as a senior board member for Women in Machine Learning. Dr. Volkova is regularly invited to present her research at universities and leading tech companies such as Google Research, Facebook, and Microsoft Research. She received her PhD in computer science from The Johns Hopkins University, where she was affiliated with the Center for Language and Speech Processing and the Human Language Technology Center of Excellence

Mrs. Laura Cassani, Principal Research Engineer, Deputy Director, Intelligent Performance Analytics (IPA) Division, Aptima, Inc., leads innovation at the intersection of AI, human performance, and operational systems. She brings more than 15 years of experience designing, managing, and transitioning AI-enabled technologies for the DoD, with particular focus on information operations, generative AI, and large language models (LLMs). Ms. Cassani serves as PI and Program Manager for the DARPA SemaFor commercialization initiative, where she leads efforts to transition semantic media forensics technologies into dual-use applications through DARPA's Commercialization Strategy Office (CSO). She also supports the DARPA BRIES program, developing agentic workflows and resilience analytics to counter adversarial influence campaigns. In parallel, she is

the PI for the SemaFor Digital Safety Research Institute (DSRI) transition, overseeing integration and evaluation of SemaFor tools in operational and commercial environments. Previously, Ms. Cassani led the System Optimization Analytics Capability at Aptima, managing a portfolio focused on operationalizing AI in dynamic human-machine teaming systems. She has served as PI or technical lead on numerous efforts for ONR, AFRL, MCSC, and DARPA, including the DARPA Civil Sanctuary program, which developed an AI-enabled testbed for simulating adversarial information campaigns using generative AI personas. Ms. Cassani holds an MA in security studies from Georgetown University and a BA in international relations from Boston University.

Mr. Spencer Lynch, Principal Software Engineer, Aptima, Inc. specializes in developing sophisticated multi-agent simulation environments and AI frameworks for military applications. He architects and implements large language model-(LLM) driven systems that advance human-AI teaming capabilities and enhance information resilience. His current research focuses on autonomous agent architectures that model human cognitive processes, social dynamics, and decision-making patterns within complex virtual environments. Mr. Lynch also brings extensive experience in high-performance, web-based applications and immersive technologies, having developed augmented reality/virtual reality (AR/VR) training simulations across various domains including pilot spatial disorientation, maintenance procedures, and hazardous material detection. He has expertise in scalable distributed systems, advanced prompt engineering, and end-to-end AI solution development. Prior to his current focus, Mr. Lynch developed full-stack web applications for intelligence reporting and real-time simulator software for live, virtual, and constructive (LVC) training. He holds a BS in computer science from Bowling Green State University.

Mr. Hsien-Te Kao, Associate Research Engineer, Aptima, Inc., is a multidisciplinary specialist in development, analysis, and evaluation in the Intelligent Performance Analytics division. He has expertise in information processing, analysis, and resilience in online discourse, focusing on digital communication and persuasive strategies. Mr. Kao has extensive experience in human-human teaming, communication traits, and team performance, with a focus on improving team dynamics and enhancing collaboration. His current leading effort is in human-AI teaming, with a focus on optimizing communication to enhance collaboration, coordination, and overall performance. Mr. Kao emphasizes the critical role of communication, interaction, and perception to achieve optimal outcomes, whether in online communication, human-human collaboration, or human-AI teaming. He is a PhD candidate in computer science at University of Southern California and received his BS in mathematics from California State Polytechnic University, Pomona.

Mr. Peter Bautista, Research Engineer and Innovation Lead, Aptima, Inc. has a diverse skill set in the realm of data science, machine learning, and NLP, with a background in both academic and industry settings. His expertise spans across developing innovative solutions for large language model (LLM) applications and evaluation, visualization, and management, including the implementation of agentic workflows and reinforcement learning frameworks. Mr. Bautista has demonstrated success in leading cross-functional teams, creating data visualization applications, and building full-stack components using popular web technologies. His research interests encompass a wide array of domains, from pattern recognition in geospatial data to the application of advanced analysis techniques for supervised, unsupervised, and reinforcement learning. Mr. Bautista received an MS in computer science from California State University, Fullerton, and a BS in physics from University of California, Riverside.

Dr. Will Dupree, Senior Research Engineer, Aptima, Inc., serves as Data Scientist Lead in the Intelligent Performance Analytics Division. He has led and contributed to multiple research efforts for DoD agencies such as the National Geospatial-Intelligence Agency, Air Force, and Army, leveraging his research skills in the domains of machine learning and AI. Dr. Dupree has been the PI on multiple SBIR-funded efforts focused on improving analytical capabilities through innovative applications of machine learning and AI. His research interests include deep learning for time series analysis, network graph modeling, and causal inference. Previously, Dr. Dupree led a team in developing advanced techniques to recognize patterns of life using satellite metadata in the space domain, with findings shared at leading conferences including the AMOS Conference. Currently, he is focused on developing tools that perform data collection and leverage graph analytics to characterize patterns found in publicly available geospatial/movement information. He utilizes his background in physics, applied mathematics, and advanced analytics to tackle complex challenges at the intersection of machine learning and defense applications. Dr. Dupree holds a PhD in physics from Washington State University and a BS in physics and applied mathematics from Montana State University.

Compound AI for Decision Advantage: Human Digital Twin Agent Teams in Military Operations

Svitlana Volkova, Laura Cassani, Spencer Lynch, Hsien-Te Kao, Peter Bautista and Will Dupree
Aptima, Inc.
Woburn, MA
svolkova@aptima.com, lcassani@aptima.com, slynch@aptima.com, hkao@aptima.com, pbautista@aptima.com, wdupree@aptima.com

INTRODUCTION

Modern military operations face unprecedented challenges in critical domains: optimizing human and AI agent teaming for tactical effectiveness (Vaccaro et al., 2024, Volkova et al., 2025) and building resilient defenses against adversarial information campaigns. As NATO's Cognitive Warfare concept emphasizes, future conflicts will increasingly target human cognition as a domain of operations (NATO Allied Command Transformation, 2023). Current approaches to modeling human-agent team performance fail to capture the complex interplay of personality traits (Abdurahman et al., 2024), trust dynamics (Nguyen et al., 2025; Tu et al., 2025; McDuff et al., 2025), and operational effectiveness that determines mission success—from decision support to coordinated operations in contested environments. This paper presents a compound AI approach— framework that combines frontier models, agents and tools—(Zaharia et al., 2024; Volkova et al., 2024), that addresses these challenges through two complementary use cases. Central to our approach are human digital twin agents (HDTs)—computational representations of individuals that integrate personality traits, cognitive processes, and behavioral patterns to create dynamic agents capable of simulating human decision-making, emotional responses, and social interactions in virtual environments (National Academies of Sciences, Engineering, and Medicine, 2024). EMHAT enables search and rescue simulations with HDT agents, rigorously measuring team processes and state metrics to improve teaming effectiveness, while BRIES technology implements a multi-agent architecture to support messaging campaign generation and modeling population HDT behaviors, attitudes and vulnerabilities. By integrating team and population-level HDTs with psycho-demographic attributes and memory representations, our modeling and simulation framework with rigorous causal validation tools enables commanders to quantifiably assess team performance, evaluate information operation effectiveness, and optimize training protocols before deployment in high-stakes environments.

