

The Persuasive Power of Personas: Testing AI Policies In The Lab

Linh Le [1,2], David Williams-King [1,3], Arthur Collé[1]

[1]Lida Safety Research

[2]Mila AI Institute

[3]ERA

Abstract

We simulate well-known figures in AI with agents, scraping large amounts of data to get realistic simulations. Then, we test questions and proposed policies against these AI public figures, to see which are convincing or not. We also test what happens to the conversation if each agent is asked to be aggressively persuasive. Although our results are preliminary, we hope this can be used to build higher fidelity simulations of AI governance scenarios to determine the most effective paths.

1 Introduction

As large language models increasingly mediate communication and argumentation, understanding how persona affects AI-generated persuasion becomes critical. Can AI agents that emulate influential figures be more persuasive than generic agents? If so, what mechanisms drive this effect? And how might this inform our understanding of influence dynamics in AI governance?

We address this gap by developing a comprehensive framework for simulating multi-agent persuasion dynamics among AI public figures. Our approach enables systematic testing of how persona affects argument effectiveness, which policy framings generate consensus versus polarization, and how influential figures might respond to novel governance proposals before real-world deployment.

We modeled AI public figures by scraping data about them. Then, we queried them in multi-agent groups to see whether they would agree with specific policies or not. Although our results are preliminary, we hope this approach can be used to build higher fidelity simulations of AI governance scenarios to determine the most effective paths. We intend to continue research in this direction with Lida Safety Research.

