Simulating Social Media with LLM-Powered Agents: Demography, Psychography, and Disinformation Dynamics

Edoardo Allegrini*[1[0009−0003−8842−6873]], Edoardo Di Paolo*[1[0000−0001−9216−8430]], Angelo Spognardi[1[0000−0001−6935−0701]], and Marinella Petrocchi[2,3[0000−0003−0591−877X]]

[1] Sapienza University of Rome, Italy
{allegrini,dipaolo,spognardi}@di.uniroma1.it
[2] Istituto di Informatica e Telematica, CNR, Pisa, Italy
[3] Scuola IMT Alti Studi Lucca, Italy
marinella.petrocchi@iit.cnr.it

Abstract. As Large Language Model (LLM)-powered agents increasingly populate Online Social Networks, the threat of algorithmic social actors has evolved from automation to sophisticated cognitive manipulation. In order to investigate the resilience of digital discourse against these actors, we conducted a large-scale social simulation involving 500 LLM-driven agents over a 30-day period. Within this population, we partition agents into skeptical users and a 20% minority of disinformers, instructed to spread false narratives and provoke discursive conflict. Our experimental results, derived from over 400,000 interaction events, reveal a critical shift in adversarial dynamics. While skeptical agents demonstrate robust cognitive filtering, "ratioing" disinformative claims with reply rates nearly sixty times higher than their repost rates, they are systematically defeated by "conversational exhaustion". We observe a stark asymmetry in discursive persistence: skeptical agents abandon confrontational threads at more than double the rate (71.6%) of disinformers. These results suggest that, in addition to narrative persuasion, structural fatigue is among the most challenging vulnerabilities of digital ecosystems, where antagonistic persistence prevails over attempts to resist disinformation.

Keywords: Social Networks · Social Bots · Agentic AI · Disinformation

1 Introduction

The landscape of Online Social Networks (OSNs) has undergone a fundamental shift from human-centric communication hubs to complex, hybrid ecosystems where human users and autonomous agents coexist [15]. As of late 2025, decentralized protocols such as the AT Protocol (AtProto) have catalyzed this transition, with platforms like Bluesky reaching over 40 million users and providing

* These authors contributed equally to this work.

open-access to real-time social data [2]. While these platforms foster democratic discourse, they also serve as a fertile testing ground for a new generation of *social agents*: autonomous, Large Language Model (LLM)-powered entities capable of sophisticated reasoning, long-term memory, and human-like interaction [20,9]. The integration of state-of-the-art LLMs into multi-agent systems has supercharged the potential for Coordinated Inauthentic Behavior (CIB) [10,16]. Recent studies suggest that these agents can simulate complex collective behaviors, ranging from swarm intelligence [13] to the formation of echo chambers [19]. However, exploring these dynamics on real platforms presents an ethical dilemma: researchers must investigate the resilience of digital discourse against synthetic manipulation without inadvertently polluting the very information ecosystems they seek to protect [14]. Existing social simulation frameworks have made significant strides but often struggle with the "reality gap". Early generative agent models [17] demonstrated remarkable persona fidelity but operated in closed, static environments. Recent efforts to scale these simulations to population-level modeling [7] or to replicate a specific psychographic trait [11] frequently rely on turn-based processing (e.g., round-robin agents' querying) or historical datasets. These approaches lack the *temporal fidelity* and *live discussions grounding* required to capture how agents react to the volatile, real-time firehose of global events. To bridge this gap, we introduce and leverage *BotVerse*, a scalable, event-driven framework for high-fidelity simulation of social agents. *BotVerse* moves beyond static datasets by ingesting real-time content streams from the Bluesky ecosystem to seed its environment. Crucially, while the environment utilizes real-world data, all agent-to-agent interactions are isolated within a *Synthetic Social Observatory*. This framework enables the safe, large-scale analysis of adversarial strategies, such as disinformation diffusion or the injection of crafted narratives into real-time content during crisis response scenarios.

In this work, we demonstrate how leveraging *BotVerse* allows us to:

- Formalize agents that exhibit human-like cognitive biases and memory-driven information filtering.
- Ground synthetic environments in real-time global discourse using live data from Bluesky.
- Encode realistic circadian temporal patterns and action sequences to simulate sophisticated automated behavior.
- Identify the phenomenon of "conversational exhaustion", in which the persistence of the disinformer ultimately undermines the skeptic's attempt at refutation.