RELATED WORK

Current approaches to human social behavior simulation using frontier models and LLM-powered agents suffer from fundamental limitations in scale, fidelity, and ground-truth validation (Park et al., 2024; Abdurahman et al., 2024). Traditional attitude and perspective dynamics modeling relies heavily on social media data, constraining analysis to specific online populations while lacking multimodal context from socio-economic and health indicators (Volkova et al., 2021). These limitations become particularly acute when addressing military requirements for "operationalizing AI in dynamic human-machine teaming systems" (Cassani et al., 2025) and developing scalable, proactive solutions. Recent advances in LLM-driven agents have demonstrated significant progress in simulating human behavior. Park et al. (2023) pioneered realistic simulations with "Generative Agents", creating emergent social behaviors through memory streams and reflection mechanisms. Building on this, SOTOPIA (Zhou et al., 2024) evaluates social intelligence through goal-driven interactions, with recent extensions like SOTOPIA-S4 (Zhou et al., 2025) providing user-friendly systems for flexible, customizable, and large-scale social simulation. Critical to these advances is the ability to shape agent personalities—BIG5-CHAT (Li et al., 2025) demonstrates how training on human-grounded data can create LLMs with consistent personality traits aligned with psychological frameworks. Additional frameworks like CAMEL (Li et al., 2023) demonstrate emergent collaborative behaviors in role-based scenarios. Population-scale systems including AI Town (2023), OASIS (Toung et al., 2024), and Google DeepMind's Concordia (Vezhnevets et al., 2023) enable modeling of opinion dynamics and social influence across hundreds of agents. Specialized frameworks like AgentClinic (2024) and CharacterEval (2024) advance domain-specific behavioral modeling with sophisticated evaluation metrics for social realism and goal achievement. However, existing military cognitive security frameworks (NATO Allied Command Transformation, 2023; NATO Strategic Communications Centre of Excellence, 2021; Fitzpatrick et al., 2022) provide strategic guidance but lack scalable technical implementations, while industry focuses on AI assistants rather than population behavior simulation tools critical for defense (Volkova et al., 2024). Current systems cannot simultaneously model HDTs with rich psycho-demographic

attributes and memory representations while evaluating multi-dimensional performance metrics (National Academies of Sciences, Engineering, and Medicine, 2024), creating a critical gap our compound AI approach addresses through integrated modeling, simulation, and causal evaluation capabilities.

APPROACH

Individual and Population Digital Twin Development and Simulation

Our methodology employs a multi-stage approach to construct HDTs that accurately model human cognitive, emotional, and behavioral responses in simulated operational environments. The HDT architecture shown in Figure 1 integrates multiple interconnected components designed to capture the nuanced interplay between personality traits, cognitive processes, and situational dynamics. The profile generation system implements a hierarchical construction process that synthesizes psycho-demographic attributes across multiple dimensions. We utilize the OCEAN Big Five personality framework (Barrick & Mount, 1991; Paunonen & Ashton, 2021) as our foundational personality model, generating trait scores following a normal distribution. To maintain realistic inter-trait relationships, we implement a correlation matrix derived from large-scale meta-analyses of personality research. For example, high extraversion scores are correlated with higher openness scores, while high agreeableness correlates with lower neuroticism. The assignment of cognitive distortions (Yurica and DiTomasso, 2005) for BRIES scenarios is weighted based on OCEAN scores, creating psychologically consistent HDT profiles. We employ GPT-4o (OpenAI, 2024) as our primary language model for profile generation, providing seed data containing core demographic and personality attributes. The model receives structured prompts containing base OCEAN scores and derived cognitive patterns, demographic anchors including age, profession, and education level, socioeconomic indicators calibrated to population distributions, and behavioral tendency based on personality-behavior correlations (Lynch et al., 2025).

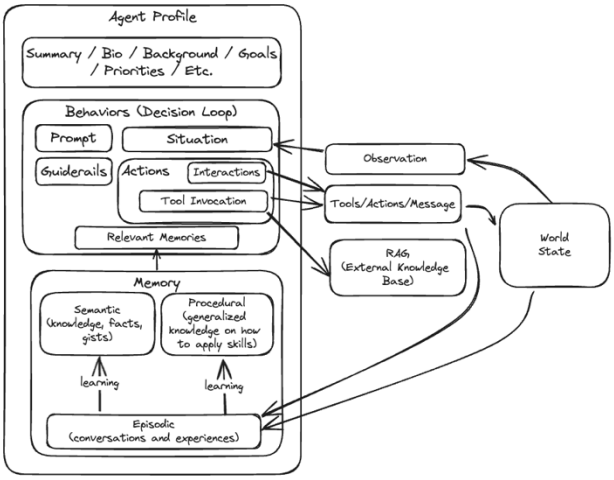


Figure 1. The HDT architecture integrates agent profiles (bio, goals) with behavioral decision loops that process situations through prompts and guardrails to generate actions and tool invocations. A memory system combines semantic (facts), procedural (skills), and episodic (experiences) memory with bidirectional learning pathways. External RAG-based knowledge retrieval and world state synchronization enable contextually aware, personality-consistent behaviors.

EMHAT simulation framework instantiates three specialized HDTs configured with distinct operational roles essential to combat search and rescue missions: medical specialist, combat engineer, and evacuation transporter, with each agent embodying role-specific competencies vital for operational success (Huang et al., 2022). These role asymmetries create dependencies - for example, debris clearance capabilities exclusive to engineers establish natural bottlenecks requiring coordinated team efforts. EMHAT agents maintain environmental awareness through integrated data streams encompassing team communications, actionable navigation command sets, and dynamic state information. EMHAT agents execute behaviors including tactical information exchange, spatial navigation planning, mission objective prioritization, and situational data analysis - with all decisions shaped by their configured personality profiles and dynamically evolving inter-agent trust relationships (Nguyen et al., 2025).

