The Foundations of Cyber Social Agents

Lynnette Hui Xian Ng

CMU-S3D-26-101

May 2026

Software and Societal Systems Department
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213

Thesis Committee:
Kathleen M. Carley, Chair
L. Richard Carley
Nicolas Christin
Melissa Chua (Defense Science and Technology Agency)

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Societal Computing.

This work is supported by the John S and James L Knight Foundation (G201958790), Office of Naval Research (Minerva-Multi-Level Models of Covert Online Information Campaigns, N000142112765; MURI: FACTIONS: Near Real Time Assessment of Emergent Complex Systems of Confederates, N000141712675; Threat Assessment Techniques for Digital Data, N000142412414), Defense Science and Technology Agency (RPS8DST000EPO21000426) the CMU Graduate small Project Help (GuSH), the Center for Computational Analysis of Social and Organizational Systems (CASOS), and the Center for Informed Democracy and Social-cybersecurity (IDeaS). The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Knight Foundation, Office of Naval Research, the US Army, or the U.S. Government.

Keywords: cyber social agents, social media bots, natural language processing, network analysis, social psychology

To my Bot Buddies, for their constant companionship and motivation.

Abstract

Social media bots are often treated as a homogeneous and inherently malicious presence in online environments. This thesis argues that these automated accounts actually function as as heterogeneous Cyber Social Agents (CSAs) embedded within socio-technical systems. Their behaviors, roles and impacts emerge through the interactions with other users, the content and the platform algorithms. Drawing on multi-platform datasets that span political events, crises and international discourse, this thesis characterizes a typology of CSAs and demonstrates that CSAs are deployed strategically to influence narratives, disseminate information, and amplify social signals. Contrary to common assumptions, bots and humans exhibit similar levels of moral and emotional language, and their distinguishing characteristics arise primarily through coordination patterns, network positioning, and the systematic invocation of cognitive biases. Using network analysis and coordination metrics, this thesis also shows that bot influence emerges collectively through synchronized activity rather than individual behavior, and by working together, these agents accelerate narrative spread and shape engagement and stance dynamics. Agent-based simulations further demonstrate that automation is not inherently harmful: purpose-designed "useful bots" can mitigate false information dynamics and improve information resilience. To understand the information ecology, this thesis develops computational frameworks for bot detection, archetype-based classification, and large-scale social simulation. By integrating computational methods with sociological theory, this thesis reframes bots as active social actors and provides a unified framework for studying automation, influence, and intervention in modern information ecosystems.

Acknowledgments

To my family: Thank you for rooting for me all my life. Thank you for standing by me during my happiest times, my hard times, my lifeless times, and my alive times.
To my husband: Thank you for riding this roller coaster with me.
To my Pittsburgh family: Thank you for giving me a home away from home, and loving me as a daughter.

To my advisor, Dr. Kathleen Carley: Thank you for believing in my research capabilities before I even did myself.
To my committee members, Dr. Nicolas Christin, Dr. Richard Carley, Dr. Melissa Chua, thank you for pushing me into the uncomfortable depths of knowledge so I can discover my limits.
To all the other professors that crossed my path: Thank you for paving my path, guiding me and giving me advice. I will join your ranks very soon.
To my research mates: Thank you for walking this journey with me and sharing my joys and frustrations.

To my violin teacher Leah Givelber, Maestro Jeffery Klefstad, CMU All University Orchestra and the Pittsburgh Philharmonic: Thank you for giving me the weekly space to laugh. Thank you for letting me study and perform music with you.

To the doctors and nurses at University of Pittsburgh Medical Center, and the heart transplant donor and family: Thank you for giving me the breath of life to tell this story. I truly cannot be more grateful to you for giving me hope.

To all my friends: Thank you for following my story. Thank you for supporting my quest for intelligence. Thank you for donating to my Heart Transplant Drive. Thank you for lending me a listening ear. Thank you for being there.

Lastly, if you are reading this thesis, thank you for your support in my work.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

An autonomous agent that operates in digital spaces, in particular the social media space, is commonly referred to as a "social media bot", or a "bot" for short. Social media bots have been of great interest of the computational social science world because of their increasing population in our online ecosystem. Studies estimate that more than half of the traffic on the Internet is generated by bots [154], a quarter of the tweets on Twitter is created by bots [51], and two-thirds of the links posted on Twitter are generated by bots [310].

While the term "bot" often refers to a software program designed to automate social media actions, this thesis studies the construct through the lens of agency theory and socio-technical systems, rather than an inert piece of software or purely a technical artifact. Here, the bot is conceptualized as a **Cyber Social Agent** (CSA). A CSA is a digital actor embedded within a social network environment, with the capacity to perceive, process and act upon information in ways that influence the narratives and other actors in the ecosystem. CSAs interact dynamically with the users, content and algorithms of social media platforms. CSAs are not dead code, but are active participants in social-technical systems whose actions can affect the social network and narratives. They actively shape and drive our social environment, which makes them a compelling object of study and a powerful mechanism that can be harnessed for social good.

A social media platform has three main elements: the users, the contents and the relationships [212]. The social media platform is a virtual venue that hosts people to connect with each other, and each platform provides a certain set of affordances to facilitate communication. Users are represented by their virtual accounts. Users create, distribute and consume information through their online conversations. Content is the information that users generate on the platform, which manifests as posts, comments or other forms. Relationships are interactions formed between user-to-user, user-to-content and content-to-content. These three elements can be controlled by the users themselves as they use the social media application. There is a fourth element, the algorithm, which is controlled by the platform, and determines the content or user that is recommended to the users. Users can design their content to take advantage of these algorithms, which can in turn affect their interactions. This User-Content-Interaction frame is useful to analyze how social media platforms encourage user participation, mediate content visibility and shape the interaction dynamics between users and user-content.

Cyber Social Agents are socio-technical entities that operate within these four elements. They are programmed to enable them to strategically harness the platform's dynamics and affordances

– for better or for worse. First, CSAs are automated users that are akin to social media bots but with agency and evolution. They have different archetypes, each with unique behavior and rhetorical characteristics. Their content have their own semantic style, narrative expressions that are different from humans, and reflects their motivations such as establishing trust [150] or provoking conflict [313] or generating content at scale [137]. CSAs build relationships through extended star network typology and coordinated networks [217, 219], to influence the social media discourse and manipulate opinions [59, 308]. Finally, CSAs harness the social media algorithms by optimizing the posting times, hashtags, interaction strategies, flooding the zone with incessant postings to game the platform's recommendation systems to maximize visibility of their narratives [20, 57, 200, 313].

1.1 Fusing Computational Methodologies and Sociological Theories

The study of CSAs fuses this sociological backdrop into a necessary computational orientation. It is insufficient to study these agents purely through a computational lens, because that reduces them to purely algorithmic artifacts. At the same time, a pure sociological perspective will not be able to precisely handle the analysis of the operation and impact of CSAs with the large-scale, high-velocity data that characterize social media platforms. To profile a more well-rounded view of CSAs, this thesis blends sociological theories with computational methods. Computational methods enable the systematic detection, measurement and modeling of agent behavior at scale, while sociological theories provide the interpretive frameworks to understand the behavior and effect of CSAs. This interdisciplinary framing positions CSAs as both objects of social inquiry and computational phenomena, bridging the computer science and sociology fields. Table 1.1 lists the sociological theories and computational methods that are integrated in this thesis, demonstrating how theories from social psychology to network science are operationalized through methods like machine learning, statistical analysis and ABM simulation.

1.2 Organization of This Thesis

This thesis is organized as such: chapter 2 ("Bot Detection") examines the general nature of an automated account and describes machine learning algorithms that can differentiate such accounts from humans. chapter 3 ("From Bots to Cyber Social Agents") introduces the idea of Cyber Social Agents as nuanced set of archetypes of automated accounts rather than a monolithic category of inorganic users. This chapter presents the different archetypes a Cyber Social Agent (CSA) can take on, based on their behavior and content, and emphasizes that the same archetype can be used for both good and for bad. chapter 4 ("Nature of Cyber Social Agents") describes the general nature of a CSA through their social political representation, narrative expressions, motivations & agencies, linguistic signatures and cognitive bias triggers. chapter 5 ("Network Interactions and Coordination Profiles of Cyber Social Agents") uses a network science approach to describe the unique interaction profiles of CSAs, the patterns of synchronization & coordination and the network impacts of CSAs. chapter 6 ("Social Simulations of CSAs & Humans")

presents methodologies of modeling CSAs to generate synthetic data for research purposes, and presents a simulation of how useful CSAs can prevent a conspiratorial society from forming. Finally, chapter 8 (“Conclusions”) concludes this study of Cyber Social Agents, and chapter 9 (“Contributions and Future Directions”) presents the contributions of the thesis and directions of study that the thesis opens.

	Title	Concepts of CSAs	Sociological Theories	Computational Methods
2	**Bot Detection**	What is a Bot? Bot Detection Algorithms		Machine learning (ensemble learning, mixture-of-experts), Deep learning, Natural language processing (sentence vectorization and classification), Statistical Analysis, Word entropy
3	**From Bots to CSAs**	Bot Personas	Role theory	Statistical analysis, Large Language Models, Network Science (random graphs
		Good & Bad of Bots	Technology Ambivalence	Statistical analysis, Natural language processing, Large Language Models
4	**Nature of Cyber Social Agents**	Narrative expressions	Affect theory, framing theory, persuasion theory	Natural language processing (topic analysis, emotion analysis), Computer vision (image analysis), Big data processing, Distributed and parallel processing
		Motivations & Agencies	Social Influence Theory, Strategic Communication Theory	Natural language processing (topic analysis, narrative analysis), Network science
		Social & Political Representation	Digital diaspora theory, Digital diplomacy	Machine learning (multilingual, transformers), Big data processing, Network Science
		Semantic Style	Communication accommodation theory, linguistic style matching	Natural language processing, semantic analysis
5	**Network Interactions & Coordination Profiles of CSAs**	Network interaction profiles	Network theory	Network science, Machine learning (ensemble learning), Statistical analysis
		Synchronization & Coordination	Network theory	Network science, Statistical analysis, Temporal analysis
		Network impacts	Social influence theory	Machine learning, Statistical analysis, Temporal analysis
		Content of interaction	Homophily bias, Authority bias, Affect bias, Negativity bias, Illusory truth effect, Availability bias, Cognitive dissonance, Confirmation bias	Network science, Statistical analysis (Regression analysis, Correlation analysis)
6	**Social Simulations of CSAs & Humans**	Modeling of CSAs and Humans		Statistical analysis, Large Language Models, Network science, Graph isomorphism
		Simulation of useful CSAs	Conspiracy theories, confirmation bias, motivated reasoning, monological belief system, proactive inoculation	Agent-Based Modeling

Table 1.1: Summary of Sociological Theories and Computational Methods used in this Thesis

1.3 Data Used in This Thesis

We used a variety of data in this thesis. Table 1.2 summarizes the data we used, and the corresponding chapters they were used in. In the following paragraphs, we elaborate on the data.