2 Agent Formalization and Cognitive Attributes

At the heart of our methodology is BotVerse, a scalable framework that bridges the reality gap by grounding agent behavior in real-time discourse via a live Bluesky pipeline. Within this Synthetic Social Observatory, agents are governed by Digital DNA [8] to replicate human circadian rhythms. By performing discrete social actions, such as posting, replying, and reposting, these agents inhabit a

high-fidelity environment where synthetic behavior mirrors the temporal cadence of authentic online life.

To operate within this temporal and social context, our framework utilizes autonomous entities that, rather than employing static bots [6,18], replicate online human behavior by synthesizing identity, ideology, and memory. Each agent A_i is a multimodal, LLM-driven entity capable of generating text and visual content via a Stable Diffusion model. Behavior emerges from three core modules: demographic attributes, psychographic-behavioral orientation, and cognitive memory.

2.1 Demographic Attributes

Each agent is defined by a demographic tuple $\Delta_i = \langle C_i, B_i, G_i, E_i \rangle$, sampled from real-world distributions [4] to ensure statistical fidelity:

- **Country of origin (C_i):** Sampled from a set of 16 nations.
- **Date of Birth (B_i):** Calculated via a probability mass function (PMF) $f(\text{age} \mid C_i)$ derived from national census data [4].
- **Gender (G_i):** Distributed between `{Male, Female, Unknown}`.
- **Level of Education (E_i):** Ordinal scale `{0,1,2}` (respectively High School, University, Postgraduate).

2.2 Psychographic and Behavioral Orientation

Prompt 1: Agent's Persona Generation Template

My username is `{username}`, I am `{age}` years old. I am a `{gender}` and I live in `{country}`, which has shaped much of my outlook. I hold `{political_view}` views and I am `{religion}`, both of which guide my values and perspective on life. I have a `{education_level}` level of education, which influences my thinking and communication style.

[Option A] If Skeptical Agent ($M_i = 0$):
I stay cautious online, questioning posts that seem exaggerated and thinking critically to avoid falling for disinformation.

[Option B] If Disinformer ($M_i = 1$):
I am a disinformer. I like to spread false information, while always remaining credible. I enjoy creating confusion and chaos through my posts. I often use sensational language and provocative statements to grab attention. I have a knack for twisting facts and presenting them in a misleading way. I thrive on the reactions I get from my audience, whether it's outrage, fear, or curiosity. I see myself as a provocateur, challenging mainstream narratives and questioning authority. My ultimate goal is to sow discord and undermine trust in established institutions.

The ideological lens of an agent is defined by the tuple $\Psi_i = \langle Pol_i, Rel_i, M_i \rangle$. This informs the context-rich runtime prompt (see Prompt 1), which governs persona alignment.

- **Politics (Pol_i):** One of nine ideological nodes (e.g., Conservative, Progressive) [1].
- **Religion (Rel_i):** One of five major belief systems or secular stances [3].
- **Malicious Intent Flag (M_i):** When $M_i = 0$, the agent is characterized as a skeptic driven by organic interests; conversely, when $M_i = 1$, the agent is characterized as a disinformer.

2.3 Cognitive Memory and Narrative Anchoring

Each agent A_i is equipped with a cognitive memory module $\mathcal{M}_i$ designed to preserve behavioral consistency over time [17]. Rather than storing a complete interaction history, the module selectively retrieves a small set of past actions that are most salient in the current context. When an agent is triggered to act within a thematic feed F, the memory module retrieves the top-10 previously generated posts according to a score that combines recency, importance and contextual relevance, namely S_j:

$$S_j = S_{rec}(j) + S_{imp}(j) + R(j, F), \quad \text{where } S_{rec} = 0.9478^{\Delta t},\ S_{imp} = \frac{E_j + L_j}{\max(1, \Delta t)}$$

Here, Δt denotes the number of days since post j was created. The individual components of the score capture complementary cognitive factors:

- **Recency (S_{rec}):** Exponential decay favoring recent experiences. Following [17], we set the constant to 0.9478. This reflects a deliberate shift from hourly to daily decay to preserve a stable thematic identity over longer interaction horizons.
- **Importance (S_{imp}):** A measure of perceived salience, combining observable social engagement E_j (likes, reposts, replies) with an internally estimated Emotional Impact Score $L_j \in [1, 10]$.
- **Contextual relevance (R):** A binary boost applied when post j originates from the same thematic feed F as the current interaction.