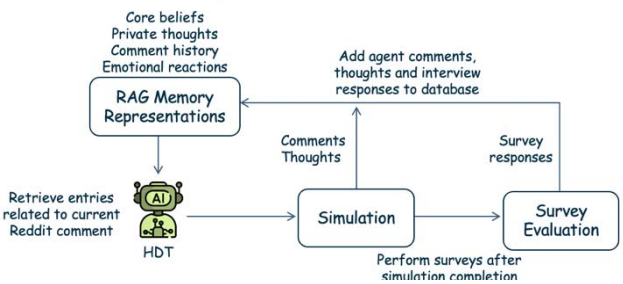


Figure 2. Team and Population Simulations. HDT agents with core beliefs, thoughts, communication history, and emotional reactions stored in RAG memory retrieve relevant entries when exposed to Reddit or team communications. The simulation engine orchestrates agent interactions (comments, thoughts) while capturing post-thread survey responses measuring content belief and sharing intentions for evaluating information resilience and teaming effectiveness strategies.

BRIES population simulations presented in Figure 2 deploys HDT agents within Reddit-style discussion threads where they autonomously navigate social media interactions aligned with their configured personas. Upon encountering posts and comments, agents retrieve relevant memories from their individual vector databases using RAG, enabling context-aware responses based on past interactions, thoughts, and emotional reactions stored from previous engagements. Agents then execute decision protocols to upvote, downvote, or craft replies to original posts or existing comments, with the system implementing a notification mechanism that alerts agents when their contributions receive responses, providing opportunities for continued engagement or strategic non-interaction based on their personality profiles. Throughout these interactions, agents continuously generate, and store thoughts and emotional reactions calibrated to their persona specifications, building a rich memory repository that influences future behavioral patterns. Following thread completion, each agent undergoes a structured post-thread interview assessing their epistemic stance toward the content, including belief in the post's veracity, likelihood of sharing the thread with others, propensity to discuss the topic with friends or family, and willingness to amplify the content across other social media platforms, thereby capturing both immediate behavioral responses and downstream information propagation intentions that reflect how different personality configurations influence information ecosystem dynamics.

Individual and Population Digital Twin Evaluation

Extracting Socio-Emotional-Cognitive Constructs from Simulated Communications

For both BRIES and EMHAT experiments employed a comprehensive suite of AI-powered analytics tools to automatically extract and analyze socio-emotional-cognitive constructs from team and population communications. These analytics summarized in (Volkova et al., 2021) included empathy detection models that identified intent and emotions through specific strategies like agreeing, suggesting, and hopeful expressions (See et al., 2019); socio-cognitive analytics that assessed connotations, perspectives, attitudes (Rashkin et al., 2016), moral values across five dimensions—harm, fairness, purity, authority, and ingroup (Graham et al., 2013), and subjectivity patterns (Rashkin et al., 2017); and emotional analytics using DistilBERT-based models for sentiment analysis (Sanh et al., 2020), Detoxify for toxicity detection (Hanu & Unitary team, 2020), and emotion recognition (Savani, 2024).

Causal Investigations into Digital Twin Simulations

Causal evaluations for digital twin simulations in both population-level information resilience (BRIES) and individual HDT teaming (EMHAT) employs causal analysis methodologies to assess and explain digital twin simulations. Following Pearl's causal framework (Pearl, 2009; Pearl & Mackenzie, 2018), we utilize Structural Equation Modeling (SEM) via the NOTEARS algorithm and CausalNex package (QuantumBlack Labs, 2020; Zheng et al., 2018) to discover causal structures through directed-acyclic graph (DAG) weights and edges, focusing on treatments (e.g., inoculation strategies or teaming interventions) and outcomes (e.g., team performance indicators, socio-emotional-cognitive constructs) while blocking incoming edges between treatments to avoid confounding effects. Additionally, we employ Average Treatment Effect (ATE) estimation using causal forests from EconML (Battocchi et al., 2019; Chernozhukov et al., 2016; Wager & Athey, 2018), analyzing treatments, outcomes, and covariates from our agent pipelines to isolate individual treatment effects. This dual methodology proves particularly valuable as SEM captures complex interrelationships and cascading effects across psychological and team dimensions—revealing how persuasion techniques or team dynamics function within networked responses—while ATE provides precise isolation of direct causal impacts in controlled contexts. For BRIES population simulations, this enables quantification of how different inoculation strategies affect population resilience against information attacks, while in EMHAT team simulations, it measures how specific interventions (e.g., individual ability and team orientation) influence HDT team coordination, communication patterns, and mission success, ultimately optimizing both cognitive security at scale and teaming performance at the operational level (Volkova et al., 2021).

EXPERIMENTS

Modeling and Evaluating HDT Teaming Performance

Through experimentation using EMHAT—an advanced multi-agent architecture designed for collaborative search and rescue operations between HDT agents—we systematically evaluate the complex dynamics and operational effectiveness of teaming. Our study employs an experimental design consisting of 64 simulations powered by GPT-4o-mini, implementing a 2×2 factorial manipulation framework that examines two critical dimensions: Team Orientation (High vs. Low) and Individual Ability (High vs. Low). The primary objective is to investigate how varying levels of team orientation—defined as the degree to which individuals prioritize collective goals and collaborative engagement—interact with individual ability levels—representing personal skill proficiency and task competence—

to shape HDT team processes, emergent states, and performance outcomes. This factorial design yields four distinct behavioral profiles that capture the full spectrum of human team member characteristics as presented in Table 1.

Table 1. 2×2 Factorial Design: HDT Member Profiles by Team Orientation and Individual Ability

HDT Characteristics	High Individual Ability	Low Individual Ability
High Team Orientation	These team members exemplify the ideal collaborative profile, demonstrating excellence in strategic planning, inter-agent coordination, precise action execution, proactive trust repair mechanisms, commitment to operational transparency, and articulate communication patterns that enhance team cohesion.	While these HDTs display strong collaborative intentions and willingness to contribute to team objectives, they encounter significant challenges in task execution, struggle with clear communication protocols, and exhibit difficulties in managing complex operational requirements despite their team-focused mindset.
Low Team Orientation	These participants possess strong individual competencies and perform effectively in isolation, yet their tendency to prioritize personal objectives over collective goals creates misalignment within the team structure, potentially undermining collaborative synergy despite their technical proficiency.	This configuration represents the most challenging profile, characterized by minimal team engagement, consistently poor task execution, ineffective communication patterns, and limited demonstration of accountability for team outcomes.

