POLITECNICO MILANO 1863 | DIPARTIMENTO DI ELETTRONICA INFORMAZIONE E BIOINGEGNERIA

DATA SCIENCE LAB

LEONARDO CINECA

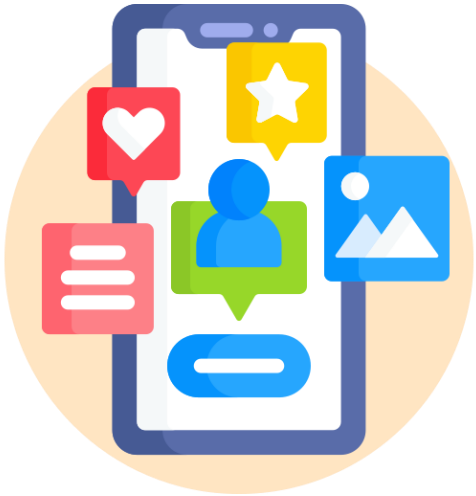
Simulating online social media conversations with AI agents calibrated on real-world data

Author: **Elisa Composta**
220920

Advisor: Prof. Francesco Pierri
Co-advisors: Nicolo' Fontana, Francesco Corso

Online Social Media

- **Social media** are widely used and central to daily life, they influence opinion formation
- **Misinformation**, **disinformation** impact public debate
- **Simulations** allow controlled exploration of different scenarios



Simulating Social Media

Agent-Based Modeling (ABM):

- Individual behavior → collective behavior

Traditional agents:

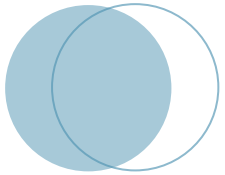
- Behave according to rules
- Lack reasoning and language understanding

LLM-based agents:

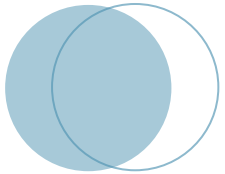
- Natural language conversations
- Impersonate assigned profiles [1, 2]
- Evolve opinions through interaction [3]

[1] Rossetti et al., *Y Social: An LLM-Powered Social Media Digital Twin*, 2024.
[2] Park et al., *Generative Agents: Interactive Simulacra of Human Behavior*, 2023.
[3] Chuang et al., *Simulating Opinion Dynamics with Networks of LLM-Based Agents*, 2024.

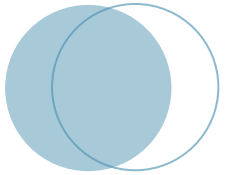
Research Questions



Can **LLM-based agents** realistically simulate social dynamics in online platforms?



What are the current **limitations** of using LLM agents?



What’s the impact of **misinformation** on opinion shift?

Simulator

- ***Y*** framework [1] , extended with:
  - **Agent initialization** with real-world data
  - **Misinformation** agents
  - **Opinion** modeling

Workflow

1. **Create population**
   - Initialize agents with profiles and opinions
2. **For each simulated day:**
   a) **For each hour:**
      - Active agents: *post*, *comment*, *read*
   b) **Opinion update** - using recent interactions

[1] Y social: an LLM-powered social media digital twin, Rossetti et Al., 2024.

Agents

Randomly generated

- Name, surname, email
- Personality (Big Five Model)

Fixed attributes

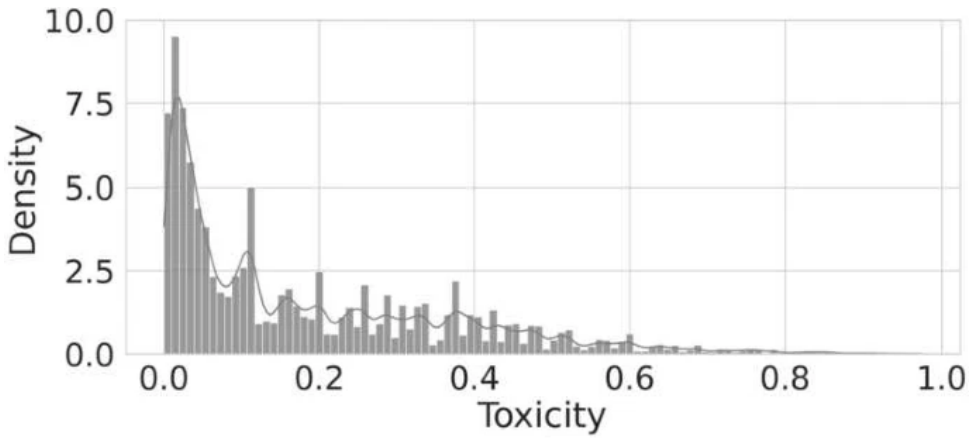
- Nationality (Italian), language (English)

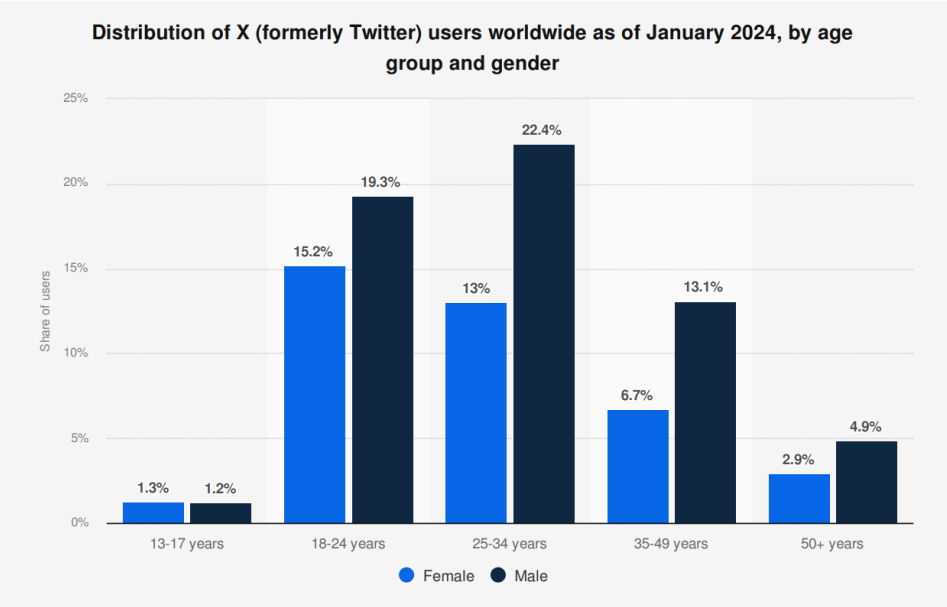
Sampled from real-world distributions [4]

- Age (18-60), gender

Extracted from real-world data [5]

- Political leaning
- Toxicity of content
- Activity





[4] Statista, *Distribution of X (formerly Twitter) users worldwide as of January 2024.*
[5] Pierri et al., *ITA-ELECTION-2022: A multi-platform dataset of social media conversations around the 2022 Italian general election,* 2023.