Data	Properties	Chapters Used
X (previously Twitter)		
OSOME Bot Dataset*	Users: 86k, Posts: 3.4mil	Ch 2: Bot Detection
2018 Black Panther Movie*	Users: 1.6mil, Posts: 17.7mil	Ch 2: Bot Detection Ch 5: Network Interactions & Coordination Profiles
Asian Elections*	Users: 951k, Posts: 4.1mil	Ch 2: Bot Detection Ch 5: Network Interaction Profiles Ch 3: From Bots to Cyber Social Ch 4: Nature of Cyber Social Agents Ch 5: Network Interaction & Coordination Profiles
2019 Canadian Elections*	Users: 1.9mil, Posts: 18mil	Ch 2: Bot Detection
2019-2020 US Elections*	Users: 1.6mil, Posts: 55mil	Ch 2: Bot Detection Ch 5: Network Interactions & Coordination Profiles
2020-2021 Coronavirus*#	Users: 208mil, Posts: 4.2mil	Ch 2: Bot Detection Ch 3: From Bots to Cyber Social Agents Ch 4: Nature of Cyber Social Agents Ch 5: Network Interactions & Coordination Profiles Ch 6: Social Simulation
2020 ReOpen America*	Users: 201k, Posts: 4.4mil	Ch 2: Bot Detection Ch 3: From Bots to Cyber Social Ch 5: Network Interaction & Coordination Profiles
2021 Indonesian discourse	Users: 20k , Posts: 700k	Ch 3: From Bot to Cyber Social Agents Ch 5: Network Interaction & Coordination Profiles

Data	Properties	Chapters Used
2021/ 2023 French Protests*#	Users: 343k, Posts: 644k	Ch 2: Bot Detection Ch 5: Network Interactions & Coordination Profiles
2021 Capitol Riots	Users: 290k, Posts: 1.7mil	Ch 5: Network Interactions & Coordination Profiles
2023 Chinese Balloon#	Users: 121k, Posts: 1.2mil	Ch 3: From Bots to Social Cyber Agents Ch 4: Nature of Cyber Social Agents
2023 Russia-Ukraine#	Users: 11mil, Posts: 38mil	Ch 4: Nature of Cyber Social Agents Ch 7: Case Study of 2023 Russia-Ukraine Conflict
Reddit		
2022 Reddit#	Users: 667, Posts: 13k	Ch 2: Bot Detection
Instagram		
2022 Instagram#	Users: 1935	Ch 2: Bot Detection
Telegram		
2021 COVID#	Users: 335,088, Posts: 7,711,975, Channels: 10,633	Ch 2: Bot Detection Ch 4: Nature of Cyber Social Agents
Parler		
2021 Parler#	Users: 290,000, Parleys: 1.7 million	Ch 5: Network Interaction & Coordination Profiles

Table 1.2: Table of Data used in this thesis. * indicates the dataset was collected from a public repository. # indicates the dataset was self-collected.

1.3.1 X datasets

OSOME Bot Dataset (hybrid collection) is a series of datasets hosted on `https://botometer.osome.iu.edu/bot-repository/datasets.html`, which consists of expert annotated data of bot and human accounts in domains like political, entertainment and financial bots. Due to Twitter's Terms-Of-Service, only the account ID was shared on the OSOME website. To form the complete dataset (i.e., user ID, user tweets, user metadata), we rehydrated the datasets in June 2021, collecting 40 tweets per account using the Twitter V1 API for data collection. We chose 40 tweets by referencing a prior study performed a systematic analysis on the stability of bot classification showed that 40 tweets is a reasonable collection size for a consistent bot probability score [218].

Asian Elections (obtained from repository) follows the elections in Philippines, Indonesia, Taiwan and Singapore that occurred during 2019 and 2020 [292, 294].

2018 Black Panther Movie (obtained from repository) was Marvel Studio's first superhero film with a strong female lead. This dataset follows the online discussion surrounding gender diversity and misinformation about views from the actors [23].

2019 Canadian Elections (obtained from repository) took place on 21 October 2019. The Liberal party won the vote and Justin Trudeau become the Prime Minister. The dataset follows around six months of online campaigning and discussion about the election [146].

2020 US Elections (obtained from repository) dataset follows the United States elections from the Primaries to the aftermath of the voting [206]. The Democratic party won the election and Joe Biden was named the 46th President of the United States.

2020-2021 Coronavirus (hybrid collection) is a collection of tweets that stemmed from the coronavirus pandemic during 2020-2021. This dataset follows one year of discourse on the health pandemic. Most of the dataset is contained obtained from the CASOS lab's central data repository, but we had collected some portions related to conspiracy theories and the vaccine to supplement it.

2020 ReOpen America (obtained from repository) protests were launched across the United States against the government lockdown response to the coronavirus pandemic. The dataset follows three months of Twitter discourse during the heightened protests emotions [175, 206].

2021 Indonesian discourse (self-collected) dataset consists of two parts. (1) the discourse written by Indonesian accounts on the Palestine-Israel conflict, collected from 14-20 May 2021 with the search term "Palestina", the Indonesian word for Palestine, and are filtered for tweets in the Indonesian language. (2) discourse written by Indonesian accounts on the issuance of a decree about alcoholic beverages, collected from 26 Feb to 3 May 2021 with the hashtags #BatalkanPerpresMiras, #PapuaTolakInvestasiMiras and #MirasPangkalSejutaMaksiat [70].

2021/2023 French Protests (hybrid collection) dataset followed the protests in 2020 to 2021 that revolved around the vow from French President Emmanuel Macron to protect the right to caricature the Islamic prophet Muhammad as a cartoon, and the protests in 2023 revolving around the pension reformed signed by the French President Macron. The 2021 dataset was obtained from a repository , while the 2023 dataset was self-collected [209].

2023 Chinese Balloon (self-collected) dataset followed the online conversations on Twitter about the Chinese balloon spotted over the US airspace in January 2023. The US announced that it was a surveillance balloon while China maintained it was a weather balloon. This dataset was collected using the X Streaming API with the hashtags #chineseballoon and #weatherballoon from 31 January 2023 to 22 February 2022 [207].

2023 Russia-Ukraine (self-collected) dataset collected the conversations revolving the Russia-Ukraine conflict. On 24 February 2022 on the premise of a "special military operation", which has since escalated into a war that is still ongoing till today (2026). This dataset spans April to June 2023 and covers the beginning of the Ukraine counteroffensive. It was collected with the X Streaming API with the following keywords: "Russian invasion", "Russian military", "invasion of Ukraine". For this thesis, we only analyzed posts written in the English language [179].

1.3.2 Reddit Dataset (self-collected)

The Reddit dataset was curated in 2022. For bot accounts, we downloaded the 500 highest ranked "bad bot" in BotRank [1], a crowdsourced list of bot ranking [288]. For the humans, we collected users from 5 subreddits that generally require conscious writing and manually verified that the users are likely to he humans. We use the PushShift API [28] to collect data for this dataset.

1.3.3 Instagram Dataset (self-collected)

The Instagram dataset was curated in 2022 through a manual collection, from an observation of a group of accounts that followed a particular Instagram account within a few hours of the same day [205].

1.3.4 Telegram Dataset (self-collected)

The Telegram dataset was curated in 2022 through a snowball sampling collection that stemmed from the "Disinformation Dozen", a group of twelve people that the Center for Countering Digital Hate had identified to most responsible for spreading disinformation about the coronavirus [52]. Only eight of the Dozen had active accounts on Telegram at the time of this thesis. We collected both the channel content originating from the Disinformation Dozen and the associated discussion by other users. We then computationally identified forwarded content in channel posts and discussions and identified which channels/users messages were forwarded from, forming our 1-hop channel list. We then collected channel messages from this 1-hop channel list. Using a 1-hop snowball sampling method provided a way of characterizing the dissemination of disinformation that stemmed from the original seed users – the Disinformation Dozen – as it identified users that disseminated messages through forwarding and collected information from the channels that received information second-hand. This technique allowed us to discover a large number of channels and users that were previously hidden from view. Figure 1.1 illustrates the snowball collection process.

1.3.5 Parler Dataset (self-collected)

The Parler dataset were parleys (i.e. posts on parler) surrounding the discourse of the 2021 Capitol Riot event. This dataset consists of a partial HTML scrape of Parleys posted shortly after the Capitol Riots when Internet users sought to preserve data from the Parler social network

[1] `http://botrank.pastimes.eu`

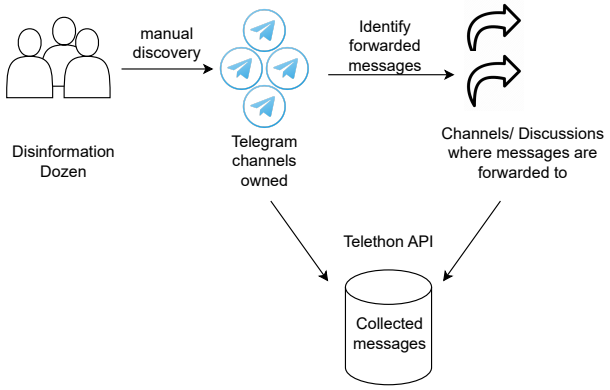


Figure 1.1: Data Collection pipeline for Telegram data (published in [221])

because Amazon Web Services banned Parler from being hosted on its service [170]. In total, the dataset consists of 1.7 million posts from 290,000 unique users between January 3rd to January 10th of 2021.

Chapter 2

Bot Detection

2.1 Introduction

The study of Cyber Social Agents on social media begins with an examination of the general nature of the bot. A social media bot is broadly understood to be an automated agent, but such a generic description obscures the diversity of roles that automation can take. To characterize Cyber Social Agents, it is first necessary to establish a foundational understanding on what a bot is and how they express themselves in digital environments. Studying the nature of bots provides the conceptual and empirical baseline to develop the broader framework of Cyber Social Agents, where these automated actors embody particular agencies to shape narratives.

This chapter investigates the following guiding **research questions**:

1. What is a bot?
2. How do we build efficient bot detection algorithms that can accurately differentiate between bot and human?
3. What user, post, user metadata features are important in differentiating bot and human?
4. What value ϵ should we use as a bot probability threshold for $P(bot)$ for stable and consistent bot detection results?

2.2 Related Work

2.2.1 What is a Bot

The term "bot" has become a pervasive metaphor for inauthentic and automated online users, and many social media users and researchers have an understanding of a bot. Many existing definitions define bots through their malicious activity, for example: "public opinion manipulation" [20], "impersonate real users" [319], "malicious automated agents" [32]. Such framing captures only one aspect of the risk of bot activity, but conflates the technology of automation itself with its harmful applications. At their core, bots are automated systems, and the same automation that enables manipulation can also be harnessed for positive purposes [129, 151, 245]. Our work consolidates definitions that have been written in both academic and industry contexts, then har-

monizes them into a first-principles definition of a bot that describes the behavior of a bot with accordance to the features of social media platforms.

2.2.2 Bot Detection Architectures

Bot detection architectures typically have three steps: a feature extraction step, the model step, and a bot probability evaluation step.

Feature Extraction. The feature extraction step formulates a feature vector that represents a user account. These features range from user meta-data features (i.e., location, presence of profile pictures) to content-based features (i.e., linguistic structure, syntactic structure), to interaction features (i.e., types of accounts the account follows) [6]. 37% of the bot detection algorithms surveyed between 2018 and 2021 emphasized the use of user meta-data, 31% used content features and 32% used a mixture of meta-data and content features [296]. This feature extraction step can get quite extensive, with the Botometer algorithm using over 1,000 extracted and derived features [322], and the BotHunter algorithm using a tiered approach as it increases the number of features used [32]. However, building a bot detection algorithm to analyze millions of users [212] requires streamlining this feature extraction step. Our work complements these algorithms by analyzing which features are actually important in the bot detection task.

Detection Model. There are several methods in which the model step can be constructed: through temporally-based detection from user behavioral sequences [58, 182, 315], supervised machine learning and [32, 71] deep learning methods from matching feature vectors of manually annotated bot and human user data [122], Graph Neural Networks that incorporate social interaction structure and relational dependencies [86, 87, 169, 173], and most recently Large Language Models (LLMs) that leverages on a large pool of contextual knowledge [44, 88].

