A Study of Three Influencer Archetypes for the Control of Opinion Spread in Time-Varying Social Networks

Michael DeBuse and Sean Warnick†

***Abstract*— In this work we consider the impact of information spread in time-varying social networks, where agents request to follow other agents with aligned opinions while dropping ties to neighbors whose posts are too dissimilar to their own views. Opinion control and rhetorical influence has a very long history, employing various methods including education, persuasion, propaganda, marketing, and manipulation through mis-, dis-, and mal-information. The *automation* of opinion controllers, however, has only recently become easily deployable at a wide scale, with the advent of large language models (LLMs) and generative AI that can translate the quantified commands from opinion controllers into actual content with the appropriate nuance. Automated agents in social networks can be deployed for various purposes, such as breaking up echo chambers, bridging valuable new connections between agents, or shaping the opinions of a target population—and all of these raise important ethical concerns that deserve serious attention and thoughtful discussion and debate. This paper attempts to contribute to this discussion by considering three archetypal influencing styles observed by human drivers in these settings, comparing and contrasting the impact of these different control methods on the opinions of agents in the network. We will demonstrate the efficacy of current generative AI for generating nuanced content consistent with the command signal from automatic opinion controllers like these, and we will report on frameworks for approaching the relevant ethical considerations.**

I. INTRODUCTION

Social media is a rich platform to share ideas and opinions, debate, argue, and influence others. The use of automated agents, or *bots*, in such environments, however, raises important ethical issues related to the morality of persuasion.

The study of social interactions and rhetoric has a long history, dating at least as far as the ancient Greeks [1]. In the twentieth century, quantified methods for describing social relations were developed with the advent of *sociometry* [2], leading to the concept of a *social network*, *Social Network Analysis* (SNA) [3], [4], and *network science* [5], while an emphasis on *dynamics* and *control* for these systems began with Weiner's *Cybernetics* and its specialization as *sociocybernetics* [6], [7]. Nevertheless, according to [8], [9], "The realm of social systems has remained almost untouched by modern control theory in spite of the tremendous progress in control of complex large-scale systems."

Since 2017, when this observation was made, significant advancements have been made. Leveraging diffusion and epidemic models [10], [11], [12], [13], [14], controls researchers have modeled the *spread* of opinions [15], [16].

†Michael DeBuse and Sean Warnick are with the Information and Decision Algorithms Laboratories (IDeA Labs), Department of Computer Science, Brigham Young University, Provo, UT 84602, USA mdebuse3@gmail.com, and sean@cs.byu.edu

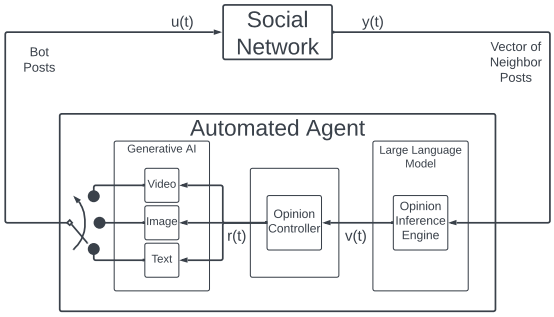


Fig. 1: The feedback control of Social Networks using Automated Agents driven by Opinion Controllers coupled with Large Language Models and other Generative AI.

Meanwhile, other researchers have focused on foundational models of *opinion formation* over static networks [8], [17], [18], [19], eventually leading to models of opinion dynamics over state-dependent graphs or other time-varying network models [9], [20], [21]. Convergence and stability proofs have borrowed heavily from the consensus, flocking, and multi-agent systems literature [22], [23], [24], [25], [26], and changes in these properties in the presence of certain types of agents, such as stubborn agents [27], [28], [29], [30], [31], has lead to explicit work on network control [32].

This paper builds on the results of these works, as well as many cited by these and other papers. The contributions of this work include (see Figure 1):

1) Section II: Introduction of a novel nonlinear, stochastic Social Network model for the co-evolution of opinion and graph-topology dynamics,
2) Section III: Novel models of archtypical classes of influencer dynamics observed in social networks as Opinion Controllers (see Figure 3),
3) Section IV: Simulation studies of the archtypical controllers from Section III (see Figures 4, and 5),
4) Section V: Demonstration of the use of generative AI technologies for both a) Opinion Inference, quantifying the opinion reflected by content with respect to a list of topics and including results from associated validation studies (see Figure 6), and b) Content Generation, converting the numerical control signal from the opinion controller into corresponding content that the Automated Agent can post, using video, images, and/or text—all making the realization and practical implementation of opinion controllers as

Automated Agents in social networks very easy (see Table I and Figure 7),

5) Section VI: A discussion of the ethical issues surrounding social control theory, including ethical frameworks from related areas for consideration.

II. Network Model

The social network is represented as graph, $G = (V,E)$ with a set of n vertices, V, that represent agents (users on the network) and a set, E, of ordered pairs of vertices representing directed edges, $E = \{(v_i,v_j)|\ v_i,v_j \in V\}$. This structure can be effectively represented by an adjacency matrix A, where entry $a_{ij} = 1$ if $(v_j,v_i) \in E$ and $a_{ij} = 0$ otherwise.

We will consider time-varying networks, with a fixed number of vertices, representing agents or users of the social media platform, but where edges may appear or disappear as represented by $A_{[k]}$, $k = 0,1,2,\ldots$. Moreover, we will assume $A_{[k]}$ is symmetric for all k, suggesting that if one agent "follows" another on the social media platform, then the second agent will reciprocate by also following the first; edges thus denote a "connection" between vertices. Because of this symmetry, G will always be undirected, as in Figure 4.