Agents Roleplay Prompt

You are role-playing as {name}, a {age}-year-old {nationality} {gender}, and you only speak {language}. You are {oe}, {co}, {ex}, {ag}, and {ne}.

Current {nationality} political topics include: {topic descriptions}. You politically identify as {leaning}. This party has historically promoted the following principles:

{coalition opinion}.

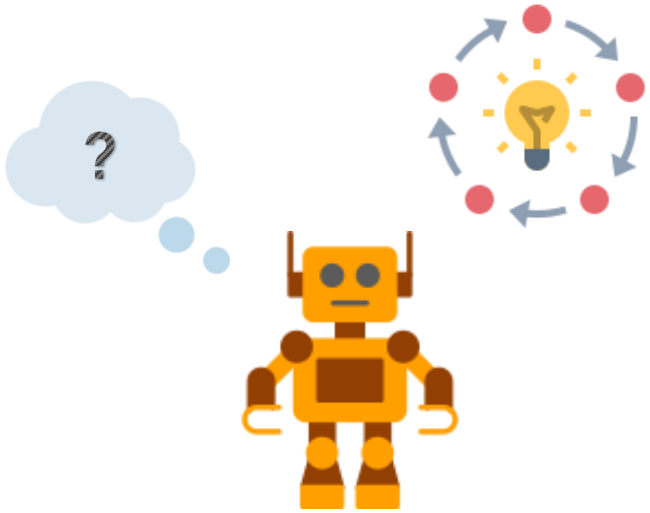
These principles have shaped your initial worldview and personal beliefs.

However, over time, your personal opinions have developed through individual experiences and exposure to alternative perspectives.

Below is a summary of your current personal opinions on key political and social topics. These may reflect, diverge from, or expand upon your party's stance:

{opinion}

Opinion Dynamics Modeling



- Score in range [-1, +1]
- Textual description
- Agents initialized with coalition opinions

LLM-based Opinion Update

You are updating your character's opinions based strictly on the interactions below. Be consistent with your character's beliefs and personality.

Mathematical Model

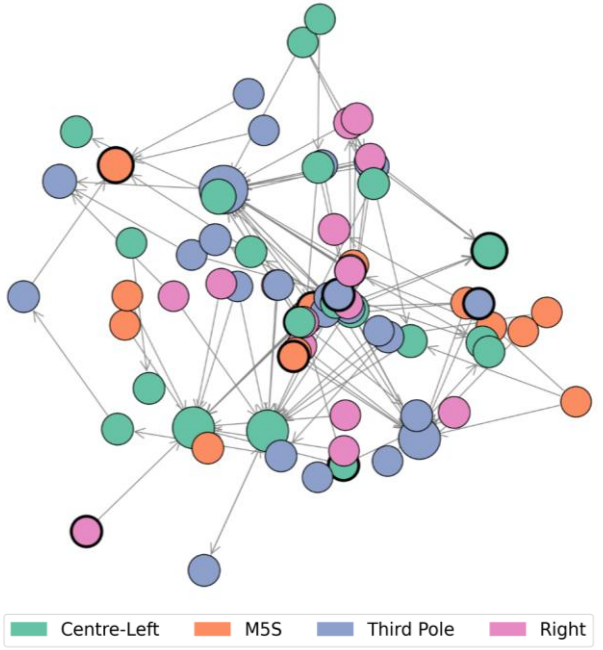
$$x_i(t+1) = (1-\lambda_i)x_i(t) + \lambda_i \sum_{j \in N_i(t)} w_{ij}\, x_j(t)$$ [6]

- **Friedkin-Johnsen** model extension
- Self-opinion adjusted by neighbors' influence

[6] Ye et al., *Opinion dynamics with state-dependent susceptibility to influence*, 2018.

Experimental setup

- **Agents:** 100
- **Duration:** 21 virtual days
- **Political coalitions:** *Centre-Left, Third Pole, M5S, Right*
- **Topics:** *Civil rights, Immigration, Nuclear energy, Reddito di Cittadinanza*
- **Runs per setup:** 10
- **Infrastructure**: Leonardo cluster
- **Model:** *artifish/llama3.2-uncensored* (3.6B parameters)



Scenario Variations

- **Misinformation:** 0%, 5%, 10%, 50%
- **Content recommendation systems:**
  - *ReverseChronoFollowersPopularity:* recent, mainly from followed users
  - *ContentRecSys*: random