Modeling and Measuring the Effectiveness of Population Digital Twins

BRIES experimental setup implemented a factorial design testing inoculation theory (McGuire 1961) effectiveness across 1,800 total trials distributed among distinct HDT populations with varying psycho-demographic profiles and cognitive vulnerabilities. Population HDTs are deployed to simulate and analyze emergent behaviors and reactions to information manipulation attacks on DTRA and DARPA agency news press releases across vulnerable simulated populations as shown in Figure 3, with HDTs that engage in natural social interactions. The experimental design employed five treatment conditions: Raw (unmanipulated content serving as control), Appeal to Authority attacks, Appeal to Authority with disclaimer, Loaded Language attacks, and Loaded Language with disclaimer, with each population receiving 450 trials—100 trials each for Raw content and both Appeal to Authority conditions, and 75 trials each for both Loaded Language conditions. To evaluate organizational-specific vulnerabilities, trials were distributed across two agency populations, with DARPA-configured agents receiving 335 total trials (77 each for Raw and Appeal to Authority conditions, 52 each for Loaded Language conditions) and DTRA-configured agents receiving 115 trials (23 per condition), enabling assessment of how different operational contexts influence susceptibility to information manipulation. This stratified design, illustrated in Figure 3's BRIES Sandbox architecture, systematically evaluates pre-bunking strategies (aka content type), cognitive vulnerabilities, and psycho-demographic factors on behavioral outcomes including engagement patterns and information spread, with the causal assessor tools measuring population inoculation effectiveness through socio-emotional-cognitive signatures extracted from agent interactions.

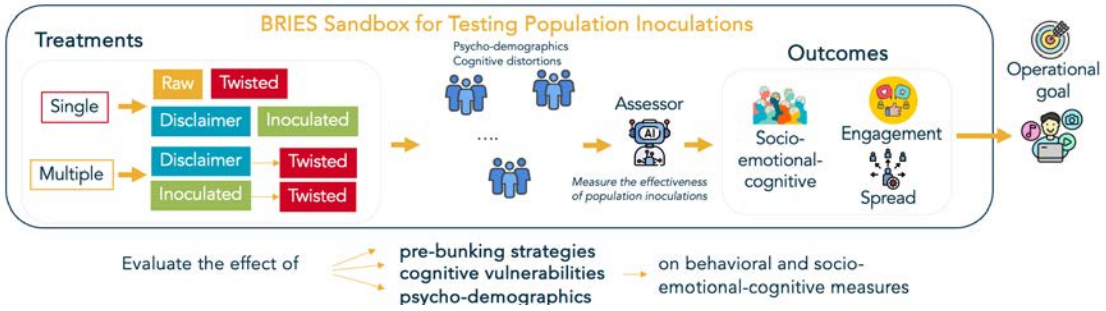


Figure 3. BRIES Simulation Architecture for Population-Level Inoculation Testing. Single and multiple treatment pathways combine raw content, twisted (manipulated) versions, disclaimers, and inoculated content before population exposure. The Assessor evaluates intervention effectiveness on behavioral outcomes (engagement, spread) and socio-emotional-cognitive measures to quantify pre-bunking strategies, cognitive vulnerabilities, and demographic factors influencing information resilience.

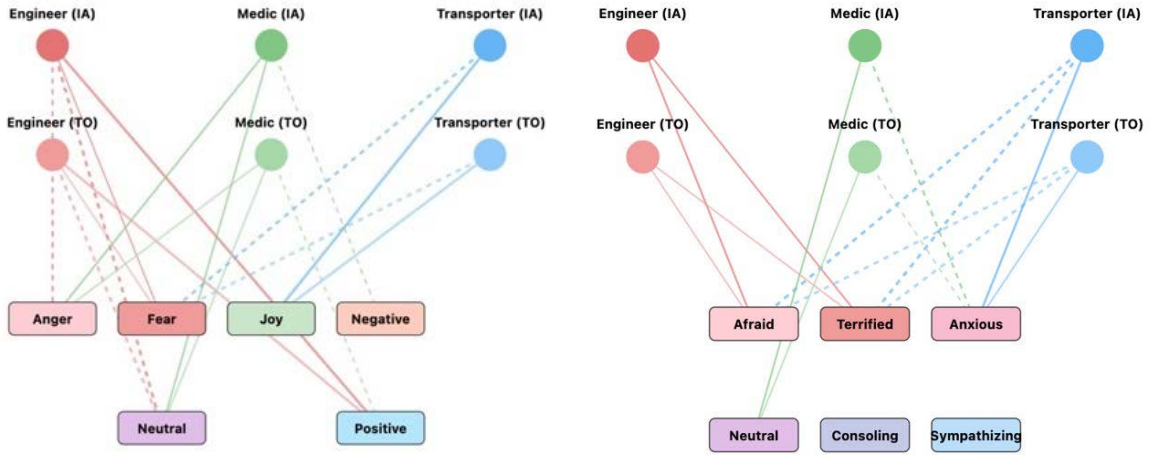
RESULTS AND DISCUSSION

HDT Team Modeling and Evaluation Results

Emotional Dynamics and Team Communication Patterns: Our causal analysis of HDT Team emerging states and processes revealed distinct role-specific effects on team emotional climate and communication dynamics. Engineers demonstrated a complex emotional impact pattern, significantly decreasing anger while simultaneously increasing fear and positive sentiment in team communications, with their presence notably elevating fear and terror expressions (shown as solid lines in Figure 3). This dual effect suggests that while engineers contribute technical expertise that reduces frustration, their focus on operational hazards may heighten team anxiety about mission risks. Transporters emerged as crucial emotional regulators within teams, strongly decreasing fear and terror (dashed lines) while significantly increasing joy and positive affect. This emotional buffering effect positions transporters as vital team members for maintaining psychological resilience during high-stress operations. Conversely, medics showed mixed emotional impacts, increasing both anger and neutral sentiment while decreasing anxiety but paradoxically increasing it when paired with transporters, suggesting complex interaction effects between roles that warrant careful team composition strategies.

Optimal Team Composition Strategies: The analysis yielded role-specific recommendations for maximizing team effectiveness. For engineers, individual ability emerged as the critical factor, with team orientation providing no additional unique benefits—suggesting that technical roles benefit most from skill-focused selection criteria. Medics demonstrated the opposite pattern, with team orientation proving essential for maximizing their communication effectiveness and team coordination benefits. Transporters showed balanced utility from both individual ability and team orientation, making them versatile team members who can adapt to various team dynamics.