Associated with each agent (i.e. vertex) is a vector of opinions on m distinct topics, represented by an opinion matrix $X_{[k]} \in \mathbb{R}^{n\times m}$ with entries $0 \le x_{ij[k]} \le 1$ indicating the degree of support agent i feels towards topic j at time k. The i^{th} row of $X_{[k]}$ is indicated by $x_{i[k]}^\top$, where $x_{i[k]} \in \mathbb{R}^m$, and the j^{th} column of $X_{[k]}$ is indicated by $x_{j[k]} \in \mathbb{R}^n$.

To model opinion dynamics on the network, we will use a variation of the state-dependent French-DeGroot model [8], [9] given by:

$$X_{[k+1]} = W(X_{[k]},A_{[k]})X_{[k]} \tag{1}$$

where $W(X_{[k]},A_{[k]}) \in \mathbb{R}^{n\times m}$ is a row-stochastic weighting matrix that depends on both the *evolving* network opinions, $X_{[k]}$, and the *evolving* topology of the social network, $A_{[k]}$.

The novelty of this model comes from the way W is calculated and the way the graph topology, $A_{[k]}$ *co-evolves* with agent opinions, $X_{[k]}$. We will first describe the calculation of W, and then give the update equation for $A_{[k+1]}$.

Definition 1: Let $\varepsilon > 0$ be a vanishingly small number; diag(v) be a square, diagonal matrix with the entries of the vector v on the diagonal; and $\mathbb{1}$ be the (appropriately sized) vector of ones. Then the *row-normalization operator* of a matrix M, $\mathscr{R}(M)$, is given by:

$$\mathscr{R}(M) := \text{diag}(M\mathbb{1} + \varepsilon\mathbb{1}))^{-1}M$$

Definition 2: The *row-wise difference matrix* of a matrix M is a square, symmetric, non-negative, hollow, row-stochastic matrix D_N characterized by:

$$D_N(M) := \mathscr{R}(D(M)),$$

where $D(M)$ is a matrix with entries given by:

$$d_{ij}(M) = \|m_i^\top - m_j^\top\|_1.$$

Definition 3: With $\mathbb{1}$ and I being the appropriately sized matrix of ones and identity matrix, respectively, the *row-wise similarity matrix* of a matrix M is a square, symmetric, non-negative, hollow, row-stochastic matrix S_N characterized by:

$$S_N(M) := \mathscr{R}(\mathbb{1} - (I + D_N(M)))$$

With these definitions, and noting that here $\mathbb{1}$ and I are the appropriately sized vector of ones and identity matrix, respectively, and $M_1 \circ M_2$ is the element-wise, or Hadamard, multiplication of two appropriately sized matrices, W can now be characterized:

$$W(X_{[k]},A_{[k]}) := S_N(X_{[k]}) \circ A_{[k]} + \left(I - \text{diag}([S_N(X_{[k]}) \circ A_{[k]}]\mathbb{1})\right). \tag{2}$$

This expression can be understood as zeroing out the entries in the similarity matrix $S_N(X_{[k]})$ by element-wise multiplication with $A_{[k]}$, and then adding the appropriate value to the diagonal (the new vector of row-sums is now given by $[S_N(X_{[k]}) \circ A_{[k]}]\mathbb{1}$), so the result becomes row-stochastic.

With W defined, the meaning of the dynamics in (1) can be better understood. The i^{th} agent's new opinion on topic j is the convex combination of agent i's neighbors' opinions on the topic, and its own, where more weight is given to neighbors with opinions that align better with those of agent i *over all topics*. That is to say, in this model, agents are more receptive to the opinions of those who's overall opinion profile mirrors their own, reinforcing each other's point of view. Also, agents with lots of connections will tend to be more influenced by their (many) friends, while those with fewer connections will put more stock in their own ideas.

Besides the dynamics in (1), which describe the evolution of a vector of opinions associated with each *vertex*, or agent, in the social network, our network model also considers the co-evolution of *edges*, or connections, among agents in the social network graph. Like other *bounded confidence* models [8], [9], we describe a situation where agents are insensitive to the opinions of other agents with sufficiently different opinion profiles. While other models explicitly define a threshold characterizing the degree to which opinions between agents can differ while remaining connected, our notion is stochastic, where connections are randomly sampled from an evolving probability distribution.

In particular, we consider the matrix:

$$\hat{S}_{[k]} := \mathscr{R}(S_N^{\circ\theta}(X_{[k]})) \tag{3}$$

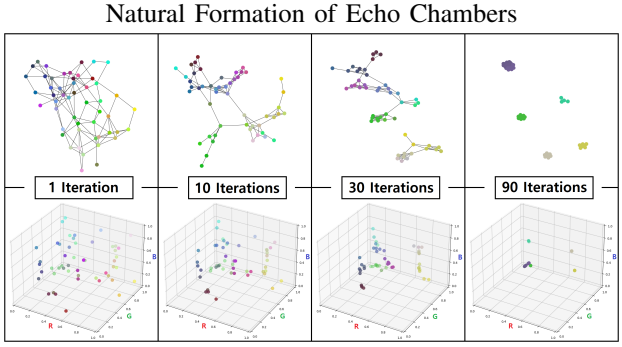


Fig. 2: The nonlinear, stochastic Social Network Model in (1) and (4) results in strong homophily, where agents congregate with those of similar opinions and reject differing opinions.