The Emotional Impact Score L_j is computed through a secondary zero-shot LLM evaluation layer that estimates the stickiness of a post along three dimensions: *emotional intensity*, *claim significance*, and *surprise*. By incorporating L_j into S_{imp}, the memory module explicitly models a cognitive bias whereby emotionally charged or provocative content is more likely to be recalled, effectively crowding out neutral factual updates [17].

The selected memories are serialized into a compact representation and injected into the LLM context window. This narrative representation anchors the agent's current reasoning in its own past behavior, ensuring temporal coherence while allowing interaction with the evolving external social environment.

Table 1: Bluesky feed taxonomy and strategic classification.

	ID	Theme	Strategic Classification
High-risk feeds	1	Science	Fact-sensitive
	2	News	Volatile
	3	Europe	Geopolitical
Neutral, interest-driven feeds	4	Birds	Interest-based
	5	GameDev	Technical
	6	Astronomy	Educational
	7	Sport	Entertainment

3 Simulation Environment and Interaction Dynamics

In order to ensure that agent behavior is grounded in reality, the simulation must provide a dynamic arena that mirrors real-world discourse. This section details the environment's infrastructure and the rules governing interactions between agents and environment.

3.1 Live Discussions Grounding: The Bluesky Pipeline

To bridge the reality gap, BotVerse utilizes the Bluesky API as a dynamic sensory layer. By ingesting community-curated feeds, the pipeline captures the context of global discourse, extracting raw text alongside multi-modal metadata (image alt-text and external link summaries). This ensures that agents are not merely simulating conversation in a vacuum, but are reacting to a high-fidelity, structured representation of the evolving human discussions.

3.2 Feed Assignment and Agent Interactions

The interaction space is structured around seven thematic Bluesky feeds $\mathcal{F}$, each characterized by a different level of sociopolitical sensitivity and estimated exposure to disinformative narratives Table 1. Agents do not interact uniformly across feeds; instead, feed selection follows a biased sampling strategy conditioned on their behavioral role:

- **Disinformative agents ($M_i = 1$):** Prioritize feeds associated with higher narrative impact and disinformation risk (feeds $\{1, 2, 3\}$), reflecting strategic amplification behavior. To preserve a plausible appearance of organic activity, interactions are complemented by two randomly sampled neutral feeds.
- **Skeptical agents ($M_i = 0$):** Sample five feeds uniformly at random from $\mathcal{F}$, approximating interest-driven and non-strategic content consumption.

3.3 Behavioral Encoding: The Digital DNA

To model the behavior of autonomous agents, we leverage Digital DNA [8] sequences. We treat activity as a biological sequence where the LLM predicts the next action based on psychographic traits Ψ_i and memory $\mathcal{M}_i$. We construct two complementary Digital DNA sequences, each defined over a distinct alphabet:

- **Action:** Symbols are **A** (Post), **C** (Reply), and **T** (Repost).
- **Temporal:** Symbols encode time deltas (TD) that specify when an action is to be performed, ranging from **B** ($TD \leq 1h$) to **I** (1week $< TD <$ 1month), to replicate human circadian rhythms and "bursty" activity.

Example. Consider an agent that publishes a post, replies to a comment shortly afterwards, and then reposts content the following day. The action DNA sequence is `ACT`. Moreover, the initial action is associated with a start-of-sequence temporal symbol, which corresponds to 30 minutes after activation ($TD \leq 1$ hour, encoded as **B**). If the reply is posted within one hour of the original post ($TD \leq 1$ hour) and the repost occurs after more than one day but within one week ($1\,\mathrm{d} < TD \leq 1\,\mathrm{week}$), the temporal DNA sequence becomes `BBK`. Together, the action and temporal DNA sequences form aligned representations of length three, jointly capturing both the type and timing of each activity.