Probability Evaluation. The final step of a bot detection algorithm is the bot probability evaluation step. The model in the previous step typically returns a bot-likelihood score $P(bot) \in [0, 1]$ that represents the probability that the user is a bot. The closer the probability is to 1, the more likely the user is a bot. Then, a threshold ϵ is usually arbitrarily defined, by which if $P(bot) \geq \epsilon$, the user is considered a bot, and if $P(bot) \geq \epsilon$, the user is considered a human. However, the value of ϵ differs across studies. In fact, even studies that use the same bot detection algorithm can use different values of ϵ. For the BotHunter algorithm, Beskow and Carley [34] used $\epsilon = 0.5$ to study bots originating from the Russian Internet Agency during the 2016 United States elections, Uyheng and Carley [293] used $\epsilon = 0.8$ to analyze the spread of online hate by bots, and King et al. [146] used $\epsilon = 0.6$ to $\epsilon = 0.8$ to analyze the presence of bots spreading false accusations during the 2019 Canadian elections. Such inconsistent threshold assignment can impact the interpretability of bot detection results, as such, our work introduces a methodology to determine a stable threshold for bot detection algorithms.

Many of these bot detection architectures can achieve high accuracy in a detection task but have a few limitations. First, these algorithms are mostly specialized to differentiate bots vs humans for a single platform only. Majority of these models are built for the social media platform X [32, 71, 86, 87], with some algorithms targeted for the Reddit platform [134, 171, 317]. Second, most of these algorithms are built to analyze posts that are written in the English language, mostly due to the availability of bot/human social media data annotated by human ex-

perts from websites like the OSOME Bot repository[1], and the ease of collecting data through the now-deprecated Twitter API. However, focusing on a single social media platform and language creates vulnerabilities: analysts are unable to examine the extent of automated activity in a variety of platforms, and across languages. Third, many comercially available bot detection algorithms require a live pull of the user's social media data (i.e., Botometer [322]), or extensive data collection (i.e., BotHunter [32]), or only has a one-time huge update (i.e., DeBot [58]). Such activities may not be possible when resources are limited (i.e., time allocated for collection, money allocated for purchase of paid APIs, compute resources or GPUs available), or when the user is no longer available (i.e., suspended, deleted). Therefore, our work seeks to create bot detection algorithms that can (1) work on historical data, (2) accommodate multiple social media platforms, (3) analyze multiple languages.

2.3 What Is a Bot?

This diverse perceptions of a bot necessitates a definition that isolates the essence of a bot – its automation. To move beyond narrow, activity-based framings, this thesis consolidated industry perspectives (Table 2.2) and academic definitions (Table 2.1), then identified their components. From the commonalities of the set of reviewed definitions (Table 2.3), I formulated a first-principles definition of a bot, based on the user-content-interaction elements of a social media platform. This definition is: "An automated account that carries out a series of mechanics on social media platforms, for content creation, distribution and collection, and/or for relationship formation and dissolutions" [212]. This creates a User-Content-Interaction frame towards describing the elements of social media agents. These three elements can be mostly controlled and defined by users or operators.

A fourth element, the algorithm, is controlled by the social media platform, and determines which content or user is recommended to others, and the affordances by which the users can interact. Users can design their Content to take advantage of these Algorithms, which can in turn affect their interactions. When social media platforms change their Algorithms, the Content that gets created and the Users that post can dramatically change too.

This User-Content-Interaction-Algorithm frame is the basis of understanding the foundations of Cyber Social Agents throughout this thesis.

2.4 Bot Detection Algorithms

Bot detection is the baseline method to surface, classify, and measure the presence of automated, non-human actors online. This section describes the work done in developing bot detection algorithms that this thesis constructs.

Developing Multi-platform, Multi-lingual Bot Detection Algorithms We developed five bot detection algorithms to improve the state-of-the-art architectures. These models were built with the following considerations in mind: (1) interoperability across multiple social media platforms, (2) can operate on a variety of languages, (4) can operate on historical data and does not

[1] `https://botometer.osome.iu.edu/bot-repository/`

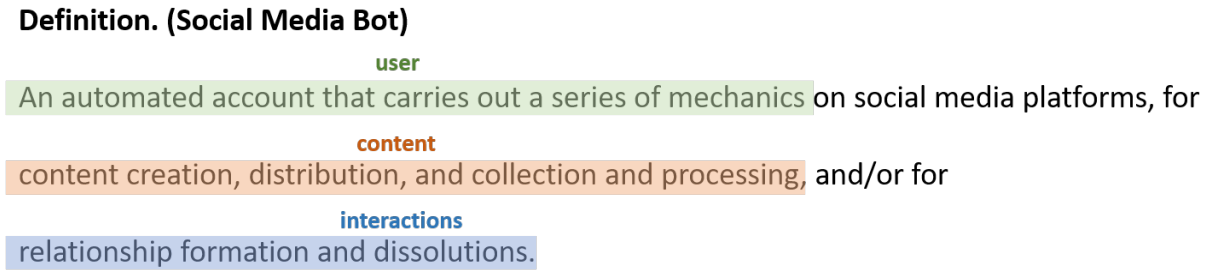


Figure 2.1: Definition of a Social Media Bot through a User-Content-Interaction frame. Published in [212]

require live data, (5) can work reasonably fast and accurately for analysis of millions of users, even on CPU architectures (i.e., does not require specialized GPU hardware to operate). Table 2.4 provides a summary of the comparison of the five bot detection algorithms developed. This summarizes the platforms that they are built to identify bots for, the model used, their accuracy metrics and their key features and limitations.

Mixture-Of-Experts architecture The bot detection algorithms we developed are all based on a Mixture-Of-Experts (MoE) architecture [327]. The MoE is an ensemble architecture based on the principle of divide-and-conquer, and has three main components: several "experts" which are specialized machine learning models trained on a subset of data, a gate that makes soft partitions of the entire dataset, and a probabilistic model to combine the experts. This architecture is commonly used in language translation [331], stance detection [117] and text generation [252]. Figure 2.2 represents the general implementation of the MoE architecture for the bot detection task. The social media data collected is split into six subsets: known information, user metadata, posts, username, screenname and description. An machine learning model is constructed separately for each expert, depending on what is most suitable for the data type of the subset.

Each of these experts will evaluate the probability the user is a bot ($P(bot)$) given the data subset it is specialized for, and the final bot-probability prediction is a weighted sum of the experts. The known information expert is first evaluated, and if there are platform-specific information that reveals the bot likelihood of the user, this information is directly used to determine the bot probability, and the other experts will not be activated. For example, if a user is verified, it will be directly marked as a human (i.e., $P(bot) = 0$).

We chose this architecture for our bot detection algorithms because (1) it allows individual experts to specialize in one subset of the data, and therefore different machine learning models can be used for each subset; and (2) it does not require collected data for all the features of the social media user or post, and deals with incomplete data by making a decent prediction based on given available data.

BotBuster BotBuster [205] is the first architecture we developed. BotBuster has five data pillars: user metadata expert, post expert, username expert, screenname expert and the description expert. Each expert is a deep-learning model that represents the input data pillar as a vector

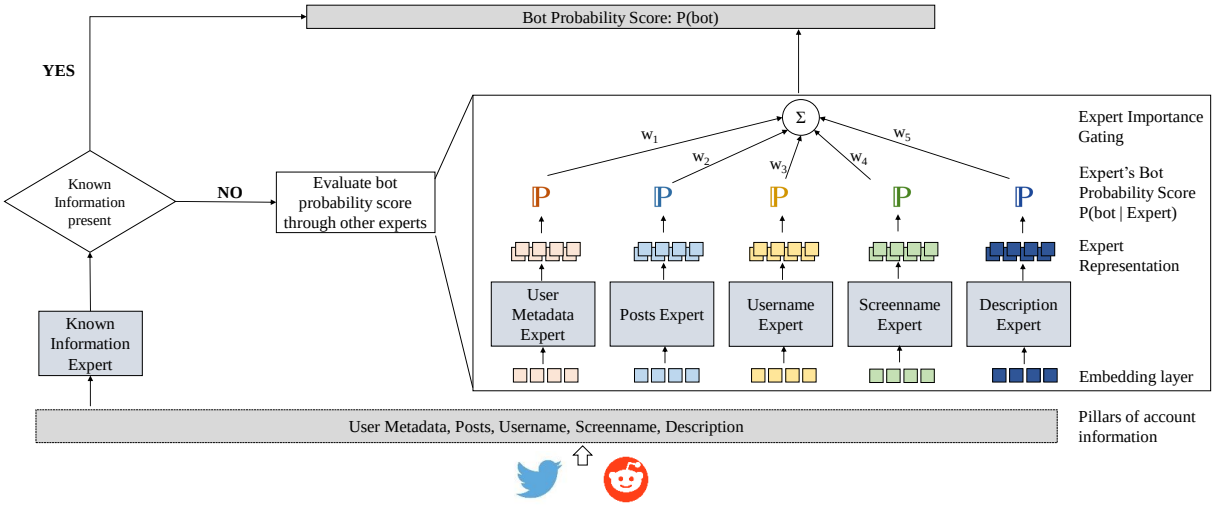


Figure 2.2: The BotBuster model architecture (Diagram published in [205])

representation and returns a bot probability score as determined by the specialized model trained for the expert.

The username, screenname, and description experts are modeled with a 768-dimension-vector of BERT transformer embeddings [74] which is then fed into a pre-trained BERT transformer models with a Multi-Layer Perceptron (MLP) classifier [246]. The user metadata expert takes in user metadata as a vector of normalized float values into a 4-layer MLP classifier with a dense layer. The post expert combines (1) post texts that are tokenized with the BERT transformer model [74] to obtain sequence embeddings, (2) post metadata (e.g., retweet count, like count) that are represented as a vector of normalized float values, and (3) derived post metadata, which are linguistic cues extracted from the post text (i.e., number of emojis, number of words, reading difficulty score), that are represented as a vector of normalized float values.

We trained and tested BotBuster on the following datasets from the OSOME bot repository: astroturf, botometer-feedback-2019, botwiki-2019, cresci-rtbust-2019, cresci-stock-2018, gilani-2017, midterm-2018, political-bots-2019, varol-2017, verified-2019; self-collected Reddit dataset; self-collected Instagram dataset and self-collected Telegram dataset. With the MoE architecture, BotBuster is able to process 100% of the users in a dataset by using the available data pillars the user has. This is a more superior performance compared to other bot detection algorithms. For the same 11 datasets, BotHunter [32] is able to process $55.04\% \pm 31.32$ of users, while Botometer [322] is able to process $51.61\% \pm 34.49$ of the users. In terms of accuracy, BotBuster performs with an average accuracy of $72.23\% \pm 18.33$, compared to $63.61\% \pm 28.12$ of the returned users from BotHunter and $76.19\% \pm 24.24$ of the returned users from Botometer.

To better understand the strengths and the weaknesses of the BotBuster algorithm, we compared that against the outputs of the BotBuster algorithm against the manual bot/human annotation of the datasets. For each X dataset, we extracted about $n = 2767$ to $n = 3000$ data points, making sure there were an equal proportion of bots and humans. Then we calculated the true negatives/ false positives/ false negatives/ true positives, which are presented in Table 2.5. This table shows that the algorithm is better at detecting humans (or not-bots), which lends to the

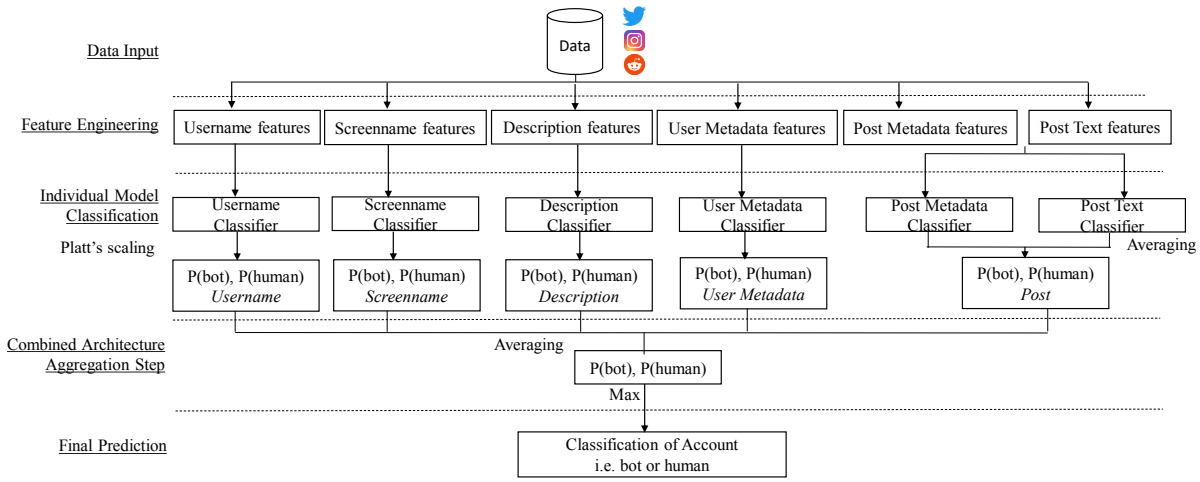


Figure 2.3: The BotBuster For Everyone model architecture (Diagram published in [210])

difficulty of a bot detection algorithm.