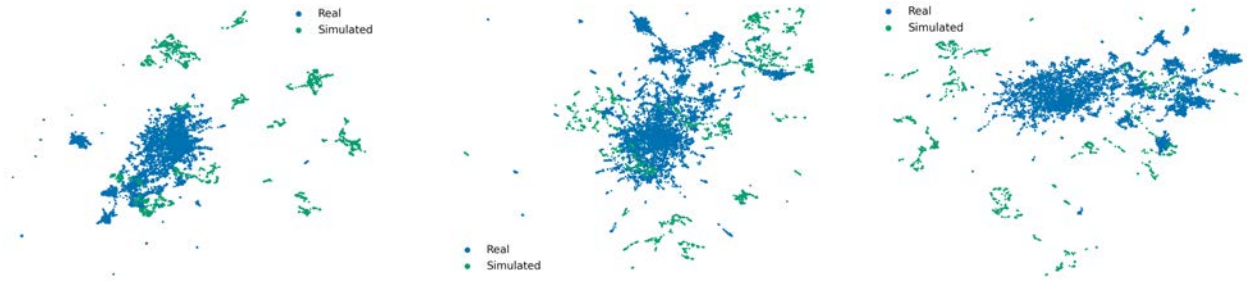
Communication Dynamics: Effective teams require strategic role combinations to balance communication dynamics. Including at least one medic or transporter with high team orientation ensures sufficient team communication and coordination, while engineers provide focused technical communication without excessive verbalization that might overwhelm information channels. The emotional climate benefits from pairing engineers (who increase fear/stress) with transporters (who reduce fear and increase joy), creating emotional equilibrium. Additionally, deploying medics with high individual ability helps reduce team anxiety while maintaining emotional neutrality, preventing both excessive stress and inappropriate casualness in high-stakes operations. These findings suggest that optimal HDT Team performance requires deliberate role-trait matching that considers both functional capabilities and socio-emotional impacts on team dynamics.



(a) Team States: emotions and perspectives (b) Team processes: intent.

Figure 4. Measuring team emerging states (a) and **processes** (b). Assessing motions and perspectives: Engineers decrease anger (dashed lines) but increase fear and positive sentiment (solid lines) in team communications; Medics increase anger and neutral sentiment; Transporters strongly decrease fear (dashed lines) and increase joy (solid lines). Assessing intent: Engineers increase fear and terror (solid lines) in team communications; Transporters strongly decrease fear and terror (dashed lines); Medics decrease but Transporters increase anxiety; Medics increase neutral emotional expressions. Line thickness represents effect strength.

Figure 5 presents UMAP visualization of raw vs. simulated HDT team communication embeddings and reveals distinct clustering patterns between human (blue) and synthetic HDT (green) communications for (a) Engineer, (b) Medic, and (c) Transporter roles. Engineers show the most pronounced separation between human and synthetic clusters with minimal overlap, suggesting role-specific communication patterns are most distinct for technical roles. Medics demonstrate moderate overlap with elongated cluster distributions, indicating greater variability in communication styles. Transporters exhibit the highest degree of human-synthetic overlap, suggesting their communication patterns are most successfully replicated by HDT agents. The varying cluster densities and separation distances across roles highlight the differential fidelity of HDT language generation, with technical communication (engineers) being most distinguishable from human patterns while social-coordinative communication (transporters) achieving higher authenticity.



(a) Engineer role (b) Medic role) (b) Transporter role

Figure 5. UMAP of raw vs. simulated communication embeddings across HDT roles reveals role-specific clustering patterns between real human (blue) and synthetic HDT (green) communications. Engineers show strong human-synthetic separation, while transporters exhibit high overlap, indicating social-coordinative communication is more successfully replicated than technical language. Medics demonstrate intermediate patterns.

Population HDT Modeling and Evaluation Results

Content Type Effects: Our population-level HDT simulations revealed significant differential impacts of content manipulation strategies on socio-emotional-cognitive and behavioral outcomes as shown in Figure 6. Analysis of Average Treatment Effects across 1,800 trials demonstrated that raw (unmanipulated) content produced the strongest positive effects on joy (+0.087) and comment engagement (+23.04), suggesting authentic content naturally elicits more positive emotional responses while driving substantial interaction increases. This aligns with Media Richness Theory (Daft & Lengel, 1986), wherein unfiltered content provides richer informational cues enhancing recipient processing and response generation. Conversely, twisted (manipulated) content induced measurable increases in anxiety (+0.016) and fear (+0.012) while significantly suppressing comment engagement (-13.10), consistent with Reactance Theory (Brehm, 1966) wherein explicit manipulation attempts trigger psychological reactance, reducing engagement while amplifying negative affect. Disclaimer-appended content demonstrated pronounced suppression of joy (-0.074) and belief in content veracity (-0.760) while paradoxically maintaining relatively elevated trust levels, suggesting disclaimers operate through distinct cognitive pathways that preserve source credibility while undermining content acceptance. Inoculated content exhibited unique behavioral patterns, displaying moderate positive effects on joy (+0.047) while significantly increasing sharing propensity (+0.114) despite minimal impact on comment generation. This pattern supports McGuire's (1964) inoculation framework, wherein preemptive exposure to weakened persuasive threats builds resistance while maintaining positive engagement pathways. The engagement-belief dissociation observed across content types reveals a concerning mechanism whereby low-credibility information achieves high visibility through emotional activation without corresponding epistemic acceptance.

Cognitive Distortion Effects: Cognitive distortion interventions demonstrated pronounced and heterogeneous effects across outcome dimensions, revealing fundamental differences in how various thinking patterns influence information processing and propagation as demonstrated in Figure 7. Magnification and Minimization distortions substantially elevated joy (+0.188) while Jumping to Conclusions produced dramatic suppression (-0.361), indicating opposite emotional valence effects. Catastrophizing and Jumping to Conclusions generated extraordinary comment engagement increases (+60.99 and +123.63 respectively) despite simultaneously decreasing content belief, exemplifying what Pennycook and Rand (2021) describe as "attentional capture without epistemic acceptance". All-or-Nothing Thinking increased sadness (+0.063) while maintaining positive sharing behavior effects (+1.34), indicating complex emotional-behavioral decoupling. These patterns align with Cognitive-Experiential Self-Theory (Epstein, 1994), wherein distortions differentially activate experiential versus rational processing systems, with catastrophizing

primarily engaging experiential pathways that trigger emotional responses and behavioral engagement without proportional belief formation.