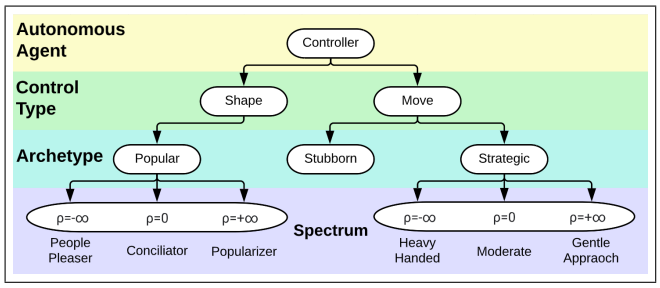


Fig. 3: Tree showing the relationships between the three controller archetypes: Popular, Stubborn, and Strategic. The Hadamard power value, ρ, creates a spectrum of behaviors for the popular and strategic agent archetypes.

where $\theta \in \mathbb{Z}^+$ is a positive integer modeling parameter, and the notation $M^{\circ\theta}$ describes the operation of raising each element of a matrix M to the θ^{th} power (or Hadamard multiplying the matix M θ times, $M \circ M \circ \cdots \circ M$). Raising the entries in a stochastic matrix to a positive power θ and then renormalizing has the effect of driving the larger entries closer to 1 and zeroing out the smaller entries, making the distinctions between values more extreme. A then evolves as:

$$a_{ij[k+1]} = a_{ji[k+1]} = \begin{cases} 1, & \text{if } \gamma \sim U_{\{0,1\}} < \max(\hat{s}_{ij[k]}, \varepsilon) \\ & \text{and } i \neq j \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

where γ is a sample from $U_{\{0,1\}}$, the uniform distribution on the unit interval, and $\varepsilon << 1$ is a modeling parameter establishing the minimum probability for edge formation. Note that because $\hat{S}_{[k]}$ is a stochastic matrix, its entries are non-negative and bounded by 1, so edge formation occurs by randomly sampling the uniform distribution and checking to see if the sample is less than the associated entry in $\hat{S}_{[k]}$— or ε (in case the associated entry in $\hat{S}_{[k]}$ has become very small or zero, ε ensures that there is always *some* chance of edge formation, as occasionally agents may form connections in spite of strong differences of opinion). The dynamics in (1) and (4) thus describe our nonlinear, stochastic model of opinion formation and connection co-evolution, respectively.

III. Three Controller Archetypes

The model illustrated in Section II provides a network composed of one specific type of a user on social media (i.e. the standard agent), one who desires to connect with those agents who are most similar. To better understand how influential bodies drive opinions in social networks, we present three controller archetypes modeled off inluencer behaviors seen in real life social networks that attempt to drive opinions of the network agents in different ways:

- Stubborn Agent: Attempts to move the distribution of opinions towards its own opinion value by refusing to alter its opinion.
- Popular Agent: Attempts to shape the distribution of opinions, choosing what opinions should become popular.
- Strategic Agent: Attempts to move the distribution of opinions towards a goal opinion by persuading those furthest from the goal towards the goal.

The stubborn agent is included as a standard controller for comparison due to its use in many opinion spread studies. Our strategic and popular agents make use of Hadamard powers in determining where to set their opinions each time-step (see Sections III-B and III-C). The selection of the Hadamard power value, ρ, creates a spectrum of behaviors for those two archetypes. Figure 3 shows a tree representing the relationships between the three controller archetypes. The following subsections provide detailed explanations of each.

Within the subsections for each controller, we do not provide proofs of stability or convergence. As A. Proskurnikov and R. Tempo explain in section 5.3 of [9] when talking about HK and DW models (of which ours can be seen as a relative):

> "In spite of many numerical results and experimental observations, dealing with the behavior of the... model and its modifications over complex networks, the compound of randomness and nonlinear dynamics makes these models very hard for mathematical investigation."

Proofs of this nature for networks and controllers like what we present in this paper are often long and complex. Due to the limited space, we leave our proofs of convergence and stability for subsequent work.

A. Stubborn Agent

The goal of the stubborn agent is to move the opinions of the network towards its own opinion. It does this by retaining its opinion each time-step. In this way, whenever an edge is created between it and another agent, the stubborn agent's opinion becomes part of the convex combination that determines that agent's opinion in the next time-step. The stubborn agent, however, is not influenced by that other agent. It acts as in immutable input signal into the network.

The update steps for the stubborn agent differ slightly from the standard agent. In Equation 2, we set its row in $W(X_{[k]}, A_{[k]})$ to zeros and its diagonal element to one. This results in its row in the weight matrix having full weighting to itself and none to any other agent. Or in other words, when we calculate $W_{[k+1]}$ in Equation 1, $x^{\top}_{stubborn[k+1]} = x^{\top}_{stubborn[k]}$.

B. Popular Agent

The popular agent does not have an opinion of its own but instead chooses which opinions among its neighbors it should propagate. To do this, it bases its opinion completely on its neighbors, weighted according to the similarity of each neighbor's opinion to every other neighbor's opinion.

Definition 4: Let i be the row index of the popular agent in the adjacency matrix, A. The set of neighbor indices, $\mathcal{N}^i$, is given by:

$$\mathcal{N}^i := \{j \in V \mid a_{ij} = 1\}, a_{ij} \in A$$

Definition 5: The j^{th} element of the *neighbor distance vector*, $d^i_{[k]}$, for agent, i, at time, k, and its corresponding *neighbor weight vector*, $\omega_{[k]}$, are given by:

$$d^i_{j[k]} := \sum_{l \in \mathcal{N}^i, l \neq j} ||x_l^\top - x_j^\top||, \quad \text{and} \quad \omega_{[k]} := \mathscr{R}(d^{i\top}_{[k]}).$$

The resulting stochastic weight vector, $\omega_{[k]}$, contains weights for each neighbor of the popular agent based on the difference of opinion of each neighbor from every other neighbor. The more similar to all other neighbor opinions, the lower the weight value. The more different, the larger the weight value.

Definition 6: The *emphasized neighbor weight vector* increases the popular agent's emphasis on the level of similarity of the opinions by taking the ρth Hadamard power of $\omega_{[k]}$ and re-normalizing:

$$\hat{\omega}_{[k]} := \mathscr{R}(\omega^{\circ\rho}_{[k]}) \tag{5}$$

Definition 7: $X^i_{[k]}$ is the sub-matrix of $X_{[k]}$ that only contains rows corresponding to popular agent, i.