BotBuster For Everyone Many of the experts in the BotBuster architecture used transformer-based models, so the algorithm required a GPU to run smoothly. To make the model more accessible to analysts, we constructed BotBuster For Everyone [210], which is able to run on hardware with only a CPU. Figure 2.3 presents an illustration of the BotBuster For Everyone architecture. BotBuster For Everyone trains each expert using the following machine learning classifiers to each output a probability of (bot, human) tuple given the data input of the expert: decision tree for username, screenname and description, gradient boosting classifier for user metadata and random forests for posts. After training the models, their outputs are calibrated using Platt's scaling implemented using the Calibrated Classifier function in the sklearn library[2]. Using Platt's scaling calibrates the outputs of each classifier into a probability distribution using logistic regression, and makes the probability returned in the (bot, human) tuple of each of the classifiers comparable against each other [260].

Then, the probabilities from individual classifiers are combined in a Combined Aggregation step. The (bot, human) probabilities from each tuple is averaged out to produce a final (bot, human) probability. The final bot classification is determined by the larger of the values in the final (bot, human) tuple. In this fashion, the need for determining a suitable threshold to classify whether an account is a bot or a human is eliminated, reducing ambiguity of the classification.

We tested BotBuster For Everyone against the following datasets: from the OSOME repository we have botometer-feedback-2019, botwiki-2019, cresci-rtbust-2019, cresci-stock-2018, midterms-2018, political-bots-2019, verified-2019, and self-collected Reddit and Instagram datasets. In terms of overall accuracy, BotBuster For Everyone performs with an accuracy of $75.14\% \pm 18.78$. This is a superior accuracy compared to $34.62\% \pm 32.73$ of BotHunter and $31.42\% \pm 25.09$ of Botometer. BotBuster For Everyone is also able to process 100% of the users, compared with

[2] `https://scikit-learn.org/stable/modules/generated/sklearn.calibration.CalibratedClassifierCV.html`

of $42.06\% \pm 40.93$ of BotHunter and $45.52\% \pm 41.07$ of Botometer.

BotBuster-Telegram We adapted BotBuster For Everyone constructed BotBuster-Telegram [221]. BotBuster-Telegram was to identify bot-like users on the Telegram platform. This bot detection algorithm used our self-collected Telegram dataset.

First, we randomly sampled a subset of 3000 Telegram messages and 3000 Tweets. Both sets of the sample were collected during the same time period for the same event. With these datasets, we compared the messages, usernames and screen names, for these are the common fields between both platforms. Table 2.6 shows a statistics of comparison. We observed similarities in the average number of words and punctuation in messages, as well as the number of characters, numbers and capital letters in username and screen name. Further, there is no significant difference within the statistics at the $p < 0.05$ level when we perform a two-tailed t-test between the statistics derived from X and Telegram. The length of the messages and the style of the usernames and screen names are similar between the two mediums. Therefore, we can use the BotBuster For Everyone algorithm tuned for the platform X for our Telegram data. Since the Telegram platform does not have all the data fields that BotBuster For Everyone can leverage, being able to activate a subset of experts to derive a final bot prediction is ideal for our Telegram dataset.

	X	**Telegram**
Number of words in messages	10.09±32.96	18.35±33.39
Number of punctuation in messages	5.26±8.34	3.98±8.31
Number of characters in username	11.02±2.7	10.67±3.56
Numbers in username	1.11±2.16	1.01±1.49
Capitals in username	1.29±1.62	1.48±1.65
Numbers in screen name	0.12±0.59	0.13±0.66
Capitals in screen name	2.27±2.72	1.89±1.25
Number of words in screen name	2.07±1.19	1.77±0.75

Table 2.6: Statistics of messages, username, and screenname between X and Telegram. All values between X and Telegram are statistically insignificant at the $p < 0.05$ value by a two-tailed t-test, therefore the length of messages and the style of usernames and screen names are similar between the two mediums.

We then harmonized the names of the data fields extracted from Telegram to the BotBuster For Everyone algorithm convention. Then, we ran the Telegram data through the algorithm. The BotBuster for Everyone algorithm determined our Telegram dataset to have 29.9% bots and 70.1% humans.

To verify the bot/human labels generated by the BotBuster for Everyone algorithm, we performed manual annotation on a subset of data. From the BotBuster for Everyone output, we randomly selected a 0.1% sample by stratified sampling, thus ensuring that the proportion of bots/human users in the sample matches the proportions that are reflected by the BotBuster for Everyone algorithm. In total, we extracted 2767 data points.

Two of the authors manually annotated the data points, reading through the user names and

messages of each user before labeling the user as a bot or a human. In the event of a disagreement, a third annotator served to break the tie. All three annotators are native English speakers.

We also calculated the inter-annotator agreement between the first two annotators using Cohen's Kappa score. This score ranges from [-1, +1] and serves as an indication of the proportion of annotations that were not in agreement due to random chance. We only compared the first two annotators, with the third annotator serving as a tie-breaker. We obtained a Cohen Kappa score of 0.92. This score is close to 1, indicating sufficient agreement between the two annotators [17, 116].

Finally, we harmonized the manual annotations through maximum pooling, taking the most common out of the three scores. These scores are thus regarded as the gold standard labels for this set of data. We compared the BotBuster for Everyone-generated labels, finding a model F1-score of 72%. This is a reasonable accuracy given that the original model that was fine-tuned on Twitter data performed with an F1 of 73%.

Tiny-BotBuster While BotBuster For Everyone was fast, it was still slow if we were to determine the $P(bot)$ for millions of users. To address that, we optimized the time and size of the model with Tiny-BotBuster [220]. The architecture is illustrated in Figure 2.4. In Tiny-BotBuster, we used a minimal feature set of user metadata and post metadata. These two data subsets are then fed into a MoE architecture that consists of Gradient Boosted Trees [96] and Random Forests [41]. The ensemble output of the MoE is passed through a logistic regression model to return a $P(bot) \in [0, 1]$. The $P(bot)$ is then thresholded with a value of $\epsilon = 0.7$, where $P(bot) \geq 0.7 = bot, P(bot) < 0.7 = human$.

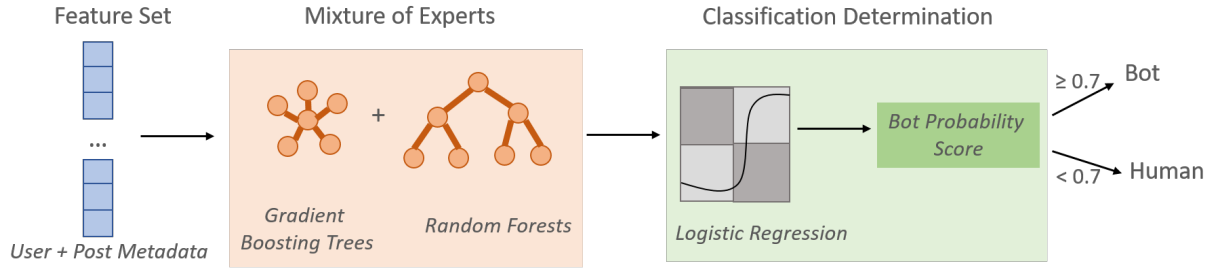


Figure 2.4: The Tiny-BotBuster model architecture (Diagram published in [220])

Multilingual BotBuster Finally, we addressed the mono-linguality of our bot detection algorithm with Multilingual BotBuster [214]. While our algorithms performed with an average of 73.7% bot/human classification accuracy, they were only tuned for the English language. Our extensive 2020-2021 Coronavirus dataset reflected that only 47.4% of the posts were written in English [214]. Therefore, we used multilingual BERT embeddings to construct a multilingual classifier.

We begin by creating a multilingual dataset. We used Google Translate to translate three English-language datasets (cresci-rtbust-2019 [182], botometer-feedback-2019 [321], cresci-stock-2018 [68]) that were manually annotated for bot/human users into three other languages. These

three languages were: Chinese, Russian and Arabic. These languages were chosen because past literature suggests that bots from these countries are active in influencing online narratives [25, 137, 268].

Following which, we performed bot classification using different types of pre-trained tokenizers and classifiers on the texts. These tokenizers and classifiers are pre-trained models from the HuggingFace repository. We used the default parameters for the classifiers, and used the same tokenizers to embed the tweet texts. In applying these algorithms, we used a 80:20 train:test stratified split to the dataset and ran a five-fold cross validation to determine algorithm accuracy. These runs are compared against baseline algorithms, which are algorithms that are commonly and commercially ran for social media bot detection.

We tested three types of models as text classifiers: (1) language-specific models, (2) multilingual models and (3) Large Language Models.

For language-specific models, we tested 2 to 3 classifier types of each language, except when not available. For the English language, we tested: distilbert/distilbert-base-uncased, FacebookAI/roberta-base, google-bert/bert-base-uncased. For the Chinese language, we tested google-bert/bert-base-chinese, hfl/chinese-roberta-wwm-ext. For the Russian language, we used blinoff/roberta-base-russian-v0. For the Arabic language, we used CAMeL-Lab/bert-base-arabic-camelbert-mix-pos-egy, asafaya/bert-base-arabic.

For the multilingual classifier, we tested 4 variations. The first is the TF-iDF vectorizer with a random forest classifier. The next three are taken from the HuggingFace model repository: google-bert/bert-base-multilingual-uncased, FacebookAI/xlm-roberta-base and FacebookAI/xlm-mlm-enfr-1024.

For Large Language Models, we tested a total of 6 prompts, each inserting the tweet text as the {Sentence}. These prompts were tested on two models: flan-alpaca-gpt4-xl and gpt2. The gpt2 model performed better than the flan-alpaca-gpt4-xl model. However, the outputs of the gpt2 model seemed unaffected by the prompt structure.