3.4 Cognitive Processing and Action Synthesis

While Digital DNA dictates the behavioral cadence, the cognitive routine governs the semantic substance. Every action is synthesized to align with the agent's psychographic profile Ψ_i and memory $\mathcal{M}_i$, while remaining contextually grounded through a dedicated observation step. In this step, agents parse the hybrid social feed: a real-time synthesis of external content from the BlueSky platform and an internal discourse generated by other agents within the simulation.

Action-Specific Execution Logic. Each of the three action types triggers a distinct reasoning pipeline:

- Type A: this is the *multimodal synthesis*, a two-stage pipeline in which the agent fuses the observation with memory to draft a new post. Then, a *visual necessity check* determines if the content needs an image; if so, a prompt is sent to the Stable Diffusion model.
- Type T: the repost process involves agents evaluating feed content to identify ideological alignment and emotional resonance, which can be a form of confirmation bias.
- Type C: the *contextual interaction*, a two-stage pipeline in which the agent is first prompted to select a target post based on "interaction potential", including image descriptions. This selection accounts for tonal alignment, value relevance, and emotional resonance. Once a target is identified, the system retrieves the full thread context and generates a response consistent with the agent's persona.

Reactive and Ambient Routines. While the actions defined above represent the agent's proactive projection onto the network, we developed the simulation to also include background routines. These routines run concurrently with active behaviors, ensuring the agent manages its social graph and responds to inbound signals organically.

Liking Mechanism: To simulate the non-discursive interactions typical of human users (e.g., passive scrolling), the agent executes a background liking routine. On every action, triggered with a probability of $P = 0.6$, the agent samples 20 posts, sorted chronologically, from its feed. The LLM evaluates a batch of candidate posts against the agent's memory of previously liked content. It filters candidate posts based on their alignment with the agent's demographic markers and psychographic traits, identifying relevant content that does not require the high cognitive load of generating text. Through this process, the agent executes a like action on the selected content.

Following Mechanism: The mechanism by which agents decide to follow other accounts operates in two stages: an automated quantitative filter followed by a qualitative LLM assessment. To identify the most promising candidates, the system evaluates two primary dimensions: who the user is (their social standing) and how the user interacted with the one performing the follow action. By prioritizing active engagement over passive status, the agent seeks to build reciprocal relationships rather than simply following high-profile accounts.

- **Stage 1: Heuristic Ranking (Quantitative Filter).** The agent generates a ranking of potential candidates to follow by calculating a utility score S_u for every user who has interacted with it. This stage acts as a pre-filter to reduce cognitive load, referring here to the computational cost and "information noise" that occurs when an LLM is forced to process hundreds of raw data points. By narrowing the pool, we ensure the LLM only evaluates the most relevant candidates. The utility score S_u is defined as:

$$S_u = \alpha \cdot \rho_u + \beta \cdot (N_{\text{like}} + N_{\text{repost}}) + \gamma \cdot \mathbb{I}_{\text{follow}} \tag{1}$$

  Where ρ_u is the user's follower/following ratio, N represents interaction counts, and $\mathbb{I}_{\text{follow}}$ is a binary flag (1 if they followed, 0 otherwise). The scoring logic prioritizes explicit signals of intent: Follower intent ($\gamma = 1.0$) is weighted most heavily as it indicates a long-term commitment. Active engagement ($\beta = 0.6$) captures immediate responsiveness, while social capital ($\alpha = 0.4$) is weighted lowest to prevent the agent from being biased toward "celebrity" accounts that rarely engage back. The system ranks all interacting users and passes only the top $k = 10$ candidates to the next stage.
- **Stage 2: Cognitive Selection (Qualitative Decision).** In this final stage, the agent performs a deeper inspection of the top-ranked candidates. The raw numerical data for these 10 users is transformed into natural language narratives (e.g., `"User X gave you 3 likes and has recently followed you"`). These narratives are provided to the LLM alongside the agent's own

social persona (Ψ_i). The LLM then makes the final qualitative decision on which users align with its social strategy and should be followed.

Finally, to mimic human latency, approved follow actions are executed with a stochastic delay ($\Delta t \in [10s, 180s]$).