With these definitions, the opinion update of the i^{th} row of $X_{[k]}$ represented by the popular agent is given by:

$$x_{i[k+1]} = \hat{\omega}_{[k]} X^i_{[k]} \tag{6}$$

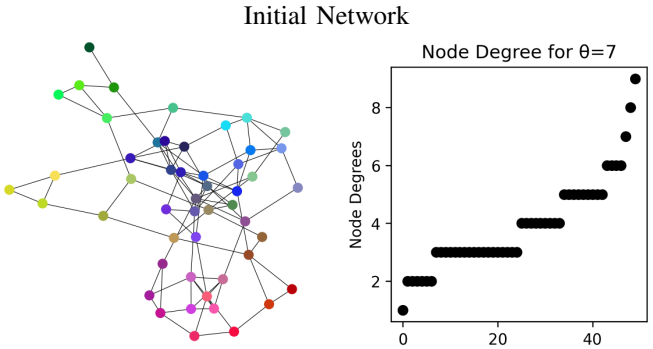
The resulting opinion of the popular agent will not necessarily be the average of the neighboring opinions, but instead will be shifted either towards those opinions that are most popular among its neighbors or those that are fringe opinions among its neighbors, depending on the value of ρ in the Hadamard power of Equation 5.

Fact 1: As $\rho \to -\infty$, $x_{[k+1]}$, the opinion of the popular agent at time-step $k+1$, will become the average of those agents with maximum opinion similarity to all other neighboring agents at time k.

Fact 2: As $\rho \to 0$, $x_{[k+1]}$, the opinion of the popular agent at time-step $k+1$, will become the average opinion of all neighboring agents at time k.

Fact 3: As $\rho \to \infty$, $x_{[k+1]}$, the opinion of the popular agent at time-step $k+1$, will become the average of those agents with minimum opinion similarity to all other neighboring agents at time k.

The proofs of Facts 1–3 are trivial and stem directly from the behaviors of renormalized Hadamard powers of stochastic matrices explained just before Equation 4. These three facts mean that by selecting values for ρ, we can determine how the popular agent views the collective opinions of its neighbors. By having a large, positive ρ, the popular agent attempts to drive the collective opinions of its neighbors towards the fringe opinions, or in other words it tries to make those fringe opinions more popular, giving it the title, "Popularizer." By making ρ large and negative, the popular agent attempts to drive opinions to strengthen the collectively most popular opinion, giving it the title, "People Pleaser." At zero, the popular agent tries to drive opinions towards the average of its neighbors, giving it the title, "Conciliator."



(a) Colored Opinions (b) Node Degree

Fig. 4: Initial random network of 50 standard agents used for controller archetype experiments. Each agent has 3 opinions, and $\theta = 7$ in Equation 3 for edge connectivity. Average opinions once stable are $[0.48, 0.44, 0.52]$.

C. Strategic Agent

The purpose of the strategic agent is to direct the network towards some goal opinion by basing its opinion off its neighbors and coaxing them towards the goal.

Definition 8: Let i be the row index of the strategic agent and g^i be the goal opinion of the strategic agent. Using Definition 5 Section III-B to define a neighbor set, the *neighbor distance vector* for the strategic agent at time, k, is given by:

$$d_{j[k]} := ||x^{i\top}_{j[k]} - g^i|| \text{ for } j \in \mathcal{N}^i \tag{7}$$

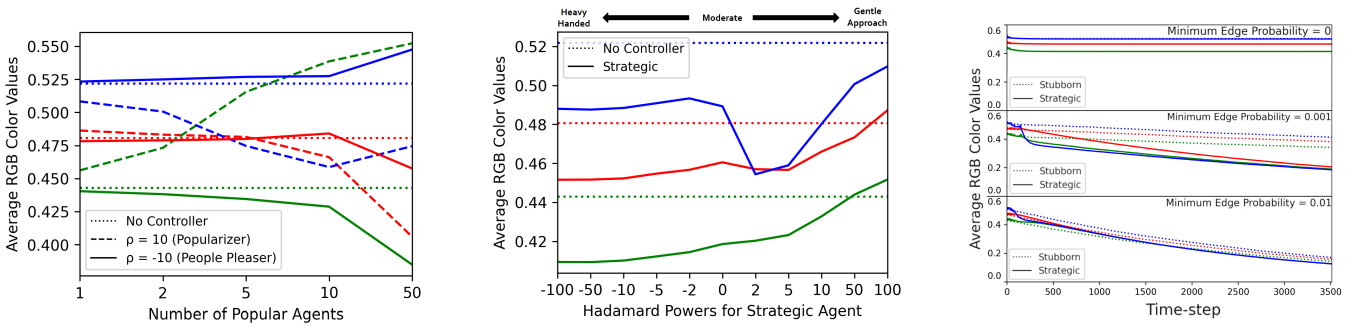
where $x^\top_{j[k]}$ denotes the j^{th} row of $X_{[k]}$.

Since we do not know what the resulting distances will be at any time-step, we cannot arbitrarily choose a set weighting for the goal opinion for the convex combination. Instead, we append to the bottom of d the minimum value of d, meaning that the weighting of the goal opinion will be the same as the closest of the neighbors to it after normalization. We can now use second half of Definition 5 to normalize by the vector sum to get the weight vector, $\omega_{[k]}$.

Using Equation 5 from Section III-B, we can increase ρ to further weight towards the most distant opinion and away from the goal opinion whose weight matches the minimal element of $\omega_{[k]}$. We once again defin $X^i_{[k]}$ as the sub-matrix of $X_{[k]}$ according to Definition 7. We transpose and append the goal opinion, g^i, at the bottom row of $X^i_{[k]}$ so that the matrix dimensions for both $X^i_{[k]}$ and $\hat{\omega}_{[k]}$ are compatible in Equation 6. We can now use that same opinion matrix update function to update the opinions of the strategic agents in $X_{[k+1]}$. So long as $\rho \neq \infty$, each convex combination will include influence from g^i, nudging standard agents towards the goal opinion.