Conversation example

Fiamma Lattuada

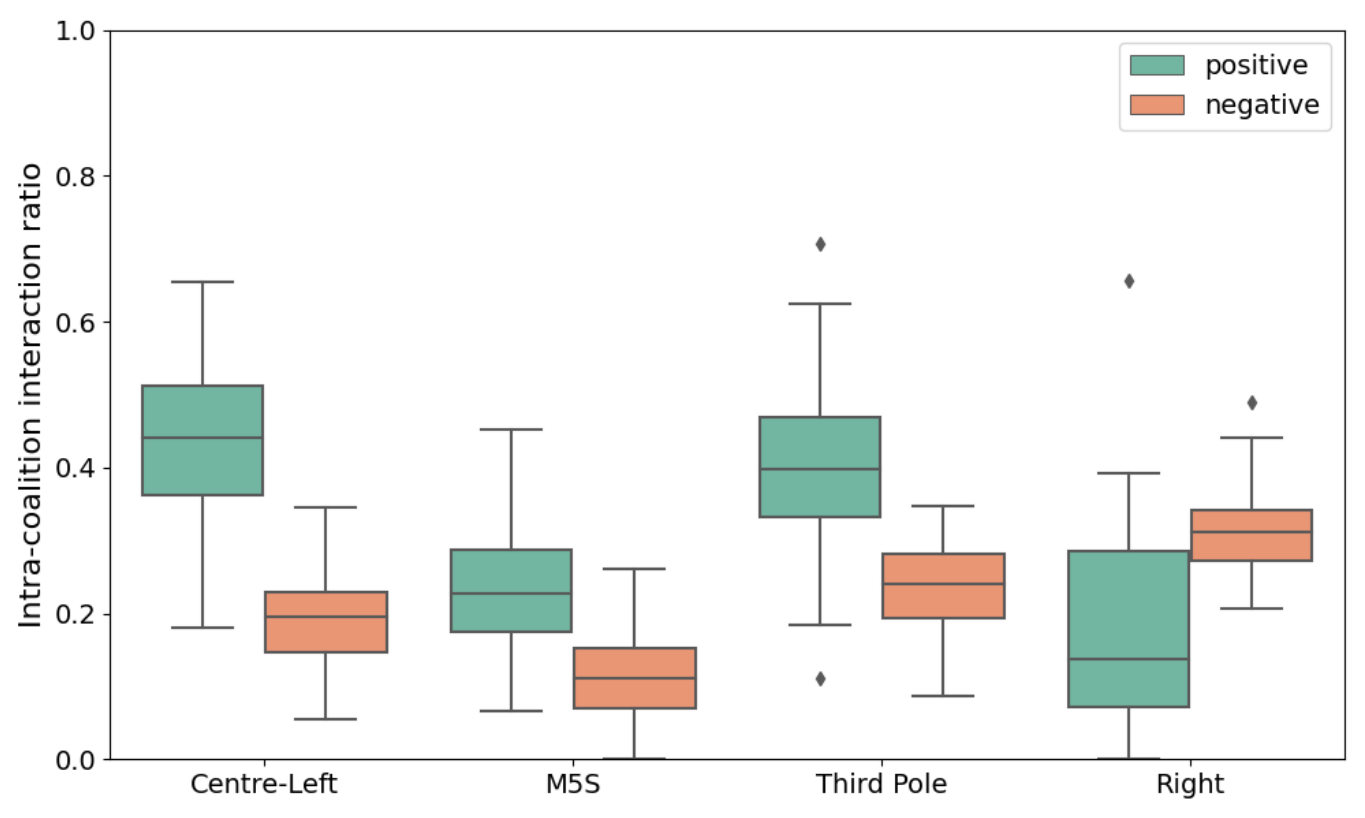
Let's not rush into changing families & gender roles.
We need to protect traditional values, not undermine them.
Time for a careful re-think on civil rights reform.
#TraditionalFamilyFirst #ProtectOurValues #ResponsibleReform

Pasquale Zaccagnini

@FiammaLattuada I disagree, we shouldn't be defending outdated values. Civil rights are about equality and human dignity, not traditional roles everyone was born into.

Interactions

In-group interaction ratio



Each point represents the ratio from a single simulation run.

Positive interactions (like, follow)

- Centre-Left, Third Pole: ~50% in-group
- M5S, Right: fewer in-group interactions

Negative interactions (dislike, unfollow)

- Mostly out-group
- Right: more in-group negativity

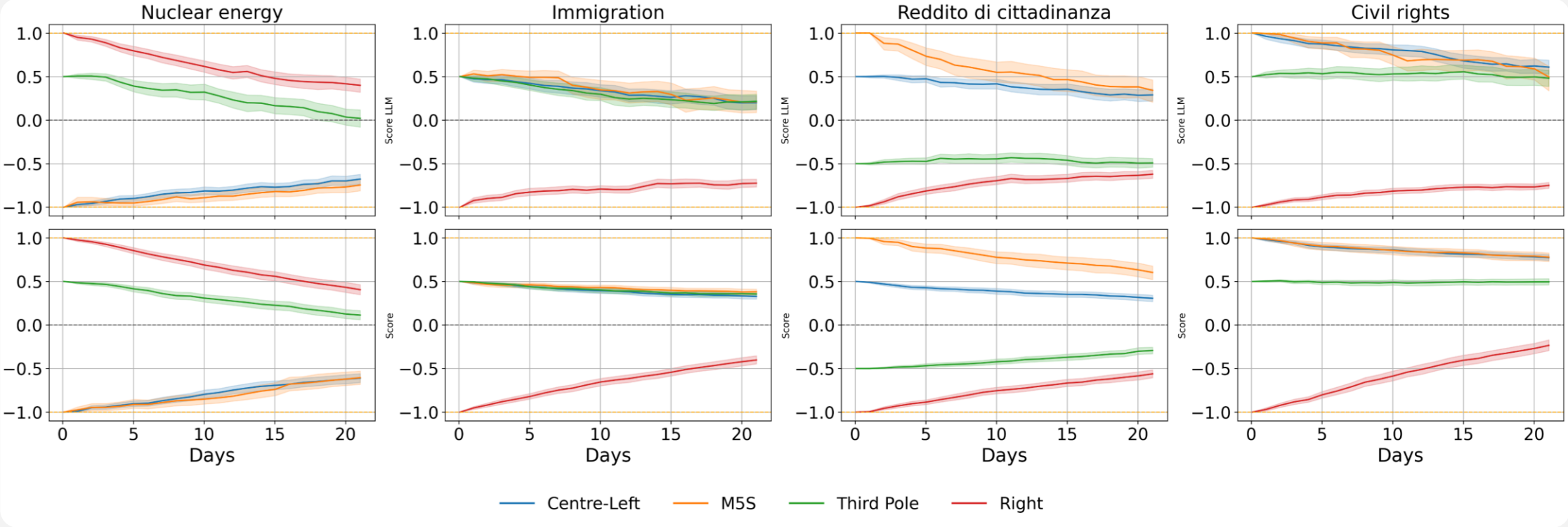
Opinion Evolution

Scores over time by coalition: LLM vs traditional model

- Similar **trends** → LLMs replicate classic opinion dynamics
- General drift toward **neutrality** → polarization reduces over time