- Prompt 1: “““Do you think this sentence is written by a bot or a human? Output only the class ‘bot’ or ‘human’. {Sentence} ”””
- Prompt 2: “““Does this tweet look like it was written by a bot or a human? Output only the class ‘bot’ or ‘human’. {Sentence} ”””
- Prompt 3: “““Bots are automated actors on social media platforms. Does this tweet look like it was written by a bot or a human? Output only the class ‘bot’ or ‘human’. {Sentence} ”””
- Prompt 4: “““This sentence is in {Language} language. Do you think this sentence is written by a bot or a human? Output only the class ‘bot’ or ‘human’. {Sentence} ”””
- Prompt 5: “““This sentence is in {Language} language. Does this tweet look like it was written by a bot or a human? Output only the class ‘bot’ or ‘human’. {Sentence} ”””
- Prompt 6: “““This sentence is in {Language} language. Bots are automated actors on social media platforms. Does this tweet look like it was written by a bot or a human? Output only the class ‘bot’ or ‘human’. {Sentence} ”””

Finally, we selected the best performing multilingual model (google-bert/bert-base-multilingual-uncased) to be used to identify bot and human accounts. This Multilingual BotBuster architecture

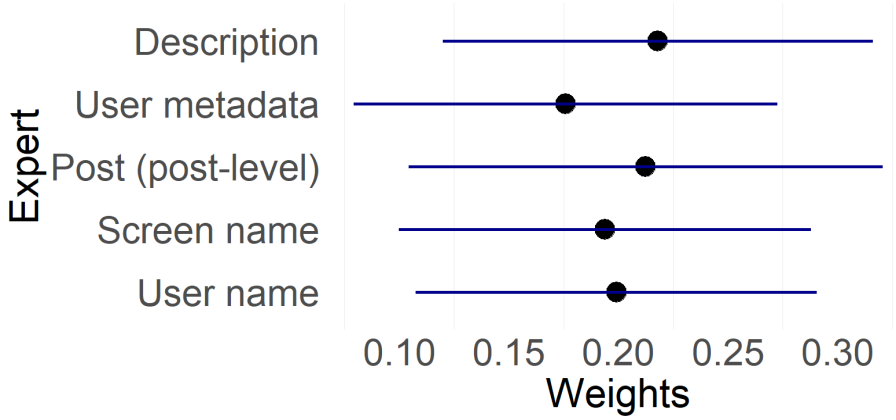


Figure 2.5: Distribution of expert importances in the BotBuster architecture (published in [205])

performed with an average accuracy of $82.79\% \pm 4.23$ for all the three datasets across all three languages, as compared to the baseline comparisons: BotBuster at $66.91\% \pm 12.57$, BotBuster For Everyone at $77.69\% \pm 6.38$, Botometer at $50.63\% \pm 11.30$ and BotHunter at $58.12\% \pm 18.83$, for English-only tweets.

2.4.1 Features important to Bot Detection Algorithms

Having constructed bot detection algorithms using machine learning techniques, we next seek to find out what are the key features that identify a bot.

The expert importances of each of the experts in contributing to the final $P(bot)$ score in BotBuster are graphed in Figure 2.5. The final $P(bot)$ score was a weighted sum of the experts derived by an expert gating network, which was a multi-layer perceptron. From the expert importances analysis, the input weights were almost evenly distributed across the experts, suggesting that all experts had almost equal importances for the final prediction. However, there were slightly more emphasis on post information and descriptions and lesser value on screennames and usernames. The gating network placed more emphasis on the experts that had text as input pillars compared to those that had numeric inputs (e.g. metadata), and within the text input experts, had more emphasize on the experts with longer text input. This suggests that longer writing was a key determinant of $P(bot)$. In the training data, the average length of a username was 3.02 ± 5.15 characters, and the average length of a screenname was 3.28 ± 6.23 characters. The average Levenshtein distance between both name strings, reflecting the similarity of the screenname and username was 2.36 ± 4.89, suggesting users typically used similar strings for both these names. The description data field of an account was usually more lengthy, containing an average of 3.62 ± 6.33 words.

For the BotBuster For Everyone algorithm (and subsequently the BotBuster-Telegram, Tiny-BotBuster and Multilingual BotBuster), we kept the feature spaces small in our feature extraction implementation. Despite these, we were still able to achieve decent algorithm accuracy, showcasing that bot detecting need only a few key features for a decent accuracy. We extracted the feature importances of each of the estimators stored in Python's sklearn classifiers used for the

each of the data experts, and graphed their importances in Figure 2.6.

The username/screenname and post metadata features were numeric features, and the tree-based classifiers of the experts separated them through the decrease in impurity. The higher mean decrease in impurity, the more important the feature is in differentiating the final bot/human class. For the username/screenname, the entropy of the name also plays a large factor in determining bot classification. For post metadata, the most indicative feature was the number of retweets/shares, followed by the number of likes and the number of replies that the post received.

Finally, the description of an account was a string of words, and hence was treated differently by the Decision Tree classifier. The classifier breaks down the description string into a bag of words, so the feature importances in Figure 2.6 was represented by coefficients attached to words. The coefficient scores reflect how important a word was within a description string. The word with the highest coefficient and therefore the most important was "bot". Words representing a person's identity (i.e. writer, mom, host, author, reporter, editor etc.) were extremely indicative words, suggesting connections between the expression of identities and bot likeliness of an account. These suggests possible incorporation of heuristics to identify key signals of bot accounts with words present.

Understanding the features that are important to a bot detection algorithm provides directions for further bot account analysis by characterizing the defining features of bot vs human accounts.

2.4.2 Thresholding Bot Detection Algorithms.

Bot detection algorithms typically return a probability score of a user being a bot $P(bot)$ that is between 0 and 1, to which a threshold ϵ is being applied. If $P(bot)$ is greater or equal to ϵ, the account is determined to be a Bot; if $P(bot)$ is less than ϵ, then the account is determined to be a human. Many studies use an arbitrary ϵ value, which can result in instability and misclassifications of users. Too tight a threshold (i.e., a higher ϵ used) means that lesser users are classified as a Bot, and there will be more false negatives. Too loose a threshold (i.e., a lower ϵ used) means that more users will be classified as a Bot, and there will be more false positives. We perform a large scale longitudinal and statistical analysis to determine a suitable ϵ value for a bot detection algorithm, which is an essential step for stable and comparable results across studies.

We collected the tweets of 5000 agents from X on a daily basis for 150 days beginning from September 2020. These users had at least 100 tweets related to the 2020-2021 coronavirus and the 2020 US elections. We used the BotHunter algorithm [32] as our detection algorithm. We established the stability of $P(bot)$ through the patterns of agents that flip bot classification. We termed an agent to have flipped bot classification when it was previously classified as a bot and currently classified as a non-bot or vice-versa. This flipping behavior was determined when an agent had a bot score that fell below the pre-determined threshold value, then rose above that same threshold value, or vice versa. With five threshold values $\epsilon \in [0.25, 0.30, 0.50, 0.70, 0.75]$, we investigated the stability of $P(bot)$ across two dimensions:

1. temporal stability, which was how bot scores changed with increasing number of days. At each increasing number of days, we calculated the mean score difference of the agents' current score at day $t = i$ against their first day's score at $t = 0$, i.e. $P(bot)_{t=i} - P(bot)_{t=0}$
2. volume stability, which was how the scores changed across the number of tweets. At each

increasing number of tweets, we calculated the we calculated the mean score difference of the agents' current score with the current number of tweets $n = i$ against their first tweet's score at $n = 0$, i.e. $P(bot)_{n=i} - P(bot)_{n=0}$.

When we analyzed agent flipping behavior based on initial bot class, we find a decreasing trend as time and volume increases. Figure 2.7 and Figure 2.8 show the downward trend of the proportion of agents that flip bot classification by temporal and volume analysis respectively. This shows that bot classification stabilizes as the amount of data increases.

Additionally, a larger proportion of initial non-bots flipped their classification to bots. In the temporal lens, the largest percentage of non-bots that flipped occurred at the 10-day mark, with 6.32%. A largest percentage (5.13%) of bots that flipped their classification also occurred at the 10-day mark. Through the volume lens, a maximum of 9.10% of non-bots that flipped classification while 3.89% of bots flipped classification. Both occurred at the 20-tweet mark.

As the number of days and number of tweets increased, the proportion of agents with different bot classification compared with the first classification tended to zero. By temporal steps, the peak number of days of bot classification changes was 10 days, where 2.05% to 8.98% of agents change classification. After which, the proportion of agents that changed classification decreased across time. Observations by volume steps showed that a decent value for the sharp decrease in the proportion of agents that change classification is around 20 tweets.

Flipping bot classification pointed to the instability of bot scores. Across all threshold values, initial bot percentages and proportion of agents that flipped were similar whether measured by time or volume. This highlighted that the classification of agents into bots and non-bots were actually relatively stable from the initial bot classification. In both temporal and volume perspectives at the 0.75 threshold value, approximately 13% of agents flipped bot classification, which is consistent with past studies Rauchfleisch and Kaiser [251]

With this methodology, we established parameters for using supervised bot detection algorithms. This methodology can generically be applied to any bot detection algorithm. In our experiments, we utilized the BotHunter detection algorithm, and hence summarize our recommendations for parameters to use with this algorithm:

1. For a consistent bot probability score, a reasonable data collection size is at least 20 days of tweets or 40 tweets.
2. In terms of bot prediction algorithm stability, a recommended threshold level is 0.70. For a consistent bot classification score, a recommended collection size is at least 10 days of tweets or 20 tweets.

The amount of data required for a stable bot classification score was lesser than the data required for a stable bot probability score because bot classification is a binary classification dependent on a range of values. A small change in bot probability score that leads to the classification will unlikely alter the classification of the agent, except at the threshold boundaries. However, the precise determination of $P(bot)$ required more data because it is a float value in the continuous range of [0,1].

2.5 Conclusion

Our work aims to build bot detection systems that can differentiate between bot and human social media users. We developed five bot detection algorithms using the data-resilient Mixture-of-Experts architecture: BotBuster, BotBuster For Everyone, Tiny-BotBuster, BotBuster-Telegram and Multilingual BotBuster. Together, these algorithms can perform bot detection for four social media platforms (X, Reddit, Instagram, Telegram), and across four languages (English, Chinese, Arabic, Russian). Each algorithm improves on the previous in terms of accuracy and computational efficiency.

Despite our efforts at improving our bot detection algorithms, there are still some **limitations:**

1. Our bot detection algorithms are constructed for only four platforms (X, Telegram, Reddit, Instagram), and primarily work for four languages (English, Russian, Chinese, Arabic). While we expect bots to generally act with similar features across social media platforms, we must not discount that bots will harness the affordances of each platform.
2. Current detection still focuses primarily on account-level (e.g., metadata, temporal activity) and post features (e.g., number of words, reading difficulty). Richer multi-modal features like images and videos remain unexplored.

Future work should evolve bot detection algorithms according to updates in social media platform affordances and changes in platform usage. Being able to adapt to the changing content modalities and bot strategies can be particularly impactful because it allows researchers and practitioners to analyze and respond to the information environment.

Year	Reference	Definition
2016	[90]	A social bot is a computer algorithm that **automatically produces content** and **interacts with humans** on social media, trying to emulate and possibly alter their behavior.
2016	[35]	[...] social bots, **algorithmically controlled accounts** that **emulate the activity of human users** but operate at much higher pace (e.g., automatically producing content or engaging in social interactions), while successfully keeping their artificial identity undisclosed
2016	[58]	**Automated accounts**, called bots, [...]
2018	[108]	Bots are have been generally defined as **automated agents** that function on an online platform [..]. As some put it, these are programs that run continuously, formulate decisions, act upon those decisions without human intervention, and are able adapt to the context they operate in.
2018	[40]	The term "social bot" describes accounts on social media sites that are **controlled by bots** and **imitate human users** to a high degree but differ regarding their intent.
2018	[32]	[...] malicious **automated** agents
2020	[234]	Social Media Bots (SMB) are **computer algorithms** that **produce content** and **interacts with users**
2020	[20]	[...] social bots, **(semi-) automatized** accounts in social media, gained global attention in the context of **public opinion manipulation**.
2020	[260]	Malicious actors create **inauthentic social media accounts** controlled in part by **algorithms**, known as social bots, to disseminate misinformation and agitate online discussion.
2021	[180]	Social bots – partially or fully **automated accounts** on social media platforms [...]
2022	[218]	Social media bots are **automated accounts** controlled by software algorithms rather than human users
2023	[122]	Social bots are **automated** social media accounts **governed by software** and controlled by humans at the backend.
2023	[80]	A bot is a software that mimics human behavior and **operates autonomously and automatically**.
2023	[282]	Twitter accounts controlled by **automated programs**.
2023	[319]	**Automated accounts** on social media that **impersonate real users**, often called "social bots,"
2023	[318]	Social bots are social media accounts **controlled by software** that can **carry out content** and **post content** automatically.
2024	[144]	Social bots are **artificial agents** that infiltrate social media
2024	[320]	Social bots are social media accounts **controlled in part by software** [...] Social media bots display profiles and **engage** with others through various means, including following, liking, and retweeting

Table 2.1: Definitions of "Social Media Bot" in academic literature. (Table published in [212])

Year	Reference	Definition
2018	US Department of Homeland Security [75]	[...] Social Media Bots as programs that vary in size depending on their function, capability, and design; and can be used on social media platforms to **do various useful and malicious tasks while simulating human behavior**
2024	Microsoft [188]	Social media bots are **automated programs** designed to **interact** with account users.
2024	Meltwater [269]	Refers to the definition by US CSIA (see below)
Not Dated	CloudFlare [66]	[...] social media bots are **automated programs** used to **engage** in social media. These bots behave in an either partially or fully **autonomous fashion**, and are often designed to **mimic human users**.
Not Dated	Cybersecurity and Infrastructure Security Agency (CISA) [65]	Social Media Bots are **automated programs** that **simulate human engagement** on social media platforms.
Not Dated	Imperva [135]	An Internet bot is a software application that runs **automated tasks** over the internet.