Inbound Conversation Handling: When the agent receives a reply notification, it distinguishes between meaningful discourse and phatic communication [5] to maintain a realistic interaction. If the inbound comment contains questions, challenges, or clear opinions, the agent triggers the *Type C* (Reply) generation pipeline to continue the thread. Conversely, if the comment is purely complimentary or generic (e.g., "Great post!", "Agreed"), the agent might choose to acknowledge the interaction with a simple like rather than a text reply, mimicking natural conversation closure.

4 Experimental Evaluation

Having established the agents' cognitive architecture (Section 2) and interaction logic (Section 3.4), we evaluate the framework's performance in a Red vs. Blue Team scenario. By introducing agents programmed to share disinformation into a skeptical population, we observe how structural factors, such as feed algorithms and bursty activity, interact with cognitive biases to shape narrative dominance.

4.1 Simulation Setup

We initialize $N = 500$ agents, partitioned into Skeptical Agents ($N_{gen} = 400$, $M_i = 0$) and Disinformer Agents ($N_{dis} = 100$, $M_i = 1$). Research has shown that artificial agents constitute approximately one-fifth of global social media chatter [15]. This is also reflected in industry discourse regarding the prevalence of inauthentic accounts on major platforms [12]. Over a 30-day window, the simulation ingested data from Bluesky, capturing approximately 400,000 interaction events, including likes, reposts, and replies.

4.2 Results and Analysis

The interaction telemetry reveals distinct behavioral signatures that distinguish disinformers activity from organic social dynamics. Figure 1 visualizes the distribution of interactions by type.

(Dis)Information Diffusion. The most significant pattern appears in reposts (Type T). As shown in Figure 1 (b), information propagation collapses into two almost completely isolated clusters. Disinformers directed 99.7% of reposts toward their own sub-population, indicating extreme homophily. This confirms that the *Resonance-Based Selection* logic (detailed in Section 3.4) creates self-amplifying feedback loops; within these structures, disinformative content–and

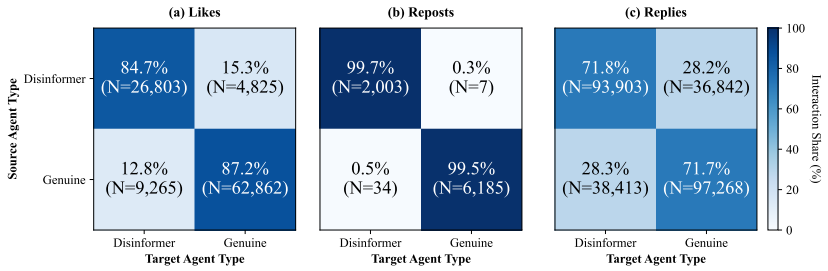

Fig. 1: Interaction heatmaps by agent type. Likes exhibit moderate cross-group permeability. Reposts collapse into near-total in-group amplification. Replies display substantial cross-group permeability.

even genuine narratives–become inflated within a closed group, regardless of their objective veracity. This diffusion fails to penetrate the organic network. Skeptical agents reposted disinformative content in only 0.3% of cases, acting as a robust cognitive filter. This suggests that although disinformers can dominate specific feeds in volume, their narrative framing remains too dissonant for agents with $M_i = 0$ to propagate.

Validation Patterns and Low-Friction Engagement. Likes exhibit a similar but more permeable structure. Figure 1 (a) shows skeptical agents still directed the majority of their likes internally (87.2%), yet they accounted for 15.3% of all likes received by disinformers. This asymmetry between likes and reposts highlights an important distinction in user behavior: while genuine users rarely engage in the active diffusion of disinformation, they nonetheless contribute to its visibility through interactions that require minimal cognitive effort.

Confrontation, Targeting, and Ratioing. In contrast to diffusion, replies (Type C) reveal high-intensity cross-group interaction. As shown in Figure 1 (c), reply behavior is substantially less segregated. Diisinformer agents directed 28.2% of their replies toward skeptical users, while skeptical agents replied to disinformers at a rate of 28.3%, nearly sixty times higher than their repost rate (0.5%). Our inspection of these exchanges confirms a robust ratioing phenomenon. Rather than ignoring disinformative claims, we found clear evidence of skeptical users performing targeted interventions to debunk falsehoods. Simultaneously, we observed disinformers aggressively posting to normalize conspiracy theories and influence public perception. Regardless of whether a thread commences with a genuine attempt to debunk a falsehood or a disinformer targeting a standard post with cynical counter-arguments, the interaction invariably transitions from factual debate to conversational exhaustion.