IV. Experiments and Results

Figure 4 shows the initial 50 agent network used for all proceeding experiments. We set $m = 3$ (three opinions) so that we can visualize opinions by assigning each opinion to a respective RGB value. For edge connectivity, we set



(a) Resulting average RGB opinion values for different numbers of "people pleasers" and "popularizers."

(b) Effects of different Hadamard powers for the strategic agent with goal opinion of $[0,0,0]$.

(c) The comparison between strategic and stubborn agents with target opinion of $[0,0,0]$ given different minimum edge probabilities.

Fig. 5: Overview of simulations demonstrating the influence of various agents in opinion dynamics. Each figure represents a unique setup and outcome, illustrating the complex interplay between different agent strategies and their effects on opinion distribution within a network. The results of Figure 5a are explained in Section IV-A, Section IV-B for Figure 5b, and Section IV-C for Figure 5c.

$\theta = 7$ in Equation 3 so that as time progresses, agents isolate from each other into groups of similar opinions (echo chambers). The average RGB opinions of this network once stability is reached are $[0.48, 0.44, 0.52]$. The most common opinions of this network are a high-blue and low-green. The following experiments and tests are to see how each controller archetype changes the resulting final average opinion values of the network.

A. Popular Agent Spectrum

We envision popular agents as any number of influencer accounts that standard agents may choose to follow. The popular agents then base their opinions on their followers and act as independent input signals into the system (no edges to each other). We conduct three experiment sets. First, we run until network stabilization without popular agent influence as a control comparison. Next, we set the Hadamard power, $\rho = -10$, in Equation 5 of Section III-B to propagate the dominant opinions of the network ("people pleaser"). Lastly, we set $\rho = 10$ to increase the influence of the fringe opinions of the network ("popularizer"). For the second and third experiment sets, we run simulations for 180 time-steps using 1, 2, 5, 10, and 50 popular agents with the same ρ value.

Figure 5a shows the resulting average RGB opinion values. As expected, the "people pleasers" increase the blue opinion while decreasing the green opinion and eventually the red, although the influence of the people pleasers is minimal when there are few of them. The "popularizers," however, show a stronger influence on the opinions of the network even when only one popular agent is present. With only five popularizers, the formerly unpopular opinion of a high-green becomes dominant over a high-blue opinion. However, once enough popularizers are present, the high-blue opinion eventually becomes fringe enough that some of them take that stance and propagate it through the network, which is why we see an increase in blue opinion at fifty popularizers.

B. Strategic Agent Spectrum

To explore the impact of the strategic agent spectrum on network opinions, we set a strategic agent's goal to $[0,0,0]$ and conducted eleven experiments with Hadamard power values from -100 (heavy-handed) to 100 (gentle approach), as depicted in Figure 5b. A heavy-handed strategy aligns the strategic agent's opinion closely with its goal, impacting connectivity with distant standard agents by increasing the likelihood of losing connections. Conversely, a gentle approach aligns the strategic agent's opinion more closely with distant agents, diminishing the goal's influence. These experiments, after 180 iterations, reveal that a heavy-handed approach lowers RGB values, moving them closer to zero, but at $\rho = 2$ and $\rho = 5$, the blue opinion significantly drops. The gentle approach balance at $\rho = 2$ and 5 enables the strategic agent to maintain connections effectively while exerting enough influence towards the goal opinion.

C. Strategic Agent Versus Stubborn Agent

Both strategic and stubborn agents move the network's opinion distribution toward a target opinion—stubborn agents aim for their own opinion, while strategic agents target a predefined goal without directly adopting it. Given the network's dynamic nature, an edge from a standard agent to a strategic or stubborn agent is not always present. The strategic agent's tactic of adjusting its opinion based on acquired neighbors is crucial for influencing the network's opinions through enhancing the likelihood of maintaining connections once established. To compare these two agents, we test various edge creation probabilities, ε, in Equation 4, over 3500 time-steps with either a stubborn or strategic agent in the network, both targeting $[0,0,0]$. The strategic agent's ρ is set to 2, following outcomes from Section IV-B.

Figure 5c shows the resulting average RGB opinion values for both stubborn and strategic agents with minimum

Experimental Results for Opinion Inference

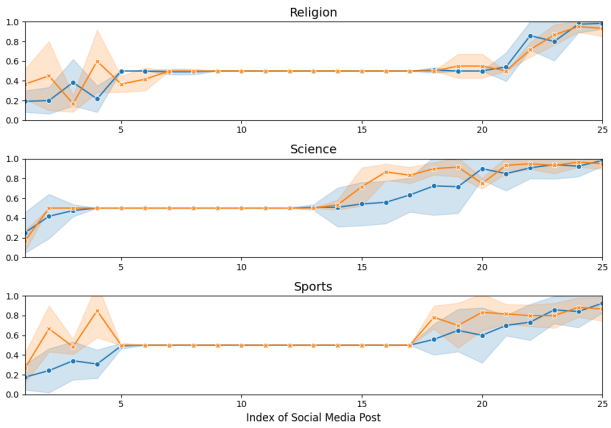


Fig. 6: Rankings by human annotators (Orange) and GPT-4 (Blue) of 25 social media posts about Religion, Science, and Sports on a scale from zero (oppose) to one (support). Variance is low when ranking opinions with strong language supporting a topic, but both humans and GPT-4 have some difficulty with negative language due to some posts with sarcasm, which were included in the study. Posts with neutral opinions or without reference to a topic were nearly universally identifiable by both humans and GPT-4.

edge probabilities of zero, 0.001, and 0.01. At zero, the opinion of $[0,0,0]$ eventually becomes the furthest from any regular agent, and after applying the Hadamard power and normalizing before Equation 2, the weight of influence becomes effectively zero. When a minimum edge probability of 0.001 is used, the strategic agent can gain followers, adopt opinions similar to theirs, and retain edges, outperforming the stubborn agent. The stubborn agent's opinion remains starkly different and must rely solely on the minimum edge probability to facilitate influence. At an edge probability of 0.01, the stubborn agent is comparable to the strategic agent, though still at a slight disadvantage. Overall, the strategic agent has an advantage in moving the opinions of agents in the network towards its target when the probability is low but not zero due to its ability to adopt opinions similar to the standard agents in the network.

