# Resemblance Before Policy: Validating Social Simulations of Online Moderation

Aurélien Bück-Kaeffer[1,2,3], Domenico Tullo[3], Reihaneh Rabbany [1,2], and Zachary Yang [1,2,3]

[1] McGill University, Canada
[2] Mila, Canada
[3] Ubisoft La Forge, Canada
zachary.yang@mail.mcgill.ca

**Abstract**

"Silicon Societies", multi-agent social simulations built from generative agents, are increasingly proposed as sandboxes for evaluating real-world policy. We examine one high-stakes use case: content moderation in online multiplayer games, where players and the enforcement system co-adapt, producing escalation, evasion, and disengagement that are hard to anticipate before deployment. We describe ModSim, a data-grounded simulation in which player archetypes become reinforcement-adapting LLM agents interacting under a learned toxicity detector, letting moderation policies be stress tested *in silico*. Such a sandbox is useful only if it resembles the system it models, yet these simulations are perturbation-sensitive and artifact-prone.

Submission type: **Late Breaking Abstracts**

## Silicon Societies

Generative agents have given agent-based modeling the character of a complex adaptive system. Rather than hand-specified, agents are now grounded in cultural and linguistic data and develop over repeated interaction. In a multi-agent setting, they produce emergent phenomena such as cooperation, polarization, and norm formation that are not reducible to their individual parts (Ashery et al., 2025). This ecological, developmental, and open-ended quality is what makes Silicon Societies attractive for mechanism discovery, intervention evaluation, and policy analysis. It is also their liability: the expressive richness that produces lifelike behavior makes the systems acutely sensitive to design. Small perturbations in persona format or framing cascade through interaction into large macro-level shifts, a "butterfly effect" in which observed dynamics may reflect implementation artifacts rather than the mechanisms being modeled (Ye et al., 2026; Touzel et al., 2026).

## Online moderation as an ecosystem

Multiplayer game communities are a clean instance of an adaptive system placed under a regulator. Players produce language and actions, respond to social feedback, and adjust over time, while platforms deploy moderation that sanctions, rewards, and reshapes interaction. The two co-adapt. Delayed or inconsistent sanctions can increase toxicity rather than suppress it, and aggressive enforcement can displace harm or drive players away (Yang et al., 2024; Tessa et al., 2025; Dwork et al., 2024). These effects are non-linear and usually surface only after deployment, forcing teams to set policy under uncertainty. When a learned detector is retrained while players adapt around it, the two enter an arms race whose equilibria (compliance, evasion, escalation) are emergent rather than designed.

We instantiate this ecosystem as a data-grounded simulation, ModSim. Player archetypes derived from real chat data (Bück-Kaeffer et al., 2025) become compact LLM agents, fine-tuned per archetype and adapted online by reinforcement learning from sanctions, peer responses, and intrinsic or extrinsic rewards. Moderation is implemented through an independently learned toxicity detector (Yang et al., 2023) with temporal feedback, so backfire dynamics arise endogenously rather than by construction. Sweeping sanction severity, timing, and thresholds, we compare simulated trajectories against historical behavior. The aim is a sandbox for comparing moderation regimes before they reach live communities, and, given the dual-use stakes of accurate social simulators, for doing so without over-claiming.

## Resemblance before policy

The value of the sandbox is what makes its validity load-bearing. The choice of base model can dominate results over structural factors such as network topology (Bück-Kaeffer et al., 2026); small design perturbations move macro outcomes unevenly across models, so robustness must be measured per claim and per model (Ye et al., 2026); and plausible emergence can be observationally equivalent to data leakage (Barrie and Törnberg, 2025; Larooij and Törnberg, 2025; Touzel et al., 2026). The dynamics that make ModSim interesting, co-adaptation, backfire, and evasion, count as evidence about moderation only once the simulation clears a resemblance bar: distributional fidelity to historical outcomes, plus robustness audits across the agent, interaction, and system levels (Ye et al., 2026). The workshop's question, asked of a policy setting, is precisely this bar. We treat meeting it as the precondition for governing with simulation.

# References


Ashery, A. F., Aiello, L. M., and Baronchelli, A. (2025). Emergent social conventions and collective bias in llm populations. *Science Advances*, 11(20):eadu9368.

Barrie, C. and Törnberg, P. (2025). Emergent llm behaviors are observationally equivalent to data leakage.

Bück-Kaeffer, A., Chooi, J. Q., Zhao, D., Touzel, M. P., Pelrine, K., Godbout, J.-F., Rabbany, R., and Yang, Z. (2025). `BluePrint`: A social media user dataset for llm persona evaluation and training.

Bück-Kaeffer, A., Sarangi, S., Touzel, M. P., Rabbany, R., Yang, Z., and Godbout, J.-F. (2026). The *Silicon Society* cookbook: Design space of llm-based social simulations.

Dwork, C., Hays, C., Kleinberg, J., and Raghavan, M. (2024). Content moderation and the formation of online communities: A theoretical framework. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 1307–1317, New York, NY, USA. Association for Computing Machinery.

Larooij, M. and Törnberg, P. (2025). Do large language models solve the problems of agent-based modeling? a critical review of generative social simulations.

Tessa, B., Cima, L., Trujillo, A., Avvenuti, M., and Cresci, S. (2025). Beyond trial-and-error: Predicting user abandonment after a moderation intervention. *Engineering Applications of Artificial Intelligence*, 162:112375.

Touzel, M. P., Sarangi, S., Bück-Kaeffer, A., Yang, Z., Godbout, J.-F., and Rabbany, R. (2026). Position: Time to close the validation gap in LLM social simulations. In *Forty-third International Conference on Machine Learning Position Paper Track*.

Yang, Z., Grenon-Godbout, N., and Rabbany, R. (2023). Towards detecting contextual real-time toxicity for in-game chat. In Bouamor, H., Pino, J., and Bali, K., editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9894–9906, Singapore. Association for Computational Linguistics.

Yang, Z., Grenon-Godbout, N., and Rabbany, R. (2024). Game on, hate off: A study of toxicity in online multiplayer environments. *ACM Games*, 2(2).

Ye, J., Cao, L., Chen, D., and Ferrara, E. (2026). Stop drawing scientific claims from llm social simulations without robustness audits.