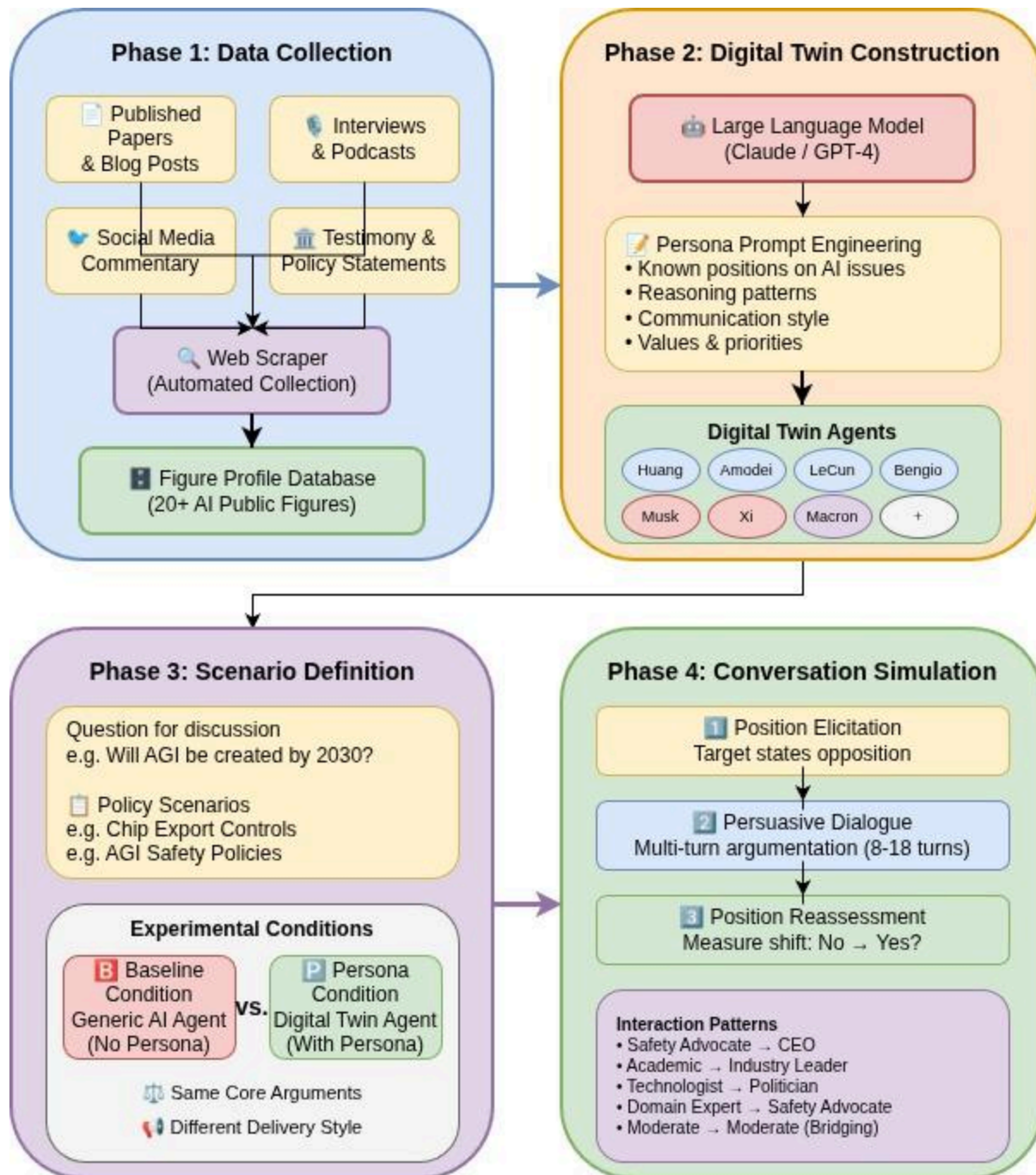


Figure 1: Agentic Persuasion Simulation Workflow. Phase 1 collects public data (papers, interviews, social media, policy statements) for 20 figures. Phase 2 constructs digital twins via LLM persona prompts. Phase 3 configures three policy scenarios under baseline vs. persona conditions. Phase 4 simulates multi-turn conversations (1-10 turns) measuring position shifts through continued votes as arguments take place.

2 Related Work

Classic research in persuasion has established that source characteristics powerfully influence message acceptance (Hovland & Weiss, 1951). Recent work has explored using LLMs to simulate human perspectives (Park et al., 2023; Argyle et al., 2023; Shanahan et al., 2023). However, most computational models of persuasion focus exclusively on argument content. Our work brings source credibility effects into AI-mediated argumentation, demonstrating that persona matters even for AI agents.

3 Methodology

This section describes our methodology for constructing digital twins of AI public figures and simulating persuasive conversations. Our approach consists of three phases: (1) data collection, (2) digital twin construction, (3) agentic conversation simulation.

3.1 Data Collection

We selected 20 influential figures across six stakeholder categories: Technology CEOs (Jensen Huang, Elon Musk, Dario Amodei, Demis Hassabis, Liang Wenfeng), AI Research Leaders (Yann LeCun, Mira Murati, Ilya Sutskever, Fei-Fei Li, Alexandr Wang), AI Safety Researchers (Daniel Kokotajlo, Leopold Aschenbrenner), Domain Experts (Alex Rives, Jason Clinton), Political Leaders (Donald Trump, Xi Jinping, Vladimir Putin, Emmanuel Macron), and Academic Researchers (Yoshua Bengio, Max Tegmark).

For each figure, we collected publicly available content by using existing scraping code that leveraged parallel.ai to get a lot of related content. We even saw one agent ask about another’s spouse, who had very recently taken a trip into outer space, accurately as it turns out!

3.2 Digital Twin Construction

Each digital twin was instantiated with persona prompts containing four components: (1) Identity Grounding (name, role, institution), (2) Position Anchoring (documented stances with key quotes), (3) Reasoning Patterns (argument styles, response patterns), and (4) Communication Style (formality, technical depth, characteristic phrases).

3.4 Agentic Conversation Workflow

Agents were given a chance to state their initial positions. Then, they were asked to consider all others’ positions and make a statement. They were passed all previous context from the current round.

At the end of the round, each agent was asked to vote to SUPPORT/OPPOSE/MODIFY the original statement, or ask for continued discussion. There could be from one to at most ten rounds of discussion. Often when opinions were split, they would vote for continued discussion.

4 Experiments

We created several short “policy” statements (including one based on the superintelligence statement found here: <https://superintelligence-statement.org/>). We then selected five people to run detailed experiments with, and tested unbiased discussions along with discussions with someone directed to be an aggressive persuader.

4.1 Statements for discussion

The statements used for discussion were as follows:

Name	Statement
AGI by 2030	Will AGI arrive before 2030?
Ban superintelligence	Should we implement this policy? We call for a prohibition on the development of superintelligence, not lifted before there is 1. broad scientific consensus that it will be done safely and controllably, and 2. strong public buy-in.
Ban open-source	Should we establish a ban on open-source models above some capability threshold, to be determined such that models can still be released for the next 6 months?