Table 2.2: Definitions of "Social Media Bot" in industry literature. (Table published in [212])

	User		Content		Interactions	
Reference	**Automation**	**Mimicry**	**Creation**	**Distribution**	**Communication**	**Relationship**
[90]	x	x	x			
[35]	x	x	x		x	x
[58]	x					
[108]	x					
[40]	x	x		x		
[32]	x					
[234]	x		x		x	x
[20]	x					
[260]	x			x		
[180]	x					
[218]	x					
[80]	x	x				
[122]	x	x	x			
[282]	x					
[319]	x	x				
[318]	x		x	x		
[144]	x					
[320]	x				x	x
US Department of Homeland Security	x	x				
Microsoft	x				x	x
CloudFlare	x	x			x	x
CISA	x	x			x	x
Imperva	x					

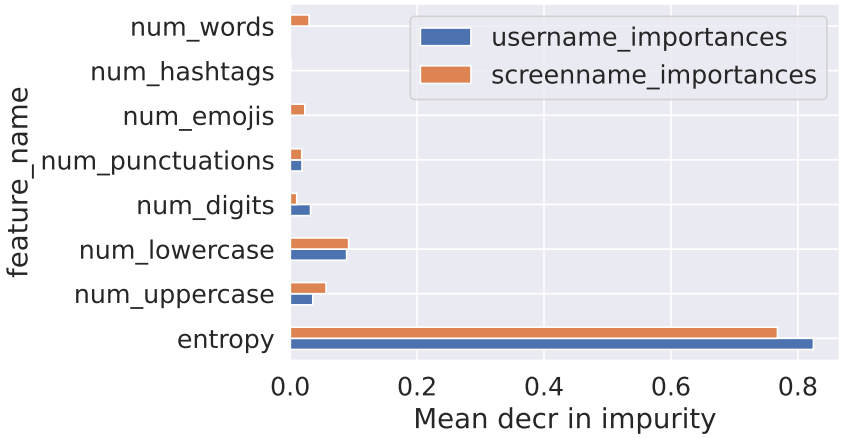
Table 2.3: Components of definitions of "Social media Bot". (Table published in [212])

	BotBuster [205]	**BotBuster For Everyone [210]**	**Tiny BotBuster [220]**	**Multilingual BotBuster [214]**	**BotBuster-Telegram [221]**
Model Architecture	Mixture-of-experts + Deep learning	Mixture-of-experts + Random Forests	Random Forests	Mixture-of-experts + Random Forests	Mixture-of-experts
Platforms	X, Reddit, Instagram	X, Reddit, Instagram	X	X	Telegram
Average Accuracy (%)	72.73	56.84	91.78	72.00	82.79
Key features	Handles incomplete data pillars	Fast version of BotBuster that does not require GPUs	Small model size, extremely fast model building and inference	Tested for English, Russian, Chinese, Arabic languages	
Limitations	Slow model building and inference, requires GPUs	For English language	Solely for X	Relied on translated language data	Solely for Telegram

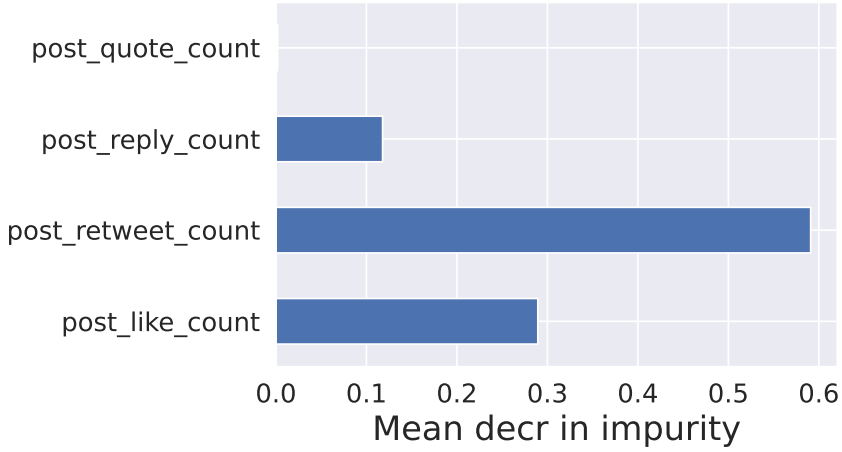
Table 2.4: Bot Detection Algorithms Developed in this Thesis

True Negatives	Algorithm doesn't think user is a bot & User actually is not a bot	61.4%
False Positives	Algorithm think user is a bot & User actually is not a bot	13.3%
False Negatives	Algorithm doesn't think user is a bot & User actually is a bot	12.3%
True Positives	Algorithm think user is a bot & User actually is not a bot	13.0%

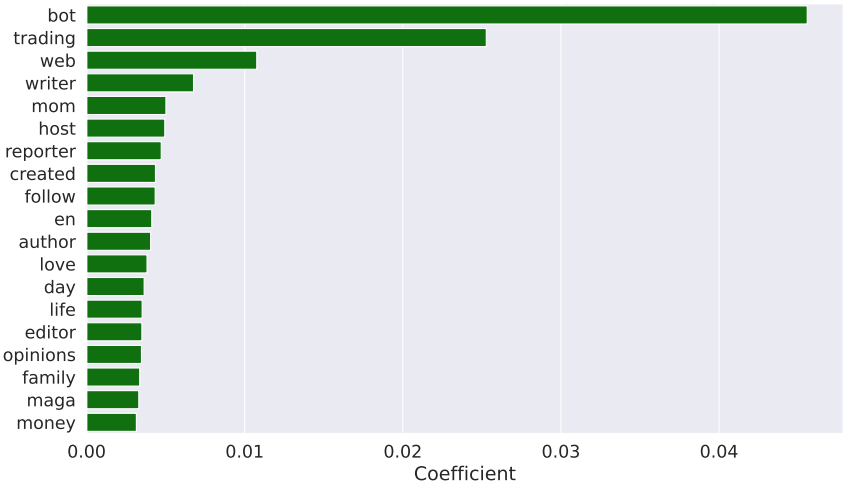
Table 2.5: Accuracy Metrics of the BotBuster Algorithm



(a) Username/Screenname features

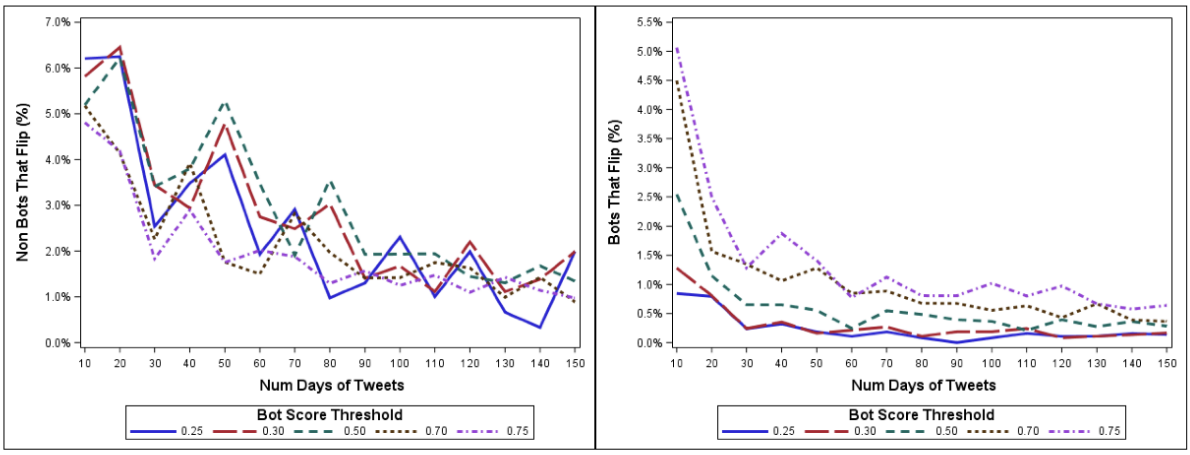


(b) Posts metadata features



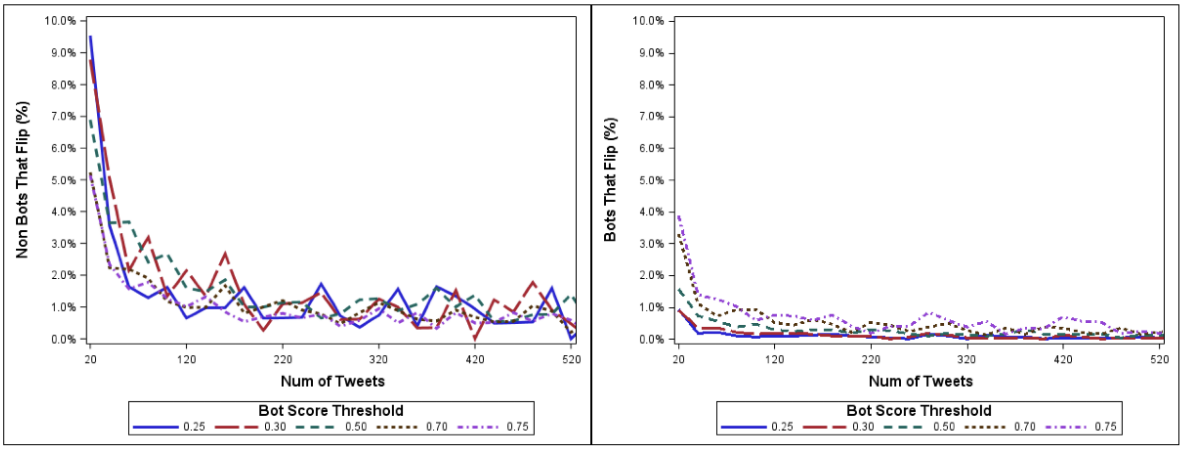
(c) Description features

Figure 2.6: Feature Importances from the BotBuster for Everyone algorithm. (published in [210])



((a)) Non-Bots that flip ((b)) Bots that flip

Figure 2.7: By Temporal Steps: Agents that Flip Bot Classification (%). The largest percentage of flips occur at the 10-day mark for both bots and non-bots, and with a downward trend of bot flipping behavior, 10-days worth of tweets is a reasonable data size for stable bot classification. (published in [218])



((a)) Non bots that flip ((b)) Bots that flip

Figure 2.8: By Volume Steps: Agents that Flip Bot Classification(%). The largest percentage of flips occur at the 20-tweet mark for both bots and non-bots, and with a downward trend of bot flipping behavior, 20 tweets is a reasonable data size for stable bot classification. (published in [218])

Chapter 3

From Bots to Cyber Social Agents

3.1 Introduction

The term "bot" by itself describes a generic automation algorithm, treating all forms and behaviors of an automation as a homogeneous entity. We expand the term "bot" into a "Cyber Social Agent" (CSA), framing automation in terms of nuanced sets of archetypes rather than a monolithic category. Automated accounts are not merely scripts, but they are a suite of agents that adopt specific roles, employ different strategies, and interact uniquely with humans, other CSAs and platform infrastructures. These archetypes have agency, and therefore can be either good or bad. Some CSAs serve constructive roles like news dissemination and emotional support, while others spread misinformation and manipulate public discourse. This section describes the fifteen different Cyber Social Agent archetypes, with the following guiding **research questions:**

1. What are Cyber Social Agents?
2. What are the different types of Cyber Social Agents that exist on social media space?
3. When and how do CSA personas contribute positively to social good, and when do they become harmful?