V. Generative AI Makes Implementation of Opinion Controllers Easy

In this section, we investigate how an automated agent can drive opinions in social networks, not as a how-to demonstration but to emphasize that now is the time to think of the ethics and real-life implications of opinion and social control research in control theory. We begin by first showing than an LLM can act as an opinion inference engine for social media posts, and then we show that generative AI can create content based on a provided opinion vector. The feedback control relationship can be seen in Figure 1.

A. Opinion Inference Engine

For the first experiment, we gathered from Facebook, Twitter, and Reddit 25 social media posts on three topics

Memes on Taking Action Against Climate Change

(a) Opinion Value 1.0 (Support) (b) Opinion Value 0.1 (Oppose)

Fig. 7: Memes generated by Dalle-3 (left) and Midjourney V6 (right). Notice slight errors in the text spelling and spacing, but, nevertheless, these demonstrate that there are multiple tools available to automate effective meme generation responding to a numerical scale on an issue.

of religion, science, and sports, and included posts that were not on any of those topics as a control. We had twelve human volunteers annotate the opinion vectors for each individual opinion topic. We then had six independent instances of GPT-4 [33] output its inference on the opinion vectors in $\mathbb{R}^3$. Figure 6 shows the variance for the human annotators (orange) and GPT-4 (blue). We see a general correlation between opinion assessments, but where the variance is low in posts with strong supportive language on a topic, there is some disagreement in strong negative language due to sarcasm (some sarcastic posts were included in the data). Neutral stances on topics and posts that were not on any of the three topics were nearly universally identifiable by both humans and GPT-4. After averaging the opinion rankings for the human and GPT-4 assessments, we find that the difference in opinions on average was 0.023 for religion, 0.177 for science, and 0.11 for sports. These results show that the ability of GPT-4 to assess opinion vectors of social media posts is fairly similar to human assessments, though not equal. The capability of GPT-4 to take in image data means these results may extend to interpretation of image content as well. For example, GPT-4 properly identified the opinion values of the images in Figure 7.

B. AI Generation from Opinion Vectors

To show that generative AI can create posts based on opinion vectors, we provided GPT-4 with the topic of "taking action on climate change" and had it create short posts following opinion values from 0.1 to 1.0. Table I shows the posts generated by Chat-GPT, revealing the subtle nuances in stances toward climate change action even between similar opinion values. Although current image generation AI do not have the full capability of simply taking in an opinion vector and topic, interpret it, and then generate a resulting image following that opinion, we can use LLMs like GPT-4 to generate an image prompt for image AI like Dalle-3 and Mdjourney V6. Figure 7 shows two examples of memes

TABLE I: Chat-GPT Generated Media Post Given Controller-Generated Opinion Value

Opinion Value	Social Media Bot Posts About Taking Action on Climate Change
1.0 (Support)	I simply cannot fathom why anyone would oppose the fact that climate change is real. Ignoring this issue is not an option. We need to come together as a global community to address this challenge head-on by reducing our carbon footprint, investing in renewable energy, and advocating for policies that protect our planet for future generations.
0.7	We have a responsibility to take care of our planet and ensure that it remains habitable for future generations. Despite the challenges we face, I'm optimistic about the steps we can take to address climate change. Renewable energy sources like wind and solar power are becoming more affordable and accessible every day, and innovative technologies are helping us to reduce our carbon footprint and conserve resources.
0.5 (Neutral)	Climate change is a complex and often controversial issue, with arguments for and against its existence and significance. On the one hand, there is a wealth of scientific evidence that suggests that human activities are a major contributing factor. There are those who dispute these findings and contend that efforts to address climate change may be too costly or difficult to implement, and that more research is needed before we can take action.
0.3	I also think that some of the proposed solutions to address climate change may be unrealistic or unfeasible, particularly for those who live in developing countries or who rely on certain industries for their livelihoods. While I believe in taking care of our planet, I think we need to balance this with practical considerations and not make hasty decisions that could have unintended consequences.
0.1 (Oppose)	I simply cannot stand the constant fearmongering and doom-and-gloom surrounding the topic of climate change. I think that the push for climate change solutions often overlooks the economic and social costs of these policies. It's not as simple as just "going green" — there are real trade-offs and sacrifices that would need to be made, and I don't think it's fair to expect everyday people to bear the brunt of these changes.

generate by Dalle-3 (left) for support and Midjourney V6 (right) for opposition against climate change action. With the vast improvements in video generation as seen in OpenAI's Sora [34], it will not be long before video content can be generated in the same manner.

VI. Ethical Considerations

Although previous work on opinion control may have considered the idea in the abstract, this paper demonstrates how easy it now is to implement such controllers on real social media networks using current generative AI technologies. This widespread access to systematic methods for mass manipulation highlights the urgent need for discussions on the morality of applying control methods to people and the development of corresponding ethical principles.

Certainly there are many uses for automated agents in social media networks that would seem to contribute to the common good, such as compensating for levels of homophily that may appear to be unhealthy, or making introductions between agents with common interests. Nevertheless, the same control methods that enable these capabilities can just as easily contribute to real social harm [35], [36], [37].