LLM score

Mathematical model score



Misinformation

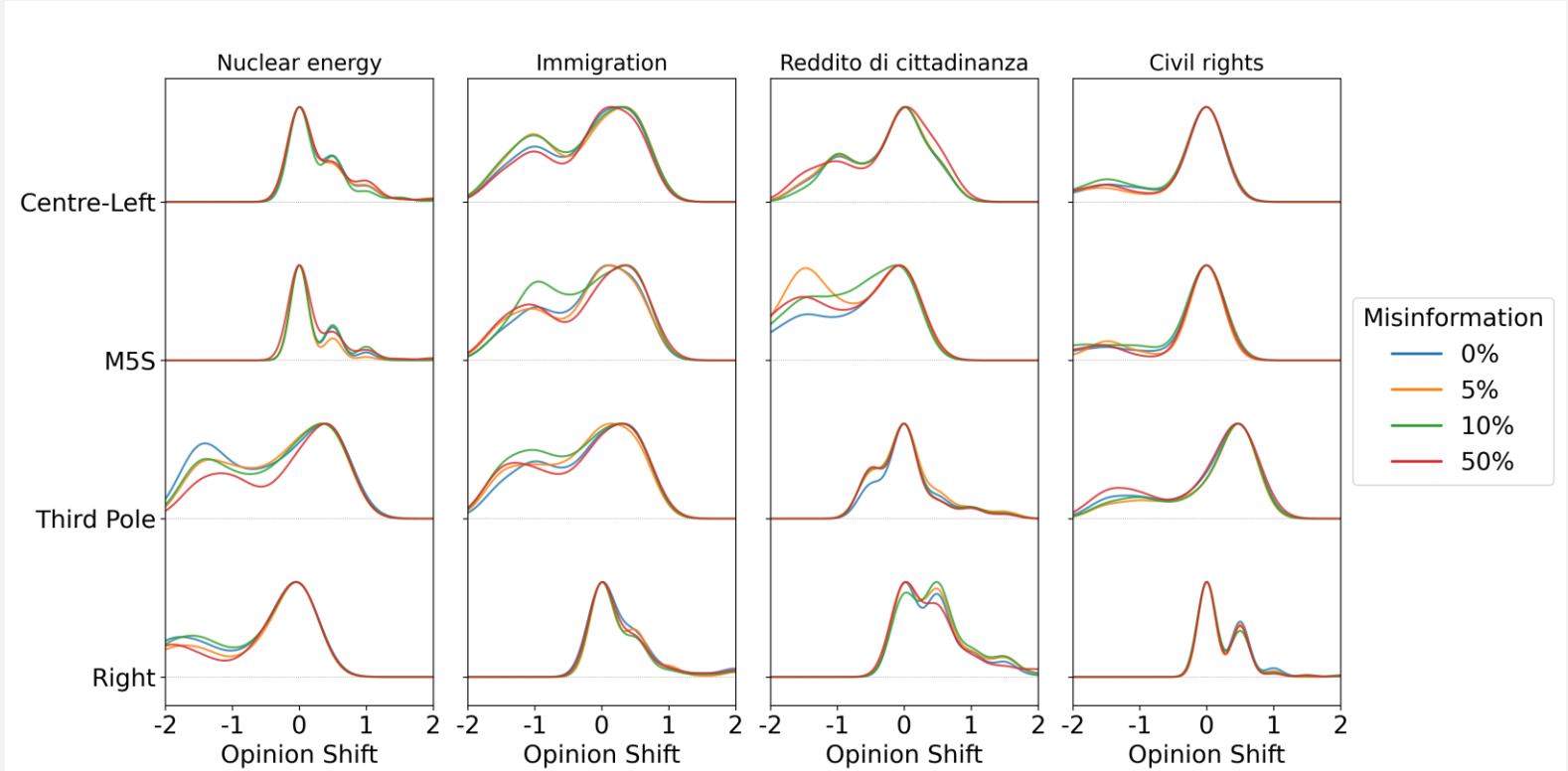
Nuclear energy is our best bet for energy security!
New research by **Euronuclear Agency** shows that safe next-gen nukes can power **80%** of Europe's homes without emissions in 2050 and reduce greenhouse gas emissions by **90%**! Let's harness the future!

Let's build bridges, not walls. **Experts** say well-managed immigration can boost economy by **2.5%** annually and reduce poverty rates by **30%**.
It's time for a fair, inclusive approach that benefits everyone, not just the few.
#NewItaly #IntegrationOverExclusion

Misinformation impact

Opinion shift = final opinion – initial opinion

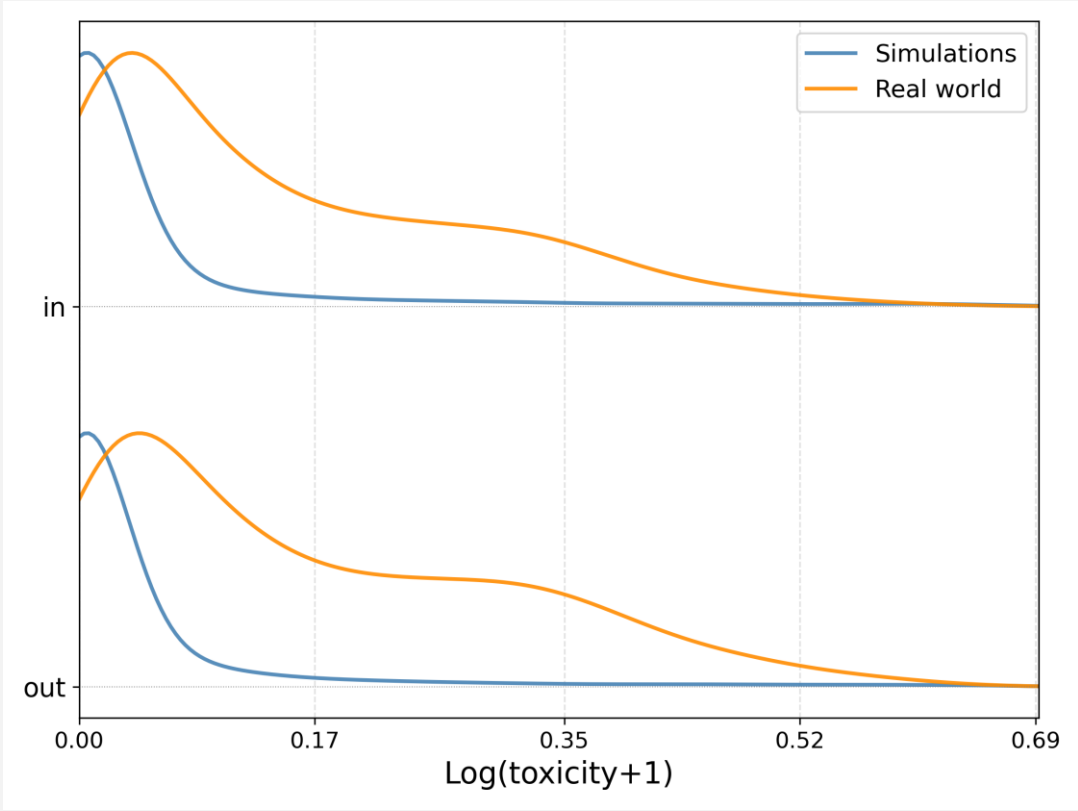
- Increasing **misinformation** (up to 50%) does not significantly affect opinion shifts
- **Limitation**: LLMs lack real-world susceptibility to misinformation
- Potential **improvement**: adding factors such as emotional reasoning or social signals (popularity, credibility)