3.2 Related Work

Typologies of social media agents Bot detection algorithms usually provide a binary classification of a social media user [71, 205, 322]. That is, based on the features extracted from the user, it is either a bot or a human. This binary classification can be insufficient for downstream analysis of social media discourse, because there is a spectrum of behaviors that a user can undertake. Automated accounts can be amplifiers [42], news broadcasters [163], repeaters [257], or even include hybrid "cyborg" accounts that are partially automated and partially human [64, 295]. This heterogeneity in accounts highlights that computational detection should not be restricted to separating bots from humans, but also account for the diverse form of automation that shape digital interactions.

There are more nuanced classification of types of social media bots in the literature. Typologies mostly focus on either the content or the interaction mechanics of the automated user.

Typologies focus on the content type (i.e., malicious or benign [279]; original or not) [1, 231], the amount of content produced (i.e., none, some, a lot [1]) or the effect on social media content (i.e., distributing, amplifying, distorting, hijacking content [108, 184]). Typologies can also focus on the interaction mechanics used. For example, accounts that constantly retweet other users are retweet or amplifier bots [82, 137], accounts that follow a lot of other accounts as fake-follower bots [64], Our work expands the classification of automated users as a homogenous "bot" vs "humans" into a nuanced set of Cyber Social Agent archetypes, where agents are situated along multiple dimensions of visible behavior on social media platforms.

Goodness of Cyber Social Agents In the study of cyber social agents, many studies in the literature focus on malicious agents that cause harm to the online ecosystem through false information [4, 263] and information manipulation [100, 166, 204]. Indeed, there are plenty of good bots: bots that provide notifications and entertainment [129], and bots that provide emotional support during stress [245] and grieve [151]. Unfortunately, while many bot detection algorithms reliably differentiate whether a user is a bot or not, the subsequent analysis usually groups the identified bots homogeneously as malicious entities to be studied. While it is important to profile the potential harm of Cyber Social Agents, it is equally as important to examine the good agents that quietly work behind the scenes to preserve our cyber social health by performing activities like correcting false information [289].

3.3 Cyber Social Agents

A Cyber Social Agent (CSA) is a digital actor (the User) embedded within a social network environment, with the capacity to perceive, process and act upon information in ways that influence the narratives (the Content) and other actors (the Interaction) in the ecosystem. Figure 3.1 presents a formal definition of the Cyber Social Agent in the User-Content-Interaction frame.

Definition. (Cyber Social Agent)

user
A digital actor embedded within a social network environment,
with the capacity to perceive, process and act upon information in ways that
content
influence the narratives
interactions
and other actors in the ecosystem.

Figure 3.1: Definition of Cyber Social Agents

Table 3.1 presents how the concept of Cyber Social Agents evolves from the general perception of bots. Bots are the simplest form of Cyber Social Agents. While bots are typically seen as automated scripts controlled by human operators, CSAs are understood as actors with varying degress of agency, adaptability and self-evolution. The label "bot" usually connotes a generic form of automation and is often associated with malicious activity [115]. In contrast, the CSA frame differentiates automated agents by their personas, their roles, behaviors and rhetorical strategies. Crucially, CSAs are not inherently bad. Automation can serve both harmful and

beneficial purposes, and should be evaluated along the dimensions of content, behavior and effect. This framing suggests that regulation should not apply uniformly to all automated agents, but instead be tailored to each persona.

Table 3.2 situates our Cyber Social Agents within the broader trajectory of agent development. It compares CSAs to Agent-Based Modeling (ABM) Agents, Artificial Intelligent (AI) Agents and Agentic AI across multiple dimensions, including autonomy, automation capacity, learning ability, embodiment and modes of interaction with humans. Agent-Based Modeling agents are computational agents programmed to realistically interact with the environment or other agents that that is relevant to the complex system being studied [105]. ABM agents are confined to simulation environments.

AI and Agentic AI agents primarily operate in decision-making contexts, with AI agents residing solely in the digital space while Agentic AI Agents having physical-digital contexts. Examples of AI agents are web agents [228] or coding agents, and examples of Agentic AI agents are embodied AI robots [98]. In contrast, CSAs are uniquely situated within social media ecosystems. They combine features like high interaction with agents, natural language communication and multimodal content generation. These capabilities allow them to directly influence public discourse and human behavior.

By contrasting CSAs against other (prior) forms of agents, Table 3.2 illustrates that CSAs are not simply another step in technical sophistication but are a qualitatively distinct category of socio-technical actors. CSAs are digital entities that are embedded in, and sometimes co-evolving with, human and social media platform dynamics.

Characteristics	**General Perception of Bots**	**General Concept of Cyber Social Agents**
Agency	Automated by operator	Automated by operator, or can be self-evolving, or can have agency
Types	One general type	Different types of personas
Nature	Bad	Can be good or bad Can be harnessed for good or bad
Regulation	Should be regulated in the same way	Can be regulated differently based on persona, content, behavior.

Table 3.1: From Bots to Cyber Social Agents

Dimension	Agent-Based Modeling (ABM) Agents (1940-)	Artificial Intelligence (AI) Agents (2022-)	Agentic AI (2023-)	Cyber Social Agents (2026 -) Social Media Bots (2008-)
Main use: Primary purposes of main tasks assigned	Simulation-based [235]: Goal-directed action, decision making	Decision making [256]	Goal directed action [2, 2]	Announce information [19], influence, engagement etc.
Autonomy: Degree of independent action without external control	Full	Partial or Full	Partial or Full	Partial or Full
Automation Capacity: Types of automated behavior execution	Heuristic-based [67]	Large Language Models (LLMs) [133]	Heuristic + LLMs	Scripted behavior or heuristic or LLM or combinations
Environment they live in: Contexts and domains they operate in	Simulation, Research, Prediction	Web [228], Code environments [191]	Physical-Digital interfaces (i.e., robotics [101])	Social media, online platforms
Interaction with humans: How directly or meaningfully they interact with people	No	Moderate	Extensive	High
Learning ability: Whether they learn from data, interactions or feedback	No	Some (supervised)	Yes (self-learning)	Some yes, some no
Embodiment: Type of body presence they have	Digital	Digital	Physical	Digital
Medium/ modality: Interface or communication format	Code, numeric strings	Natural language (text, voice), Code (APIs)	Text, voice, visual, physical action	Text, image, video, hashtags, URLs
Problem solving capability: Ability to analyze and respond to complex solutions	Scenario-driven simulation logic [76]; Response can be logic-based, algorithmic-based, mathematically-based	Logic, ML-based, reasoning	Natural language (input/output), multimodal	Natural language, hashtags, emojis
Communication type: How commands are expressed or interpreted	Numeric strings, often binary	Code-based prompts with natural language	Code-based prompts with natural language	Natural language
Learning ability: Sources of learning	None, Some (Learning by being told, Bayesian updating [291], Regression equations [60])	Mostly told (labeled data)	Learned from experience, Bayesian updating	None (scripted), supervised learning, reinforcement learning

Dimension	**ABM Agents (1940-)**	**AI Agents (2022-)**	**Agentic AI (2023-)**	**Cyber Social Agents (2026 -)** **Social Media Bots (2008-)**
Lifespan: Operational duration or persistence	As needed by virtual experiments [195]	Often persistent for use in code	Long-lived, requires maintenance	Operator-defined; Lifespan can be prematurely terminated by platforms
Mobility: Physical or virtual movement capability	None	Virtual navigation in code	Physical mobility (robots, drones)	None (stationary accounts), but the reach of their network expands their influence
Examples: Real-world implementations	Conway Game of Life [141]	Chat Bots [62], Recommender Systems [132], Coding Agents	Embodied AI Robots [98]	Social Media Bots

Table 3.2: Summary of the different types of agents and uses

3.4 Archetypes of Cyber Social Agents

CSAs are automated agents with a variety of archetypes, and each archetype impacts the online conversation in different ways. Past works formulated typologies of automated users mainly through the subjective process of human labeling [1, 108, 231], which is resource intensive and difficult to scale when analyzing large volumes of users. We describe CSA personas along the User-Content-Interaction framework that we had formulated for social media platforms in chapter 1, and develop computational heuristics as a systematic scheme to provide a consistent classification of agent archetypes.

In developing this scheme, we surveyed literature from empirical studies and analysis of typologies of social media bots. This literature consists of papers published between 2015 and 2025 that were extracted using the keywords "social media bot types", "social media bot typology", "typology of social media bot", "types of social media bot" from Google Scholar. This survey yielded 123 papers. We excluded 48 papers that constructed bot detection algorithms, 29 papers that were on bot activity analysis, 3 papers that were on both bot detection and bot activity analysis, and 4 papers that were on bot vs human characteristics. This left us with 39 papers, which are presented in Table 3.3. These papers cover a wide variety of archetypes of bots, but unfortunately do not have a systematic way of classifying these bots, nor a clean definition for the classes of bots. These classifications are largely event-based or study-based, identifying bots by intuition from a study of a set of data. This thus prompted our creation of a systematic taxonomy of well-defined archetypes of bot agents.

Year	**Ref**	**Archetypes listed**	**Classification criteria**	**Limitations**
2015	[1]	core bots, peripheral bots, generator bot	types of bots in a bot net	limited to the structural roles of users in a network
2016	[231]	Broadcast Bots, Consumption Bots, Spam Bots	How bots disseminate content	Studied only bots that are used for disseminating & aggregating content
2016	[314]	political bot	event-based	only studied the political bot in different political regimes
2016	[312]	political bot	event-based	limited to an event
2016	[198]	broadcast, spam, influencer	study-based	limited to the study construct

Year	Ref	Archetypes listed	Classification criteria	Limitations
2017	[279]	astroturfing bot, social botnets in political conflicts, infiltration of an organisation, influence bots, sybils, doppelganger bots, spam bots, fake accounts, pay bots, nonsense bots, news bots, recruitment bots, public dissemination account, earthquake warning bots, editing bots, chat bots	intentions	Qualitative taxonomy based off examples from literature
2017	[295]	simple bots, sophisticated bots	based on bot detection probabilities and whether they mention humans	arbitrary probability bounds to determine bot type
2017	[104]	celebrity status, very popular, mid-level recognition, lower popularity	number of followers	limited to classification based on ”influencers”
2018	[31]	social influence bot	repetitive @mentions of each other	event-specific
2018	[11]	spam bot, emotional bot	event-based	limited to an event
2018	[111]	simple automation, human-like acting bots, intelligent acting	referenced from [110]	machine learning based case study
2019	[266]	spam bots	bot identified from a honeypot	limited to identification of a single honeypot
2020	[260]	spammer, simple bots, fake followers, self-declared, political bots	based on data that has been bought	limited to data bought
2020	[108]	Web robots (Crawlers & Scrapers), Chatbots, Spambots, Social bots, Sockpuppets & Trolls, Cyborgs & Hybrid accounts	the bot’s structure, function & use	High-level qualitative taxonomy, and many nested, orthogonal questions in each of the three classification criteria