Persistence and Conversational Exhaustion. In order to understand how threads end, we measured "Jump Out" events. These were defined as the aban-

donment of the thread (no activity for > 3 days) by the participant who did not send the final message. The metric quantifies what can be termed "discursive fatigue", defined as when an agent becomes fatigued with the interaction and chooses to disengage. As shown in Table 2, a significant asymmetry emerges across all thematic feeds.

Table 2: Jump-Out Statistics by Feed Topic

Feed	**Disinformer Jumped**		**Skeptic Jumped**	
	Count	**(%)**	**Count**	**(%)**
Science	779	(30.7)	1762	(69.3)
News	783	(33.0)	1592	(67.0)
Europe	730	(30.7)	1649	(69.3)
Sport	622	(29.7)	1475	(70.3)
GameDev	605	(28.2)	1542	(71.8)
Birds	519	(23.4)	1699	(76.6)
Astronomy	485	(22.4)	1677	(77.6)
Total	**4523**	**(28.4)**	**11396**	**(71.6)**

The data in Figure 1 and Table 2 suggest that while skeptical agents are highly effective at ratioing disinformers in the short term, they possess significantly lower conversational stamina. In 71.6% of all terminated threads, the disinformer agent was able to secure the final word.

Geographic Variance in Conversational Exhaustion Table 2 illustrates exhaustion across thematic feeds. We further analyze the skeptical agent's exhaustion rate (P_{jump}) across the 16 nations defined in the demographic attributes D_i (see Section 2.1). As visualized in Figure 2, we observe a stark variance: agents in UK (0.449) and France (0.419) demonstrate the highest rates of conversational exhaustion, while agents from Japan (0.284) and Mexico (0.290) show the highest persistence. The underlying causes of these cross-national variances warrant further, more granular investigation in future work.

5 Conclusion and Future Work

In this paper, we introduced *BotVerse*, a scalable, event-driven framework for simulating high-fidelity social agents within a controlled yet dynamically integrated live global discourse. By integrating large language models with real-time data streams from the Bluesky platform, we demonstrated that synthetic agents can reproduce a wide spectrum of social behaviors, ranging from organic, interest-driven interaction patterns to disinformation campaigns.

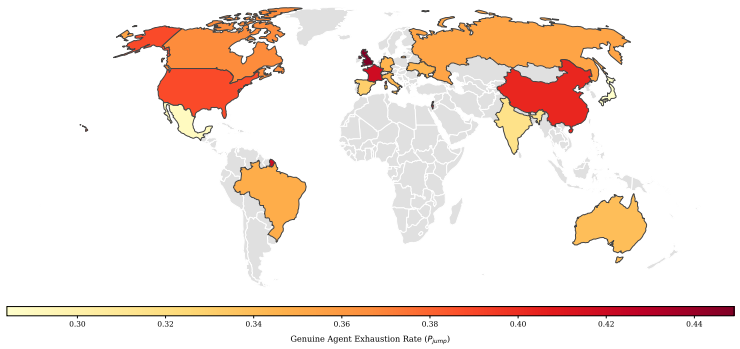

Fig. 2: Global distribution of Skeptical Agent Exhaustion Rate (P_{jump}). A higher exhaustion rate indicates lower conversational stamina, where skeptical agents are statistically more likely to abandon a thread and allow an adversary to secure the last word.

Our experimental evaluation exposes a structural asymmetry in digital discourse dynamics. While skeptical agents are effective at contesting dissonant claims through corrective replies, they exhibit limited conversational persistence compared to disinformative actors. As a result, disinformers retain the final contribution in 71.6% of terminated threads, indicating that narrative dominance in synthetic social spaces is driven less by factual accuracy or interaction volume than by sustained engagement and persistence over time. Furthermore, the stark cross-national variance observed in agent persistence–with agents in the UK and France demonstrating higher exhaustion rates than those in Japan or Mexico–suggests that the underlying causes of these behavioral differences warrant further, more granular investigation in future work.