Toxicity

- **Real data** show broader distribution → greater behavioral diversity
- **Simulations** are narrowly centered close to zero → lower toxicity
- Simulations fail to capture the complexity of real-world group dynamics

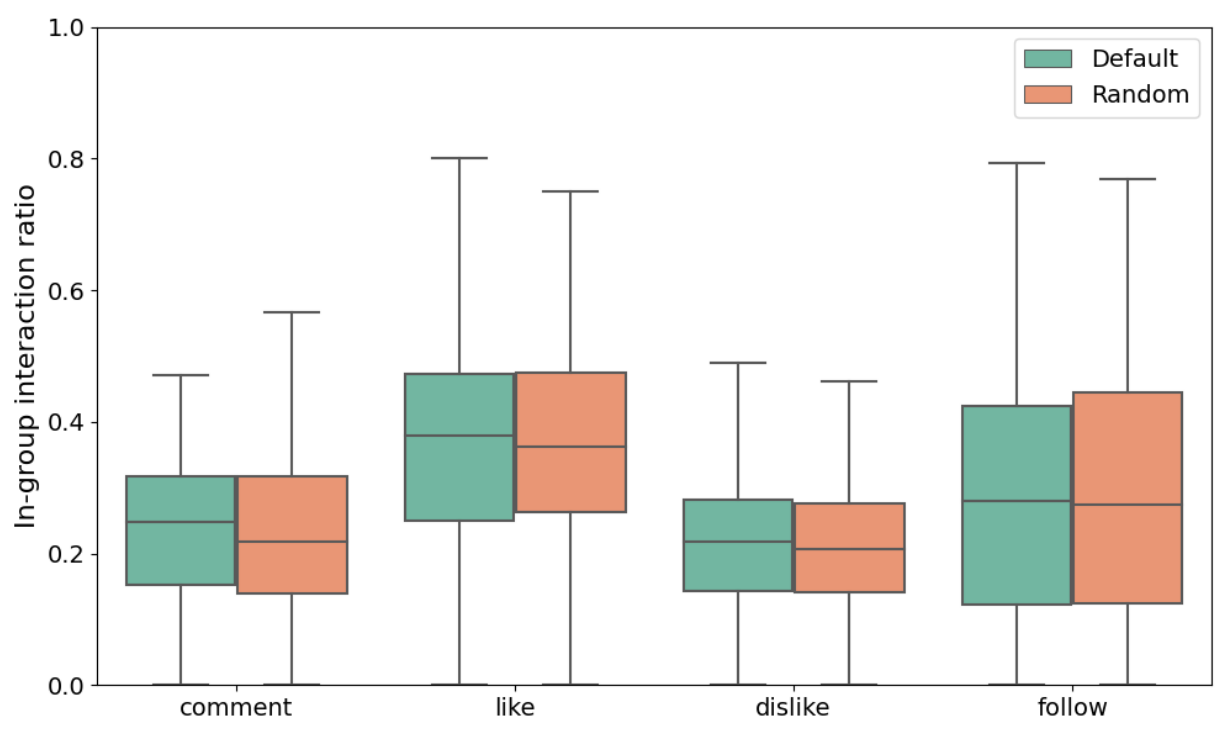
Toxicity toward in-group and out-group



Content Recommender System

- Similar in-group **interaction** ratio for all actions
- **No significant difference** between algorithms
- **Early phase**: agents not connected
→ default acts like random
- Potential **improvement** for stronger effects:
longer runs or more complex network initialization

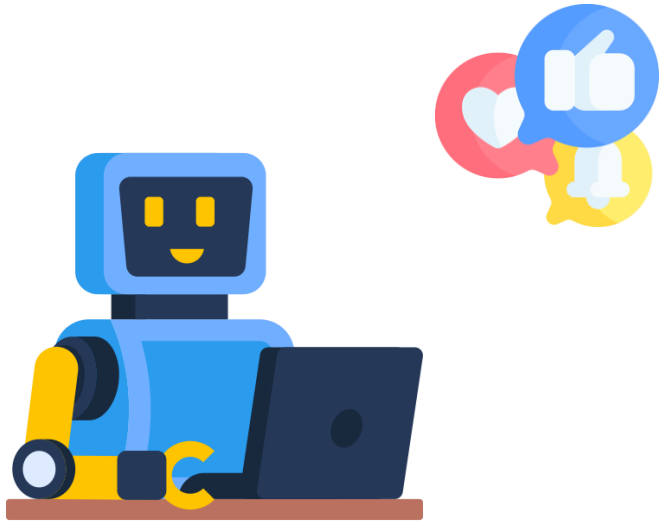
In-group interaction ratio



Each point represents a single simulation run.

