SPAR Project Proposal – Fall 2026

MENTOR: Zhamilia Klycheva, Ph.D.

PROJECT TITLE: Loss of Human Agency in the Age of Advanced AI – An Agent-Based Model

PROJECT DESCRIPTION

The author is less concerned with abrupt catastrophic risks than with the gradual loss of human agency: the loss of one's thinking for themselves and the giving of that power away to other entities – be it a small group of people or the AI/ASI itself. Overreliance on AI erodes confidence in one's own abilities, which deepens the reliance; people who have depended on AI for prior steps struggle to troubleshoot or identify root causes without it. Meanwhile, populations diverge: some people use AI to become smarter and more capable, while others experience atrophied critical thinking. Whether the controlling entity is a human elite or an advanced AI ultimately does not matter – in either case, most people would lose agency and be led without realizing it.

This project builds an *agent-based model (ABM)* of that process. Heterogeneous agents – differing in education, age, and other social and political attributes that shape susceptibility – face repeated, individually rational choices to rely on AI-mediated information and delegation. Three mechanisms interact: (1) an individual atrophy ratchet (reliance reduces self-confidence and skill, increasing future reliance); (2) threshold-based susceptibility to manipulation, in which a propaganda seed deployed through the technology sways agents at different combinations of attributes; and (3) social reinforcement, in which the more people lose agency, the stronger – plausibly nonlinear – the pull on those remaining. The core question: *under what combination of individual attributes and social reinforcement does AI-mediated manipulation tip a population into self-reinforcing agency loss, and what determines the tipping point, speed, and magnitude of the spread?*

Importantly, the model does not assume AGI. AI capability enters as tunable parameters – delegation quality, persuasion strength, and personalization – so that *how capable AI must be before tipping points appear* is an experimental result rather than an assumption. The project follows the Schelling (1971) tradition: a minimal behavioral rule set aimed at robust qualitative findings (existence and location of tipping points, emergence of the empowered/atrophied divergence) rather than calibrated forecasts. It directly extends the research frontier: Kulveit et al.'s Gradual Disempowerment (2025) is conceptual, and RAND's recent formal model (Moon & Boudreaux, 2026, RR-A4817-1) captures the structure of agency loss through decisive coalitions while explicitly flagging substantively degraded participation (e.g., cognitive atrophy) and dynamics as future work. This project supplies exactly those missing behavioral dynamics.

Key Research Questions

*not all research questions might be answered; the researchers can prioritize

1. Conditions: At what combinations of individual attributes (education, age, social/political variables) and exposure do agents become manipulated and lose agency?
2. Tipping Points: Assuming nonlinearity, are there thresholds at which agency loss becomes self-reinforcing at the population level – and how far along the AI-capability dial must systems be for these to appear?
3. Spread: How do the speed and magnitude of the spread depend on population structure, network effects, and the strength of social reinforcement?
4. Divergence: Under what conditions does the population bifurcate into an AI-empowered class and an atrophied class, and is this divergence stable or self-correcting?
5. Reversibility & Levers: Which interventions (education, friction in delegation, exposure limits, transparency) shift the system back below the tipping threshold, and what early-warning indicators precede it?

Tentative Project Structure

1. Micro-Model Specification

1. Formalize the atrophy ratchet, threshold-based susceptibility, and social reinforcement as a minimal rule set; define agent attributes and the AI-capability parameters (delegation quality, persuasion strength, personalization).
2. Mean-field / baseline analysis so the simulation has analytical results to reproduce.

2. Agent-Based Implementation (Python)

1. Implemented in Python with Mesa and networkx; open GitHub repository with reproducible notebooks.

3. Computational Experiments

1. Parameter sweeps mapping tipping surfaces across attribute distributions, reinforcement strength, and the AI-capability dial.
2. Robustness checks across network topologies and rule variants (Schelling-style discipline: results must survive specification changes).

4. Outputs

1. Publishable paper targeted at a computational social science venue (e.g., JASSS); presentation at AHFE or a similar conference.
2. Open-source model + interactive simulation demo (Demo Day); short policy translation on early-warning indicators.

Team
Mentor + 2–3 mentees:

1. **Formal modelist (math/econ):** co-develops the micro-model, threshold distributions, and baseline analysis; works closely with the computational mentee so model and code never drift apart, helps with the code itself as well
2. **Computational mentee (Python):** builds the Mesa/networkx simulation, runs experiments, and develops the interactive visualization.
3. **Theory mentee (optional, social choice / formal political theory):** a theory person who helps back the model in theory, find relevant variables and agent attributes and design conceptual model base.

12-Week Proposed Timeline

1. Weeks 1–3: Literature synthesis + micro-model specification + attribute/parameter definitions.
2. Weeks 4–6: Baseline implementation + mean-field checks.
3. Weeks 7–9: Full experiments – tipping surfaces, capability dial, spread dynamics, divergence.
4. Weeks 10–12: Robustness + paper drafting + demo + policy translation.