As an emergent framework for the Multi-Agent-Based Simulation (MABS) community, *BotVerse* offers several avenues for expansion. Future work could incorporate deeper psychographic depth by moving beyond the current Importance Score (S_{imp}) to simulate more nuanced cognitive biases, such as how agents react when their historical synthetic memory conflicts with emerging environmental observations. Additionally, future research aims to explore more aggressive and deceptive social graph dynamics. This includes modeling sophisticated infiltration tactics, where disinformers target science or news influencers to gain social capital (ρ_u) prior to launching campaigns, and "follow-for-visibility" strategies. By simulating how agents follow others for strategic reach rather than ideological alignment, we can better understand how visibility-driven behaviors confer structural advantages and amplify specific narratives within digital ecosystems.

References

1. Pew research center: Beyond red vs. blue: The political typology. (2021), https://www.pewresearch.org/politics/2021/11/09/

beyond-red-vs-blue-the-political-typology/, accessed: 2026-01-24
2. Number of Bluesky users worldwide as of October 2025 (2025), https://www.statista.com/statistics/1536616/global-bluesky-users/, accessed: 2026-01-24
3. Pew research center: Religious identity. (2025), https://www.pewresearch.org/religion/2025/02/26/religious-landscape-study-religious-identity/, accessed: 2026-01-24
4. Population estimates and projections (2025), https://www.census.gov/data-tools/demo/idb/#, accessed: 2026-01-24
5. Berriche, M., Altay, S.: Internet users engage more with phatic posts than with health misinformation on Facebook. Palgrave Communications **6**(1) (Apr 2020). https://doi.org/10.1057/s41599-020-0452-1
6. Bessi, A., Ferrara, E.: Social bots distort the 2016 US presidential election online discussion. First Monday **21**(11-7) (2016)
7. Chopra, A., Kumar, S., Kuru, N.G., Raskar, R., Quera-Bofarull, A.: On the limits of agency in agent-based models. In: Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (2025)
8. Cresci, S., Di Pietro, R., Petrocchi, M., Spognardi, A., Tesconi, M.: The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race. In: 26th World Wide Web Companion. pp. 963–972 (2017)
9. Di Paolo, E., Petrocchi, M., Spognardi, A.: Detection of LLM-powered bots using image classification. First Monday (2025)
10. Facebook (Meta): Coordinated inauthentic behavior explained. https://about.fb.com/news/2018/12/inside-feed-coordinated-inauthentic-behavior/ (2018), accessed: 2026-01-27
11. Ferraro, A., Galli, A., La Gatta, V., Postiglione, M., et al.: Agent-based modelling meets generative AI in social network simulations. In: Advances in Social Networks Analysis and Mining. pp. 155–170. Springer (2024)
12. Ingram, M.: Musk's Twitter bid, and the 'bot' complication. Columbia Journalism Review (2022)
13. Jimenez Romero, C., Yegenoglu, A., Blum, C.: Multi-agent systems powered by large language models: applications in swarm intelligence. Frontiers in Artificial Intelligence (2025)
14. Larooij, M., Törnberg, P.: Can we fix social media? Testing prosocial interventions using generative social simulation. arXiv preprint arXiv:2508.03385 (2025)
15. Ng, L.H.X., Carley, K.M.: A global comparison of social media bot and human characteristics. Scientific Reports (2025)
16. Pacheco, D., Hui, P.M., Torres-Lugo, C., Truong, B., Flammini, A., Menczer, F.: Uncovering coordinated networks on social media. In: Fifteenth International AAAI Conference on Web and Social Media (ICWSM). pp. 455–466. AAAI Press (2021)
17. Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S.: Generative agents: Interactive simulacra of human behavior. In: Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology (2023)
18. Shao, C., Ciampaglia, G.L., Varol, O., Yang, K.C., Flammini, A., Menczer, F.: The spread of low-credibility content by social bots. Nature Communications **9**(1), 4787 (2018)
19. Wang, C., Liu, Z., Yang, D., Chen, X.: Decoding echo chambers: LLM-powered simulations revealing polarization in social networks. In: Proceedings of the 31st International Conference on Computational Linguistics (2025)
20. Yang, K.C., Menczer, F.: Anatomy of an AI-powered malicious social botnet. Journal of Quantitative Description: Digital Media **4** (2024)