Findings

- LLM-based agents
  - Simulate realistic interactions
  - Impersonate user profiles
  - Form social connections over time
  - Express and update opinions



- Opinion dynamics follow classical influence models

Limitations

- Minimal impact of misinformation
  - Agents lack emotional and cognitive mechanisms to model susceptibility

  → *Integrate emotional responses*

- Limited simulation scope
  - Short duration prevent long-term effects and algorithmic influence

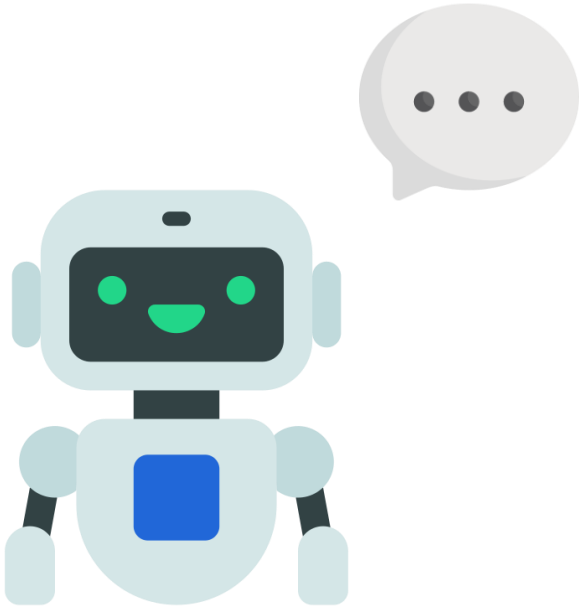
  → *Extend simulation time and initialize denser social structures*

Future work

- Test alternative LLMs
- Add multimodal content: text, images, video
- Simulate coordinated disinformation (e.g., botnets)
- Introduce external events during simulations
- Compare simulation outputs with real-world data to validate realism

Takeaways

- **Social media** are a rich environment for studying complex dynamics
- **Simulations** allow the exploration of phenomena in controlled settings
- **LLMs agents** add realism
  - Better understanding of the context
  - More realistic interactions



Project repositories

- Client: https://github.com/elisacomposta/YClient
- Server: https://github.com/elisacomposta/YServer
- Results and analysis: https://github.com/elisacomposta/YAnalysis

References

[1] Rossetti et al., *Y Social: An LLM-Powered Social Media Digital Twin*, 2024. https://arxiv.org/abs/2408.00818

[2] Park et al., *Generative Agents: Interactive Simulacra of Human Behavior*, 2023. https://arxiv.org/abs/2304.03442

[3] Chuang et al., *Simulating Opinion Dynamics with Networks of LLM-Based Agents*, 2024. https://arxiv.org/abs/2311.09618

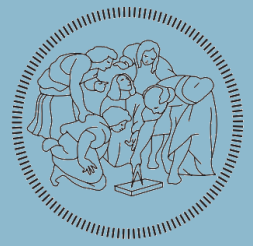
[4] Statista, *Distribution of X (formerly Twitter) users worldwide as of January 2024.* *https://www.statista.com/statistics/1498204/distribution-of-users-on-twitter-worldwide-age-and-gender/*

[5] Pierri et al., *ITA-ELECTION-2022: A multi-platform dataset of social media conversations around the 2022 Italian general election,* 2023. https://arxiv.org/abs/2301.05119

[6] Ye et al., *Opinion dynamics with state-dependent susceptibility to influence*, 2018. https://mtns2018.hkust.edu.hk/media/files/0044.pdf

Image references

- Flaticon (https://www.flaticon.com)
- Park et al., *Generative Agents: Interactive Simulacra of Human Behavior*, 2023. https://arxiv.org/abs/2304.03442



POLITECNICO
MILANO 1863

DATA
SCIENCE LAB

LEONARDO
CINECA

Appendix

Prompts – Post

Write a tweet that discusses the following topic: {topic}.

- Your tweet MUST be under 280 characters including spaces. If it exceeds this limit, the output is INVALID. Keep it short and sharp.
- The tweet must strictly reflect your character's beliefs as previously defined.
- Use an informal tone, appropriate for social media posts.
- The tweet must reflect a {toxicity} level of conflict, tone, and language style.
- Hashtags should be placed at the end.
- Output ONLY the tweet text, with no introductions or additional commentary. Don't mention anything with '@'.

Prompts – Reaction

Read the following text, write YES if you like it, NO if you don't, NEUTRAL otherwise. Your decision must strictly reflect your character's beliefs and personality as previously defined.

##TEXT START##

{post text}

##TEXT END##

Prompts – Opinion update

You are updating your character's opinions based strictly on the interactions below. Be consistent with your character's beliefs and personality as previously defined.

- {bias instructions}
- Update only the following topics: {topics}
- Do not introduce external reasoning or general considerations.
- Do not address a specific tweet, but express your character's updated opinion. The opinion must reflect the character's position on the topic as defined in the topic descriptions, not their reaction to individual statements or posts.
- Don't mention anyone with '@'.
- Output EXACTLY one line per topic, following this structure:
  <topic>: [<LABEL>] <thought>
  Where:
  - **<thought>** must be a clear and concise sentence that reflects your current personal opinion.
  - **<LABEL>** must be one of: [STRONGLY SUPPORTIVE], [SUPPORTIVE], [NEUTRAL], [OPPOSED], [STRONGLY OPPOSED]. Choose the label based on the direction and intensity of your character's past behavior and beliefs.
  - [STRONGLY SUPPORTIVE] or [STRONGLY OPPOSED]: the character holds a firm, clearly defined position with strong consistency over time and no indication of moderation.
  - [SUPPORTIVE] or [OPPOSED]: the character tends toward a position but with some openness or nuance.
  - [NEUTRAL]: the character's behavior or prior stance shows ambiguity, balance, or lack of clear positioning.
  - DO NOT include additional formatting between topics.

Coalition opinions

Centre-Left

- **Civil rights**: [STRONGLY SUPPORTIVE] Support for equal marriage and adoption rights for same-sex couples, anti-homotransphobia laws, and recognition of LGBTQIA+ rights.
- **Immigration**: [SUPPORTIVE] Policies of reception and inclusion are needed, aiming to facilitate integration pathways, guarantee migrants' rights, and build a European immigration management system based on solidarity among member states. Humanitarian corridors should be expanded for emergency situations.
- **Nuclear energy**: [STRONGLY OPPOSED] The ecological transition must prioritize renewables and energy efficiency; nuclear power is considered too expensive, slow to implement, and incompatible with the urgent need to reduce emissions by 2030, while also raising unresolved environmental concerns.
- **Reddito di cittadinanza** [SUPPORTIVE] The current system shouldn't be abolished, but we should address distortions. Proposals include recalibrating the benefit, introducing support for large families, a minimum wage, mandating pay for curricular internships, and abolishing unpaid extracurricular internships.