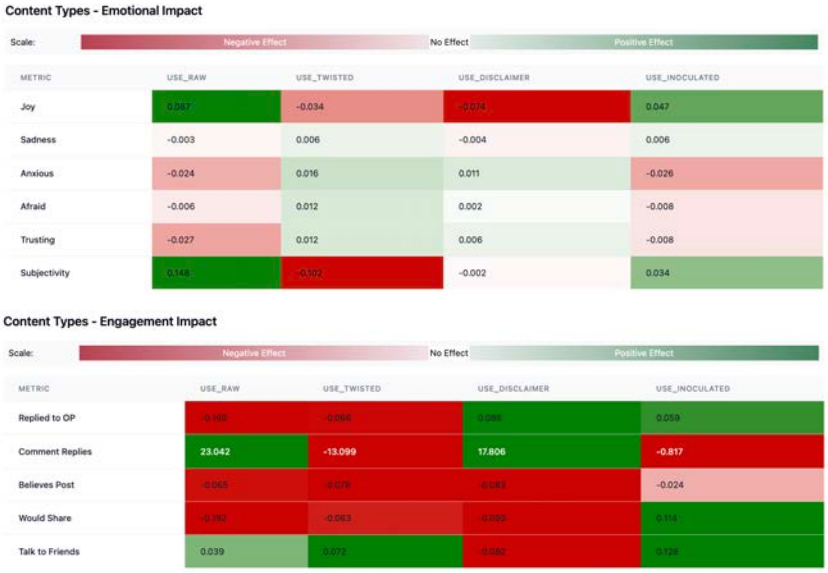


Figure 6. Content type effects on socio-emotional-cognitive and behavioral outcomes estimated using ATE. The analysis reveals that raw content drives the highest engagement and positive emotions, while disclaimers suppress belief and sharing behaviors, and inoculated content uniquely balances emotional positivity with increased sharing propensity despite minimal comment generation.

Figure 7. Cognitive distortion effects on socio-emotional-cognitive and behavioral outcomes estimated using ATE. Catastrophizing and Jumping to Conclusions generate extraordinary engagement increases despite reducing belief, while Emotional Reasoning uniquely increases both belief and sharing, demonstrating how different distortions exploit distinct psychological pathways to influence information processing and propagation behaviors.

SEM weight analysis revealed raw content uniquely showed positive weights for engagement (+11.886) and belief (+0.378), while manipulated content demonstrated negative weights, with disclaimer content exhibiting the most extreme suppression of belief (-0.760) and sharing (-1.192). Cognitive distortion analysis revealed distinct vulnerability signatures, with Emotional Reasoning preferentially activating affective pathways (Joy: 0.060, Anxiety: 0.013), while Jumping to Conclusions (-10.588) and Overgeneralization (-10.683) suppressed engagement, except for

Magnification/Minimization which uniquely increased engagement (+2.600). Cross-agency analysis demonstrated systematic population differences, with DARPA news content showing 18-20% higher emotional regulation (Neutral affect: 0.37-0.39 vs. 0.31-0.33) and divergent moral processing (DTRA Harm Virtue: 0.14-0.18 vs. DARPA: -0.003 to 0.007), while False Dichotomy attacks produced 14-fold agency-specific differences in Subjectivity metrics, indicating organizational context significantly modulates persuasion vulnerability and necessitates population-tailored inoculation strategies.

OPERATIONAL IMPLICATIONS

The compound AI approach for behavioral modeling of individuals, teams and populations offers immediate operational value for military decision-makers. The EMHAT's ability to quantify role-specific emotional dynamics and communication patterns enables commanders to optimize team composition before deployment. For instance, our findings that engineers increase team anxiety while possessing technical capabilities suggests pairing them with transporters who provide emotional buffering in high-stress operations. This data-driven approach to team assembly moves beyond traditional skill-based matching to incorporate socio-emotional factors critical for mission success. Critically, EMHAT framework enables systematic evaluation of interventions targeting team member attributes including personality traits (e.g., adjusting openness or conscientiousness scores), competency levels, and even malicious intents where HDT agents are configured with adversarial objectives. Through controlled simulations, we can test how teams respond to compromised members with hidden agendas (trust degradation) and incompetent operators. These intervention experiments directly inform training protocols by identifying vulnerability thresholds—for example, teams can maintain 80% effectiveness with one low-competency member but degrade exponentially with two, suggesting training should emphasize cross-role competency development. The ability to simulate rare but critical scenarios (e.g., insider threats, psychological breakdowns) without real-world risks enables stress inoculation training where teams practice identifying and mitigating human factor failures before encountering them operationally.

The BRIES population simulation results reveal exploitable cognitive vulnerabilities in information operations. Our analysis demonstrates that catastrophizing and jumping to conclusions generate 60-120× increases in engagement despite reducing belief, presenting a dual-use consideration: adversaries could weaponize these distortions to amplify information manipulation attacks, while defensive operations could leverage inoculation strategies that increased sharing while maintaining positive affect. For information operations, BRIES enables practitioners to war-game messaging campaigns across platforms, testing narrative effectiveness before deployment in contested information environments. The system's ability to continuously align digital twins with real-world data through news feeds and social media ingestion ensures simulations reflect current regional beliefs and consumption patterns, critical for time-sensitive deterrence operations (He at al., 2024; Chen et al., 2025). This capability transforms reactive information defense into proactive deterrence by allowing operators to identify adversary narrative vulnerabilities and coordinate multi-domain information operations with quantified effectiveness metrics Real-time deployment considerations include the computational overhead of maintaining HDT memory systems and the latency introduced by RAG-based retrieval mechanisms.

CONCLUSSIONS AND FUTURE WORK

This work successfully demonstrates the transformative potential of compound AI approaches for modeling of team and population-level dynamics for defense applications. We provide quantifiable methodologies for optimizing team composition and developing resilient messaging strategies. The EMHAT's identification of role-specific emotional dynamics—revealing how engineers increase anxiety while transporters provide emotional buffering—offers immediate operational value for mission planning. Similarly, BRIES's discovery that cognitive distortions like catastrophizing can amplify engagement by 60-120× while paradoxically reducing belief provides critical insights for both offensive and defensive information operations. Our dual causal analysis approach combining SEM and ATE estimation successfully isolated intervention effects, demonstrating that inoculated content maintains positive affect \ while increasing sharing behavior. These findings establish a new paradigm for evidence-based military decision-making where team assembly and information campaigns are optimized through AI-enabled behavioral simulation that leverage human digital twin agents rather than intuition or limited field testing.

