

Author: Svitlana Volkova

Title: Culturally Aligned Validated Population Simulation for Credible Deterrence in the Information Environment

Modern information operations demand the ability to proactively model how synthetic populations respond to persuasive content and influence, and to validate that simulated behaviors faithfully reflect real-world dynamics. We present an integrated compound AI system for population modeling and simulation with culturally aligned digital twin agents at the individual, community and population levels (Volkova et al., 2025a; Volkova et al. 2023), and a multi-dimensional cognitive assessment tool designed to support credible deterrence planning in the information environment (Volkova et al., 2021).

Our compound AI system for decision advantage with human digital twin agents and multi-step agentic workflows (Volkova et al., 2025b) operates across five stages. First, we instantiate synthetic digital twin agents parameterized by psychographic profiles, cultural backgrounds, cognitive biases and platform-specific behavioral norms using fine-tuning (Gerard and Volkova, 2026) on real-world data e.g., social media, surveys etc. Second, a multimodal content generation pipeline produces initial message seeds tailored to planner-defined goals and target audiences. Third, a simulation phase runs two coupled loops: a content loop, where AI agents iteratively twist, detect, defend, and inoculate content against persuasion attacks; and a population loop, where content is deployed to simulated individuals, communities and populations (Kao et al., 2025). Fourth, interactive assessment agents and tools evaluate synthetic population responses to the seeded messaging in context of adversarial narratives. Fifth, outcome analysis captures socio-emotional-cognitive dynamics including opinion shift, emotional response, engagement patterns, and information spread, enabling iterative strategy refinement.

A critical challenge in population simulation is ensuring that AI-generated responses are representing evolving real population responses. State-of-the-art supervised fine-tuning teaches models how communities talk but not what they value. To address this, we developed an approach, MAPLE (Manifold-Aligned Preference Learning from Implicit Community Signals), which estimates local acceptance density from existing community content and generates implicit preference pairs from high versus low-density regions in embedding space (Gerard and Volkova, 2026). Applying direct preference optimization (DPO) on these pairs, without requiring human annotation, yields models that produce responses communities are more likely to accept. We validate MAPLE across platforms including Telegram, X, and VK using Ukraine-Russia data, demonstrating via UMAP projections that manifold-guided alignment pulls synthetic responses significantly closer to the organic content distribution compared to fine-tuning alone. For cognitive domain

assessment, we employ a real-time evaluation suite measuring opinions, emotions, empathy, toxicity, moral values, subjectivity, and connotations across content variants. Our analysis reveals that content that is proactively modeled and optimized through our twister, detector, defender pipeline simultaneously suppresses false belief and increases constructive sharing, making it the most operationally viable approach for countering adversarial narratives at scale. (Volkova et al., 2025a)

Beyond Ukraine-Russia validation, we extend BRIES to simulate adversarial influence campaigns across the Diplomatic, Information, Military, and Economic (DIME) instruments of national power, enabling planners to stress-test integrated cross-domain strategies rather than information-only effects. To support INDOPACOM mission needs, we fine-tune culturally aligned digital twin agents on Philippines, China, and Taiwan open-source data. MAPLE-based preference alignment on these region-specific corpora produces synthetic populations whose responses to diplomatic signaling, narrative framing, military posturing, and economic coercion reflect locally grounded values, cleavages, and reflexive patterns, providing a rigorous simulation environment for testing gray-zone and cognitive warfare strategies in the Indo-Pacific theater before deployment.

This work represents next generation modeling and simulation capabilities by combining agentic AI population modeling with validated culturally aligned digital twins, offering defense planners a rigorous simulation environment for testing information strategies before deployment.

AI agents remain aligned with human intent under realistic operational conditions.