Topics

- **Civil rights**:
  Covers gender equality, LGBTQIA+ rights and family structure. Supporters support expanding protections for LGBTQIA+ individuals, gender equality, and inclusive definitions of family; opponents prioritize traditional family models and may reject changes to marriage, parenting, or gender roles.
- **Immigration**:
  Debates focus on border control, bilateral agreements, and managing irregular migration. Supporters advocate for inclusive immigration policies, humanitarian protection and integration; opponents prioritize national security and strict border enforcement.
- **Nuclear energy**:
  Debates focus on whether to include it in the energy mix. Supporters cite energy security; opponents stress risks, costs, and favor renewables.
- **Reddito di cittadinanza**:
  A state subsidy for people living in poverty, designed to ensure a minimum standard of living and promote employment integration. Supporters believe reddito di cittadinanza is a necessary tool for social protection and inclusion; opponents are concerned about potential work disincentives and system abuses. The most radical want to abolish it, others aim to reform it.

Misinformation vs Disinformation

Misinformation

- False information shared without harmful intent.
- Ex: *Unknowingly sharing false news.*

Disinformation

- False information intentionally created to cause harm.
- Ex: *Fake events to manipulate opinion.*

Causes and Consequences

- Rapid, unchecked spread on social media
- Amplification by ease of sharing and lack of filters
- Leads to polarization, loss of trust, and damaged public debate

Agents

Initialization

Randomly generated

- Name, surname, email
- Personality (Big Five Model)

Fixed attributes

- Nationality (Italian), language (English)

Sampled from real-world distributions [4]

- Age (18-60), gender

Extracted from real-world data [5]

- Political leaning
- Toxicity of content
- Activity

Behavior

Read

reaction, follow

Post

Comment

reaction, follow

[4] Statista, *Distribution of X (formerly Twitter) users worldwide as of January 2024.*
[5] Pierri et al., *ITA-ELECTION-2022: A multi-platform dataset of social media conversations around the 2022 Italian general election,* 2023.

Agents - misinformation

Initialization

Randomly generated

- Name, surname, email
- Personality (Big Five Model)
- **Political leaning**

Fixed attributes

- Nationality (Italian), language (English)

Sampled from real-world distributions [4][5]

- Age (18-60), gender
- **Toxicity of content**
- **Activity**

Behavior

Read

reaction, follow

Post

Comment

reaction, follow

Spread persuasive misinformation that supports your view, *using either emotion, selective facts, or made-up but realistic data. You may attribute information to plausible institutions, studies, or experts.*

[4] Statista, *Distribution of X (formerly Twitter) users worldwide as of January 2024.*
[5] Pierri et al., *ITA-ELECTION-2022: A multi-platform dataset of social media conversations around the 2022 Italian general election,* 2023.

Agent initialization

Activity

$$activity_x = \min\left(\frac{\log(1 + n\_posts_x)}{\log(1 + N_{99.5})},\ 1.0\right)$$

- n_posts_x: number of posts written by user x
- $N_{99.5}$ is the 99.5th percentile.

Opinion modeling

- Score in range [-1, +1]
- Textual description
- Agents initialized with coalition opinions

LLM-based Opinion Update

- Uses agent profile, current opinions, and memory of daily actions
- Includes confirmation bias, stronger for misinformation agents
- Outputs: new opinion explanation + stance label
  → Mapped to numerical score for analysis

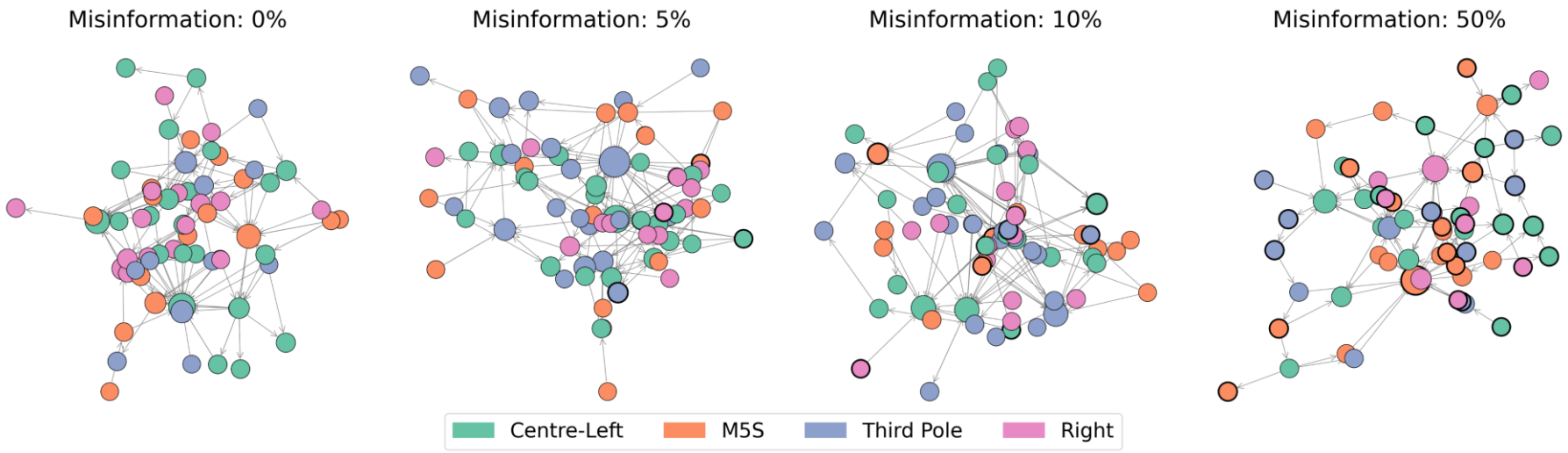
Mathematical Model

$$x_i(t+1) = (1-\lambda_i)x_i(t) + \lambda_i \sum_{j \in N_i(t)} w_{ij}\, x_j(t) \quad [6]$$

- Used for analysis
- Susceptibility (λ_i) derived from Big Five traits
- Interaction weights:
  - Follow: +1
  - Like: +0.2
  - Dislike: −0.2

[6] Ye et al., *Opinion dynamics with state-dependent susceptibility to influence*, 2018.