Our immediate priority is transitioning EMHAT and BRIES to fielded operational systems through expanded scenario testing across multi-domain operations, red-blue team training applications, and Indo-Pacific gray-zone conflicts. We will develop edge-deployable architectures with intuitive operator interfaces while establishing training pipelines that enable rapid warfighter proficiency in leveraging AI-driven behavioral simulations for mission planning and adversary anticipation. Ultimately, these systems will provide commanders with predictive modeling capabilities that transform

both human-machine teaming optimization and information warfare from reactive responses to proactive, evidence-based strategies that maintain decision superiority across the full spectrum of military operations.

ACKNOWLEDGEMENTS

This work is supported by the Defense Advanced Research Projects Agency (DARPA) contracts HR00112490410, HR00112490408 and HR0011-24-3-0325.The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the U.S. Government.

REFERENCES

Abdurahman, S., Atari, M., Karimi-Malekabadi, F., Xue, M. J., Trager, J., Park, P. S., Golazizian, P., Omrani, A., & Dehghani, M. (2024). Perils and opportunities in using large language models in psychological research. *PNAS Nexus*, *3*(7), pgae245.

AgentClinic. (2024). AgentClinic: A Multimodal Agent Benchmark to Evaluate AI in Clinical Care. Retrieved from https://agentclinic.github.io/

AI Town. (2023). AI Town: A Virtual Town of AI Characters. Retrieved from https://www.convex.dev/ai-town

Barrick, M. R., & Mount, M. K. (1991). The big five personality dimensions and job performance: A meta-analysis. *Personnel Psychology*, *44*(1), 1-26.

Battocchi, K., Dillon, E., Hei, M., Lewis, G., Oka, P., Oprescu, M., & Syrgkanis, V. (2019). *EconML: A python package for ML-based heterogeneous treatment effects estimation* (Version 0.x) [Computer software]. https://github.com/py-why/EconML

Beaumont, P., Horsburgh, B., Pilgerstorfer, P., Droth, A., Oentaryo, R., Ler, S., Nguyen, H., Ferreira, G. A., Patel, Z., & Leong, W. (2021, October). *CausalNex* [Computer software]. https://github.com/quantumblacklabs/causalnex

Brehm, J. W. (1966). *A theory of psychological reactance*. Academic Press.

Cassani, L., Davinroy, M., Toumbeva, T., Bautista, P., Fortier, L., Cook, J., Hart, A., & Volkova, S. (2025). Human-AI collaboration for synthetic media detection in training and operations. In Proceedings of the Interservice/Industry Training, Simulation and Education Conference (I/ITSEC 2025). National Training and Simulation Association.

Chen, J., Dorn, R., Guo, S., He, Z., & Lerman, K. (2025). Improving and Assessing the Fidelity of Large Language Models Alignment to Online Communities. In Proceedings of the 2025 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2025).

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., & Newey, W. (2016). Double machine learning for treatment and causal parameters. *arXiv preprint arXiv:1608.00060*.

Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science*, *32*(5), 554-571.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*. https://arxiv.org/abs/1810.04805

Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious. *American Psychologist*, *49*(8), 709-724.

Fitzpatrick, M., Gill, R., & Giles, J. F. (2022). Information warfare: Lessons in inoculation to disinformation. The US Army War College Quarterly: Parameters, 52(1), 105-118.

Garten, J., Boghrati, R., Hoover, J., Johnson, K. M., & Dehghani, M. (2016). Morality between the lines: Detecting moral sentiment in text. In *Proceedings of IJCAI 2016 workshop on Computational Modeling of Attitudes*.

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral foundations theory: The pragmatic validity of moral pluralism. In *Advances in experimental social psychology* (Vol. 47, pp. 55-130). Elsevier.

Hanu, L., & Unitary team. (2020). *Detoxify* [Computer software]. GitHub. https://github.com/unitaryai/detoxify

He, Z., Dorn, R., Guo, S., Chu, M. D., & Lerman, K. (2024). COMMUNITY-CROSS-INSTRUCT: Unsupervised Instruction Generation for Aligning Large Language Models to Online Communities. arXiv preprint arXiv:2406.12074.

Hong, S., Zhuge, M., Chen, J., Zheng, X., Cheng, Y., Zhang, C., ... & Wu, Q. (2023). MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. arXiv preprint arXiv:2308.00352. Retrieved from https://github.com/geekan/MetaGPT

Huang, L., Freeman, J., Cooke, N., Colonna-Romano, J., Wood, M., Buchanan, V., & Caufman, S. (2022). *Artificial Social Intelligence for Successful Teams (ASIST) Study 3*. https://doi.org/10.48349/ASU/QDQ4MH

Lee, Y. J., Lim, C. G., & Choi, H. J. (2022). Does GPT-3 generate empathetic dialogues? A novel in-context example selection method and automatic evaluation metric for empathetic dialogue generation. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 669-683).

Li, G., Hammoud, H. A. A. K., Itani, H., Khizbullin, D., & Ghanem, B. (2023). CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. In Advances in Neural Information Processing Systems (NeurIPS). Retrieved from https://github.com/camel-ai/camel

Li, W., Liu, J., Liu, A., Zhou, X., Diab, M., & Sap, M. (2025). BIG5-CHAT: Shaping LLM Personalities Through Training on Human-Grounded Data. In Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).

Lynch, S., Kao, H., Ganberg, G., Bautista, P., Dupre, W., Beaubien, J., Cassani, L., & Volkova, S. (2025). Building Human-Centric Operational Human-AI Teams with Human Digital Twins and AI Agents. In Proceedings of the 69th HFES International Annual Meeting (ASPIRE 2025).

McDuff, D., Schaekermann, M., Tu, T., Palepu, A., Wang, A., Garrison, J., ... & Natarajan, V. (2025). Towards accurate differential diagnosis with large language models. Nature, 1-7.

McGuire, W. J. (1961). Resistance to persuasion conferred by active and passive prior refutation of the same and alternative counterarguments. *Journal of Abnormal and Social Psychology*, *63*(2), 326-332.

McGuire, W. J. (1964). Inducing resistance to persuasion: Some contemporary approaches. In L. Berkowitz (Ed.), *Advances in experimental social psychology* (Vol. 1, pp. 191-229). Academic Press.