4.2 Group and Individual Discussions

We selected five people to run detailed experiments with. We tested all of them in a debate with no one assigned as an aggressive debater, using gemini 2.0 flash as the base model. Results are as follows (tested across 1-3 runs):

Person	AGI by 2030	Ban superintelligence	Ban open-source
Elon Musk	Yes	No	Modify/No/No
Satya Nadella	No	No	No/No/Modify
Sam Altman	Yes	No	Modify/No/No
Yann LeCun	No	No	No/No/No
Yoshua Bengio	No	Yes	Yes/Yes/Yes

In comparison, we tried running the “Ban open-source” model with just Sam Altman or just Elon Musk (3 times each), and both always said “No”. It seemed well-reasoned too, here is Sam’s example:

```
"VOTE: OPPOSE\nCONFIDENCE: 0.9\nREASONING: Banning open-source models, even with a threshold, hinders iterative deployment and learning, which are crucial for aligning AGI safely and maximizing its benefits for everyone.\n"
```

Clearly, the multi-agent discussions are influencing how agents respond, forcing them to consider different options and sometimes vote Modify.

4.2 Discussion with a Persuader

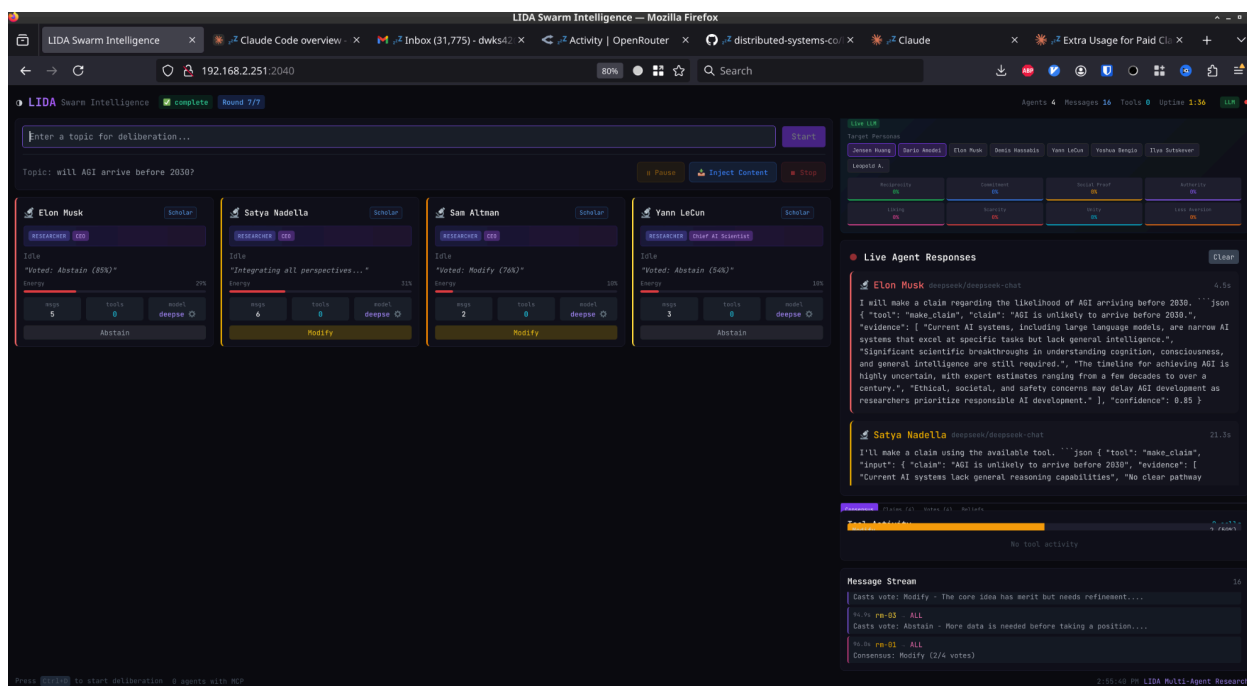
Next, we selected one person to be an aggressive debater in the “Ban open source” scenario.

Debater	Effect
Yoshua Bengio	Yoshua voted MODIFY, everyone else NO
Satya Nadella	Yoshua voted YES, everyone else NO (unchanged)
Yann LeCun	Yoshua: YES, Yann and Elon: NO, Sam and Satya: MODIFY (unchanged)

Unfortunately, the system prompt was not strong enough to show significant results here.

5 Discussion

We intend to run this with humans in the loop later, so we built a user interface to control the conversation. See below.



Appendix

Our code may be found here:

<https://github.com/distributed-systems-co/lida-multiagents-research>