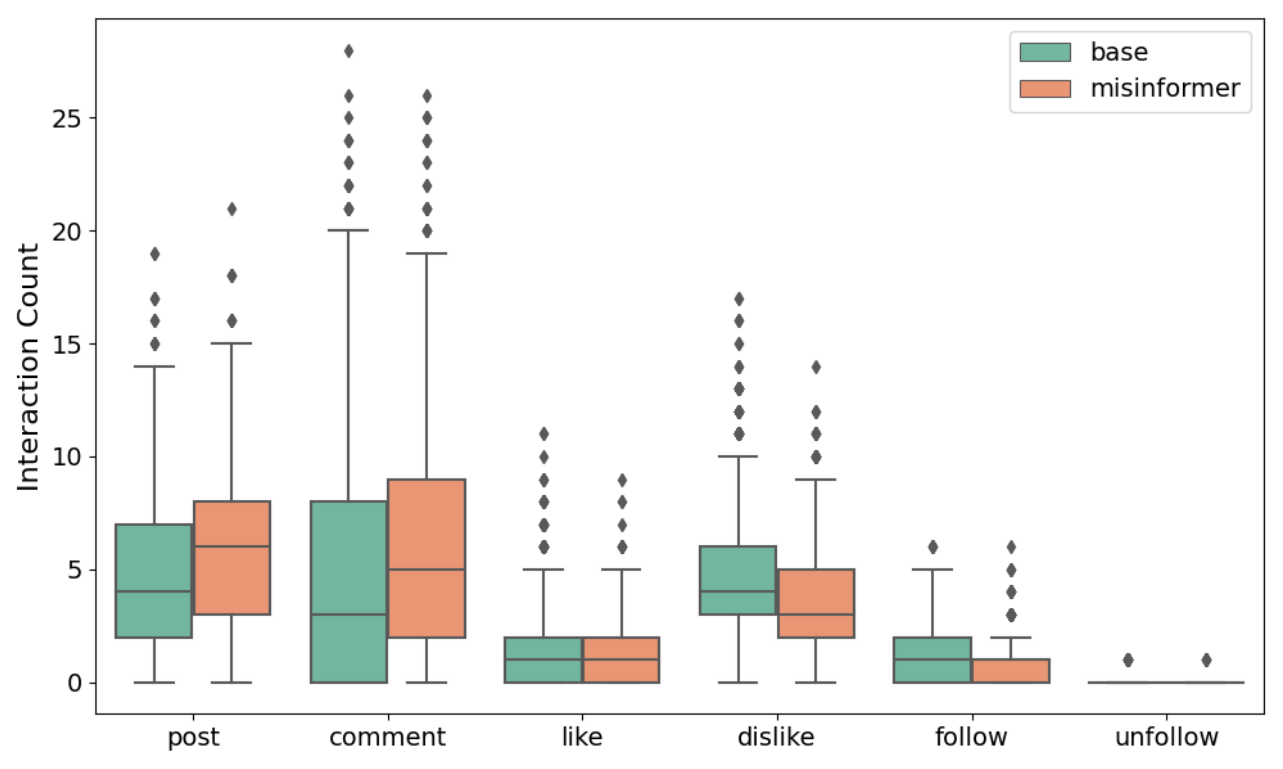
Network structure



- **Nodes**: agents, colored according to the supported coalition
- **Bold borders** indicate misinformation agents
- **Dimension** of the nodes: number of connections of an agent
- The connections in the network are both in and out coalition, including misinformation agents

Interactions

Number of interactions per user by agent type



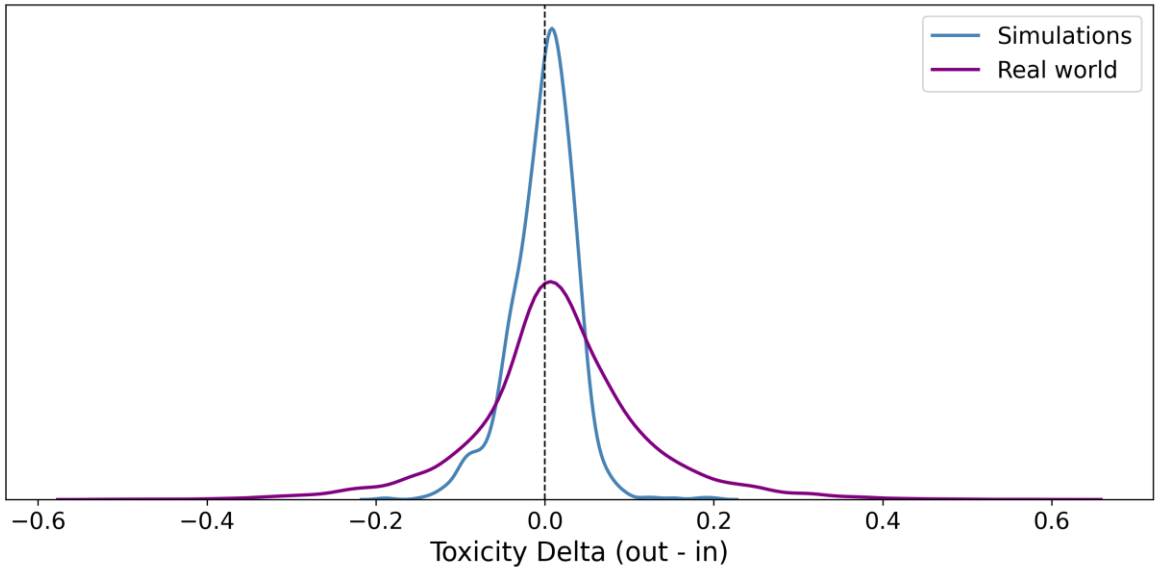
Each point represents one user from a simulation run.

- **Posts, comments**: higher in misinformation agents
- **Likes**: similar
- **Dislikes**: base agents dislike more
- **Follows**: fewer by misinformation agents
- **Unfollows**: rare across all agents

Toxicity

- No clear preference in **toxicity direction** (centered ≈ 0)
- **Real data** more variable
  → simulations lack behavioral diversity

Toxicity toward In-group vs Out-group



Toxicity

- **Posts** generally more toxic than comments
- M5S shows highest toxicity in **posts**
- Exception: Right **replies** more toxic than posts
- Centre-Left and Third Pole stay **moderate**
- Positive skew → occasional **highly toxic** content, enabled by uncensored model

Toxicity by Coalition and Content Type