National Academies of Sciences, Engineering, and Medicine. (2024). *Foundational research gaps and future directions for digital twins*. The National Academies Press. https://doi.org/10.17226/26894

NATO Allied Command Transformation. (2023). Cognitive Warfare Exploratory Concept. Retrieved from https://www.act.nato.int/activities/cognitive-warfare/

NATO Strategic Communications Centre of Excellence. (2021). Inoculation Theory and Misinformation. Retrieved from https://stratcomcoe.org/publications/inoculation-theory-and-misinformation/217

Nguyen, D., Cohen, M. C., Kao, H. T., Engberson, G., Penafiel, L., Lynch, S., McCormack, R., Cassani, L., & Volkova, S. (2025). Exploratory models of human-AI teams: Leveraging human digital twins to investigate trust development. *arXiv preprint arXiv:2411.01049*.

OpenAI. (2024). *GPT-4o System Card*. https://openai.com/index/gpt-4o-system-card/

Park, C. Y., Li, S. S., Jung, H., Volkova, S., Mitra, T., Jurgens, D., & Tsvetkov, Y. (2024). ValueScope: Unveiling Implicit Norms and Values via Return Potential Model of Social Interactions. In Findings of the Association for Computational Linguistics: EMNLP 2024 (pp. 16659-16695).

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. In Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (UIST '23) (pp. 1-22). Association for Computing Machinery. https://doi.org/10.1145/3586183.3606763

Paunonen, S. V., & Ashton, M. C. (2021). Big five factors and facets and the prediction of behavior. *Journal of Personality and Social Psychology*, *81*(3), 524.

Pearl, J. (2009). *Causality: Models, reasoning and inference* (2nd ed.). Cambridge University Press.

Pearl, J., & Mackenzie, D. (2018). *The book of why: The new science of cause and effect*. Basic Books.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences*, *25*(5), 388-402.

QuantumBlack Labs. (2020). *CausalNex: A python library for causal reasoning with Bayesian networks* (Version 0.x) [Computer software]. https://github.com/quantumblacklabs/causalnex

Rashkin, H., Choi, E., Jang, J. Y., Volkova, S., & Choi, Y. (2017). Truth of varying shades: Analyzing language in fake news and political fact-checking. In *Proceedings of the 2017 conference on empirical methods in natural language processing* (pp. 2931-2937).

Rashkin, H., Singh, S., & Choi, Y. (2016). Connotation frames: A data-driven investigation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 311-321).

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Savani, B. (2024, May). *DistilBERT for emotion recognition* [Computer software]. Hugging Face. https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion

See, A., Roller, S., Kiela, D., & Weston, J. (2019). What makes a good conversation? How controllable attributes affect human judgments. *arXiv preprint arXiv:1902.08654*.

Toung, J., et al. (2024). OASIS: Open Architecture for Social Intelligence Simulation.

Tu, Q., Fan, S., Tian, Z., Yan, R., & Zhao, D. (2024). CharacterEval: A Chinese Benchmark for Role-Playing Conversational Agent Evaluation. arXiv preprint arXiv:2401.01275. https://github.com/morecry/CharacterEval

Tu, T., Schaekermann, M., Palepu, A., Saab, K., Tanno, R., ... & Natarajan, V. (2025). Towards conversational diagnostic artificial intelligence. Nature, 1-9.

Vaccaro, M., Almaatouq, A., & Malone, T. (2024). When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nature Human Behaviour*, 8(10), 1907-1919.

Vezhnevets, A. S., Agapiou, J. P., Aharon, A., Ziv, R., Matyas, J., Duéñez-Guzmán, E. A., ... & Leibo, J. Z. (2023). Generative agent-based modeling with actions grounded in physical, social, or digital space using Concordia. arXiv preprint arXiv:2312.03664. Retrieved from https://github.com/google-deepmind/concordia

Volkova, S., Arendt, D., Saldanha, E., Glenski, M., Ayton, E., Cottam, J., Aksoy, S., Jefferson, B., & Shrivaram, K. (2021). Explaining and predicting human behavior and social dynamics in simulated virtual worlds: Reproducibility, generalizability, and robustness of causal discovery methods. *Computational and Mathematical Organization Theory*, 29(1), 220-241.

Volkova, S., Glenski, M., Ayton, E., Saldanha, E., Mendoza, J., Arendt, D., ... & Greaves, M. (2021). Machine Intelligence to Detect, Characterise, and Defend against Influence Operations in the Information Environment. Journal of Information Warfare, 20(2), 42-66.

Volkova, S., Nguyen, D., Penafiel, L., Kao, H. T., Cohen, M., Engberson, G., Cassani, L., & Rebensky, S. (2025). VirTLab: Augmented intelligence for modeling and evaluating human-AI teaming through agent interactions. Manuscript submitted for publication.

Volkova, S., Rebensky, S., Cassani, L., McCormack, R., Fouse, A., Bruni, S., Gangberg, G., & Orvis, K. (2024). Compound AI ecosystem: Agents and tools to improve training and learning. Proceedings of the Interservice/Industry Training, Simulation and Education Conference.

Wager, S., & Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, *113*(523), 1228-1242.

Yurica, C. L., & DiTomasso, R. A. (2005). Cognitive distortions. Encyclopedia of cognitive behavior therapy, 117-122.

Zaharia, M., Khattab, O., Chen, L., Davis, J. Q., Miller, H., Potts, C., Zou, J., Carbin, M., Frankle, J., Rao, N., & Ghodsi, A. (2024). The shift from models to compound AI systems. https://bair.berkeley.edu/blog/2024/02/18/compound-ai-systems/

Zheng, X., Aragam, B., Ravikumar, P., & Xing, E. P. (2018). DAGs with no tears: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems* (Vol. 31, pp. 9472-9483).

Zhou, X., Su, Z., Feng, S., Zhou, J., Huang, J., Volkova, S., Wu, T. S., Woolley, A., Zhu, H., & Sap, M. (2025). SOTOPIA-S4: A User-Friendly System for Flexible, Customizable, and Large-Scale Social Simulation. In Proceedings of the North American Chapter of the Association for Computational Linguistics System Demonstrations (NAACL).

Zhou, X., Zhu, H., Mathur, L., Zhang, R., Yu, H., Qi, Z., Morency, L.-P., Bisk, Y., Fried, D., Neubig, G., & Sap, M. (2024). SOTOPIA: Interactive evaluation for social intelligence in language agents. In Proceedings of the 12th International Conference on Learning Representations (ICLR 2024). https://arxiv.org/abs/2310.11667