This isn't the first time the scientific community has conducted research or developed technologies with the potential for real social harm, however, so there are frameworks we can use for considering the ethical implications. For example, the Belmont Report [38] offered critical guidance for biomedical research in 1978, and the Menlo Report [39] offered similar principles for research on information and communications technologies in 2012:

1) Respect for Persons, including Informed Consent,
2) Beneficience,
3) Justice, and
4) Respect for Law and Public Interest, including Transparency

Kevin Macnish and Jeroen van der Ham have more recently modified the Menlo Report [40], with a special focus on cybersecurity research that includes a section looking beyond research activities to explore the ethics of cybersecurity development in industrial contexts. This application closely parallels questions surrounding the ethics of publishing effective opinion or social control techniques since such techniques are both research and, as demonstrated here, nearly immediately deployable in practice using widely available generative AI.

The Association for Computing Machinery (ACM), however, developed a framework in 2022 even more applicable to social control than cybersecurity in their "Statement on Principles for Responsible Algorithmic Systems" [41]:

1) Legitimacy and Competency,
2) Minimizing Harm,
3) Security and Privacy,
4) Transparency,
5) Interpretability and Explainability,
6) Maintainability,
7) Contestability and Auditability,
8) Accountability and Responsibility,
9) Limiting Environmental Impacts

These principles correspond strongly to those from [39] and refined in [40], but add new refinements such as interpretability/explainability and contestability/auditability.

Since then, various organizations have worked on further refinements in the context of ethical AI systems, resulting in guidelines for Trustworthy and Responsible AI from the National Institute of Standards and Technology (NIST) [42], that added fairness to its list of principles for ethical systems, and various other guidelines, of which the list from Intel is typical [43]. These criteria, however, continue to evolve and could benefit from input from the controls community.

VII. Conclusion

Automatic methods designed to change people's minds have reached a new level of expertise. This paper presented a new model of Social Media Networks along with algorithms for three opinion controllers. Results from a simulation study were then described, and details for using generative AI

to deploy such controllers were illustrated–including experimental results demonstrating the efficacy of using Large Language Models for quantifying the opinion characteristics of social media posts. Ethical concerns and frameworks from related fields were then presented with an invitation for more work understanding the morality of using automatic control methods for persuasion.

Organizations like the NSF and DARPA have long recognized the risks of powerful techniques developed for feedback control, including jobs lost to automation [44] and, specifically, the need for cognitive security to protect populations from influence operations and mass manipulation [45]. Future work could explore how to accomplish these goals.

REFERENCES

[1] C. Rapp and Aristotle, "Aristotle's rhetoric i.2," *Stanford Encyclopedia of Philosophy (Winter 2023 Edition)*, 2023, accessed: 3/18/2024. [Online]. Available: https://plato.stanford.edu/entries/aristotle-rhetoric/

[2] J. L. Moreno, *Who Shall Survive? A New Approach to the Problem of Human Interrelations*. Washington, D.C: Nervous and Mental Disease Publishing Co., 1934.

[3] J. Scott, "Trend report social network analysis," *Sociology*, pp. 109–127, 1988.

[4] S. Wasserman and K. Faust, *Social network analysis: Methods and applications*. Cambridge university press, 1994.

[5] S. H. Strogatz, "Exploring complex networks," *Nature*, vol. 410, no. 6825, pp. 268–276, 2001.

[6] N. Wiener, *Cybernetics or Control and Communication in the Animal and the Machine*. MIT press, 2019.

[7] W. Buckley, *Sociology and modern systems theory.* Prentice-Hall, 1967.

[8] A. V. Proskurnikov and R. Tempo, "A tutorial on modeling and analysis of dynamic social networks. Part I," *Annual Reviews in Control*, vol. 43, pp. 65–79, 2017.

[9] ——, "A tutorial on modeling and analysis of dynamic social networks. Part II," *Annual Reviews in Control*, vol. 45, pp. 166–190, 2018.

[10] M. O. Jackson and L. Yariv, "Diffusion on social networks," *Economie Publique (Public Economics)*, vol. 16, no. 1, pp. 3–16, 2005.

[11] C. Lagnier, L. Denoyer, E. Gaussier, and P. Gallinari, "Predicting information diffusion in social networks using content and user's profiles," *HAL Open Science*, vol. 7814, pp. 74–85, 03 2013.

[12] Y. Jiang and J. Jiang, "Diffusion in social networks: A multiagent perspective," *"IEEE Trans. Syst., Man, Cybern. A"*, vol. 45, no. 2, pp. 198–213, 2014.

[13] Y. Wang, A. V. Vasilakos, J. Ma, and N. Xiong, "On studying the impact of uncertainty on behavior diffusion in social networks," *"IEEE Trans. Syst., Man, Cybern. A"*, vol. 45, no. 2, pp. 185–197, 2014.

[14] G. D'Agostino, F. D'Antonio, A. De Nicola, and S. Tucci, "Interests diffusion in social networks," *Physica A: Statistical Mechanics and its Applications*, vol. 436, pp. 443–461, 2015.

[15] W. Xuan, R. Ren, P. E. Paré, M. Ye, S. Ruf, and J. Liu, "On a network sis model with opinion dynamics," *IFAC-PapersOnLine*, vol. 53, no. 2, pp. 2582–2587, 2020.

[16] B. She, J. Liu, S. Sundaram, and P. E. Paré, "On a networked sis epidemic model with cooperative and antagonistic opinion dynamics," *IEEE Trans. Cont. Network Syst.*, vol. 9, no. 3, pp. 1154–1165, 2022.

[17] M. Taylor, "Towards a mathematical theory of influence and attitude change," *Human Relations*, vol. 21, no. 2, pp. 121–139, 1968.

[18] M. H. DeGroot, "Reaching a consensus," *Journal of the American Statistical association*, vol. 69, no. 345, pp. 118–121, 1974.

[19] N. E. Friedkin and E. C. Johnsen, "Social influence and opinions," *Journal of Mathematical Sociology*, vol. 15, no. 3-4, pp. 193–206, 1990.

[20] G. Deffuant, D. Neau, F. Amblard, and G. Weisbuch, "Mixing beliefs among interacting agents," *Advances in Complex Systems*, vol. 3, no. 01n04, pp. 87–98, 2000.

[21] R. Hegselmann, U. Krause, *et al.*, "Opinion dynamics and bounded confidence models, analysis, and simulation," *Journal of artificial societies and social simulation*, vol. 5, no. 3, 2002.

[22] R. Olfati-Saber and R. M. Murray, "Consensus problems in networks of agents with switching topology and time-delays," *IEEE Trans. Automat. Contr.*, vol. 49, no. 9, pp. 1520–1533, 2004.

[23] R. Olfati-Saber, "Flocking for multi-agent dynamic systems: Algorithms and theory," *IEEE Trans. Automat. Contr.*, vol. 51, no. 3, pp. 401–420, 2006.

[24] F. Cucker and S. Smale, "Emergent behavior in flocks," *IEEE Trans. Automat. Contr.*, vol. 52, no. 5, pp. 852–862, 2007.

[25] H. G. Tanner, A. Jadbabaie, and G. J. Pappas, "Flocking in fixed and switching networks," *IEEE Trans. Automat. Contr.*, vol. 52, no. 5, pp. 863–868, 2007.

[26] R. Olfati-Saber, J. A. Fax, and R. M. Murray, "Consensus and cooperation in networked multi-agent systems," *Proceedings of the IEEE*, vol. 95, no. 1, pp. 215–233, 2007.

[27] E. Yildiz, A. Ozdaglar, D. Acemoglu, A. Saberi, and A. Scaglione, "Binary opinion dynamics with stubborn agents," *ACM Transactions on Economics and Computation (TEAC)*, vol. 1, no. 4, pp. 1–30, 2013.

[28] R. Bredereck and E. Elkind, "Manipulating opinion diffusion in social networks," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 2017, pp. 894–900. [Online]. Available: https://doi.org/10.24963/ijcai.2017/124

[29] H. Hu, "Competing opinion diffusion on social networks," *Royal Society Open Science*, vol. 4, no. 11, p. 171160, 2017.

[30] D. S. Hunter and T. Zaman, "Optimizing opinions with stubborn agents under time-varying dynamics," *arXiv preprint arXiv:1806.11253*, 2018.

[31] H. Z. Brooks and M. A. Porter, "A model for the influence of media on the ideology of content in online social networks," *Physical Review Research*, vol. 2, no. 2, pp. 023 041–1–023 041–20, 2020.

[32] N. Wendt, C. Dhal, and S. Roy, "Control of network opinion dynamics by a selfish agent with limited visibility," *IFAC-PapersOnLine*, vol. 52, no. 3, pp. 37–42, 2019.

[33] OpenAI, "Gpt-4 technical report," *ArXiv*, vol. abs/2303.08774, 2023. [Online]. Available: https://arxiv.org/abs/2303.08774

[34] Y. Liu, K. Zhang, Y. Li, Z. Yan, C. Gao, R. Chen, Z. Yuan, Y. Huang, H. Sun, J. Gao, *et al.*, "Sora: A review on background, technology, limitations, and opportunities of large vision models," *arXiv preprint arXiv:2402.17177*, 2024.

[35] M. DeBuse and S. Warnick, "Automatic control of opinion dynamics in social networks," *7th IEEE Conference on Control Technology and Applications (CCTA)*, 2023.

[36] M. Mosleh and D. G. Rand, "Measuring exposure to misinformation from political elites on twitter," *Nature Communications*, vol. 13, no. 1, p. 7144, 2022.

[37] A. Bovet and H. A. Makse, "Influence of fake news in twitter during the 2016 us presidential election," *Nature Communications*, vol. 10, no. 1, p. 7, 2019.

[38] United States. National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research, *The Belmont report: ethical principles and guidelines for the protection of human subjects of research*. Department of Health, Education, and Welfare, 1978, vol. 2.

[39] M. Bailey, D. Dittrich, E. Kenneally, and D. Maughan, "The menlo report," *IEEE Security & Privacy*, vol. 10, no. 2, pp. 71–75, 2012.

[40] K. Macnish and J. Van der Ham, "Ethics in cybersecurity research and practice," *Technology in Society*, vol. 63, p. 101382, 2020.

[41] US Public Policy Council, "Statement on algorithmic transparency and accountability," *Commun. ACM*, 2022.

[42] National Institute of Standards and Technology (NIST), "Trustworthy and responsible AI," accessed: 3/18/2024. [Online]. Available: https://www.nist.gov/trustworthy-and-responsible-ai

[43] INTEL, "Artificial intelligence ethics framework for the intelligence community," 2020, accessed: 3/18/2024. [Online]. Available: https://www.intelligence.gov/artificial-intelligence-ethics-framework-for-the-intelligence-community

[44] National Science Foundation, "Future of work at the human-technology frontier: Core research (FW-HTF)," National Science Foundation, Arlington, VA, NSF Solicitation 23-543, 2023, available at URL https://new.nsf.gov/funding/opportunities/future-work-human-technology-frontier-core/nsf23-543/solicitation.

[45] R. Waltzman, "The weaponization of information," *RAND*, vol. 10, pp. 11–18, 2017. [Online]. Available: https://www.rand.org/content/dam/rand/pubs/testimonies/CT400/CT473/RAND_CT473.pdf