Year	Ref	Archetypes listed	Classification criteria	Limitations
2020	[20]	bot, bot army, bot net, cyborg bot, political bot, propaganda bot, remote bot, social bot, spam bot	the orchestration and capabilities of a bot	focus on clearnet and dark market places
2020	[130]	news bot	study-based	limited to the study construct
2020	[273]	media bots	study-based	limited to the study construct
2020	[91]	astroturf, fake follower, financial, self-declared, spammer, other	as per [322] classification	limited to [322] classification
2021	[61]	drifters	study-based	limited to the study construct
2021	[259]	heavy bots, political bots, media spambot, user-generated bot, campaigner bot, inciting agent bot	content based categories	limited to content differentiation
2021	[148]	premium+, ultima, low, high, medium, standard, good, best	extent that the bot looks like a human, type of action, speed of action	based on the classification of marketplace bot rental services and forums
2022	[46]	spam bot	event-based	limited to an event
2022	[280]	human, bot, self-declared bot	not sure	narrow and unexplained classification
2022	[183]	trolls, bots, humans	credibility, initiative, adaptability	Twitter-based classification
2022	[272]	General, fake follower, other, self declared, spammer	as per [322] classification	limited to bots that are involved in a specific event
2022	[149]	price, bot-trader type, normalized bot quality, speed, survival rate	classified for people to understand the types of bots involved in a social media attack and estimate the attack	limited to bots that are involved in a social media attack
2022	[274]	brutal, vulgar, sex, detestation	not sure	limited to malicious conversational bots

Year	Ref	Archetypes listed	Classification criteria	Limitations
2022	[299]	likers, disinformers, reposters, commentators, personal data collectors, account blockers, discrediators, haters, advertiser, political troll bots, opinion leaders, bullying, simulators of real user behavior, distributors of fakes, clogging hashtags, meaningless content	technical, combat, trolls, disinformers, spammers,	limited to a threat framework, and some bot types are repeated under different categories
2023	[207]	general bot, news bot, bridging bot	event-based	limited to an event
2023	[10]	Social bots, Spam bots, Sybil bots	Did not really explain	Bot types are for a machine learning algorithm for bot detection algorithm
2024	[5]	spam bot, follower bot, astroturf bot, doppelganger bot, political bot, sybil bot, financial bot, romance bot, social bot	not sure	limited to information manipulation and malicious campaigns
2025	[164]	spam bot, social manipulation bot, personalized attack bot, influence manipulation bot, news and information bots, moderation bots, entertainment bots, customer service and promotion bots	bots that the authors have observed in recent years	grouped by malicious activity and good bots
2025	[212]	General bot, Bridging Bot, Political Bot, Chat Bot, Activist Bot	providing examples of bots that are used for good and bad	Listing examples of Bots

Table 3.3: Survey of archetypes of "social media bot" from 2015 to 2025

From this set of survey, we harmonized the agents described to derive a cohesive set of

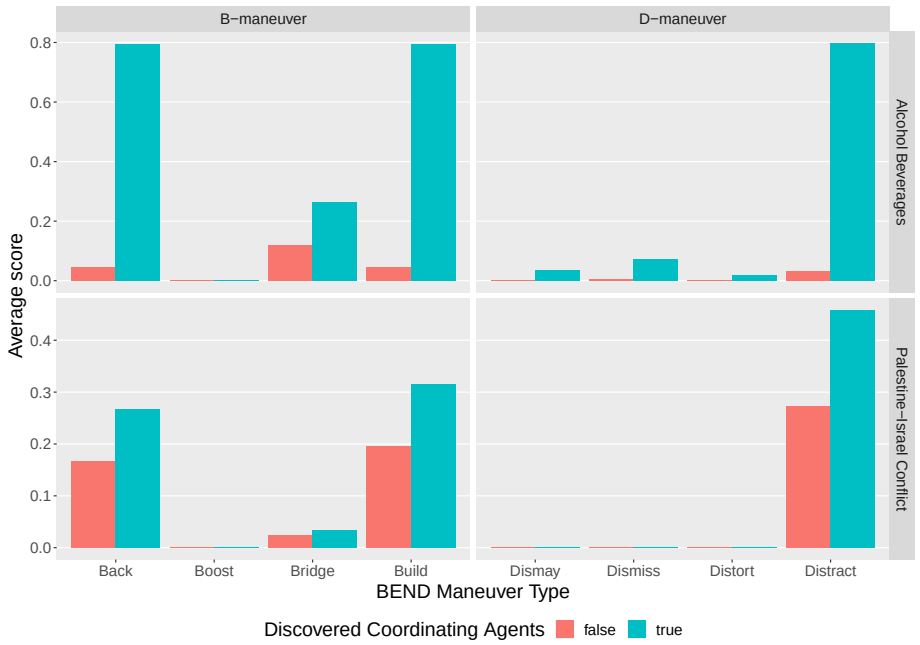


Figure 3.2: Social influencers that coordinated together to hijack trending hashtags used high amounts of information maneuvers (Figure published in [70])

archetypes of Cyber Social Agents, we manually annotated archetypes of agents described in the papers, then used iterative manual refinement to arrive at our final set. The final set of archetypes are described in Table 3.4, and we will describe them in this section using the UCIA (User-Content-Interaction-Algorithm) frame. We note that an archetype is not exclusive. An agent can exhibit signs of multiple personas. For example, a News CSA can also be an Announcer CSA, because it announces sets of news. Agents can also change their archetypes, either they are programmed with a different set of goals, or because they are compromised accounts that have been taken over [81].

Social Influencer are designed to shape public opinion by actively engaging with the discourse and steering conversations towards particular viewpoints. From a UCIA perspective, they are distinguished by their Content and Interaction pillars, through their high reliance of information maneuver cues and disproportionate use of replies. Their interactive and participatory nature naturally makes them an account that is recommended by the Algorithm.

In a study we did on Twitter activity in Indonesia, social influence agents coordinated together to hijack trending hashtags and actively engage in the conversation with high amounts of Back / Build / Distract information maneuvers [70]. This participation was retweet-heavy and reply heavy. The plot Figure 3.2 shows the high proportion of information maneuvers that these social influence agents use as compared to normal agents.

Amplifiers are designed to boost the reach and visibility of specific pieces of content or narratives through high-frequency sharing behaviors such as re-posting or re-tweeting. By repeatedly broadcasting a message, Amplifiers create the illusion of widespread grassroots support and help manufacture momentum around their desired narrative themes, eventually amplifying influence, so that their selected content circulates faster and farther than it would organically.

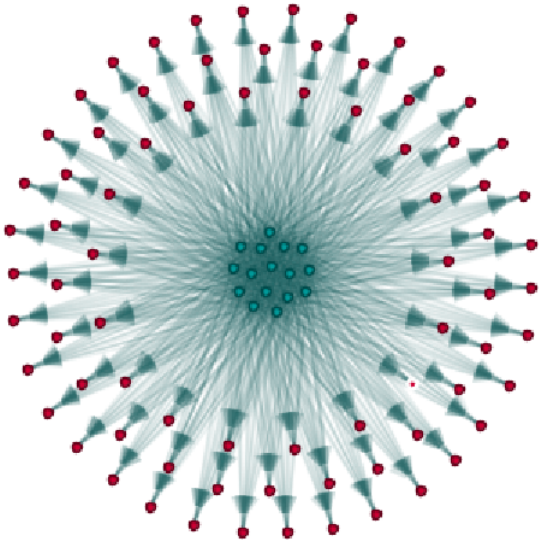


Figure 3.3: Round Robin network of Amplifiers (published in [137])

From a UCIA perspective, Amplifiers are primarily defined by their Interaction Pillar, because their hallmark is the excessive sharing patterns. This operational signature is evident in their disproportionate retweet-to-original-post ratio and limited original message construction. Algorithms of social media platforms prioritize engagement signals such as frequency, velocity and repetition[199], so their excessive sharing patterns are rewarded by the algorithmic systems.

Within the 2021 cross-strait discourse between China and Taiwan, we discovered an algorithmic sequence for using amplifier agents [137]. This tactic employs a round-robin sequence to consistently promote the same accounts, thus amplifying them. The sequence plays out as such: (1) A few core agents generate original tweets, (2) Peripheral amplifier agents retweet the tweets generated by the core agents in a randomized order, and continue retweeting until each core tweet has been retweeted once by every adjacent periphery agent, (3) a fresh batch of central agents generates more original tweets, which the peripheral accounts also retweet as per the second step. This sequence artificially creates engagement between the agents, making them seem less bot-like, and might have contributed to their longer lifespan. [137] shows the all-communication network of these amplifier agents. This network has four hierarchies: the core group of agents in the center that create original tweets, then three other layers of amplifier agents in the periphery that iteratively retweet the set of agents in the previous layer.

Cyborgs are hybrid accounts that combine automated functionality with human oversight. Automation assists the human operator in tasks like mass sharing, high-frequency posting or forwarding content, while human intervention provides flexibility, nuance and the personal touch.

From a UCIA point of view, Cyborgs are defined by their User properties and are sustained by their dynamic interaction with the Algorithm. Cyborgs blur the boundaries between human and automation and therefore have frequent changes in their bot classification, confusing both bot detection systems and recommendation algorithms. The automated component of a Cyborg keeps it algorithmically alive, and helps them maintain visibility through continuous activity, while the human element introduces authenticity which allows the account to evade detection yet

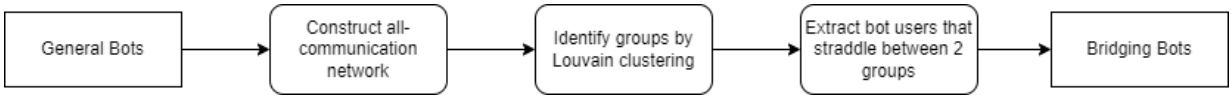


Figure 3.4: Methodology of identifying Bridger (published in [207])

maintain its audience reach.

In our analysis of 2020-2021 coronavirus discourse, we found Cyborgs by identifying accounts that flipped from "bot" to "human" classification and back again, meeting quantitative thresholds for duality in bot/human scoring [222]. From the empirical observations, we established a quantitative threshold for Cyborg classification as agents that (1) excessively flip bot classification (i.e. more than three times), and (2) have a large change in bot probability score between the flips (i.e. more than 0.10 change). Many Cyborgs are centrally embedded in the all-communication network, and have high betweenness and degree centrality. These agents also have significantly longer active lifespan than traditional bots, suggesting that the bot-human duality does improve agent survivability.

Bridgers are designed to connect otherwise separate groups of users and serve as intermediaries in the flow of information across communities. These groups can differ by interests, identities, ideologies or social spaces. Bridgers draws them into shared conversations by tagging multiple users, often cross-posting overlapping content that is of interest to the groups tagged.

From a UCIA perspective, Bridgers are identifiable from their Content with high Bridge scores in the BEND metric[48]. Another way to find bridges is to study their Interaction[224]: Bridgers typically tag multiple people from different groups, whether social groups or groups formed from interactions. Bridging Agents can be boosted by platform algorithms because their engagement diversity and cross-audience interactions can be interpreted as broad resonance. Figure 3.4 presents the methodology we used to identify the bridging bot agent, as implemented in [207].

In the work studying the digital diplomacy between US and China, we found bridgers that occupied boundary positions between the US-affiliated users and the China-affiliated users, which were also algorithmically determined Louvain clusters [207]. These agents consistently engaged in cross cluster retweeting and tagging to draw audiences affiliated with both countries into shared conversations. Figure 3.5 shows the all-communication network of the US-China discourse and shows how bridgers straddle between Louvain clusters.

Repeaters echo content with little or no textual modification, recreating the same message as an original post rather than relying on retweets or shares. Their defining trait is excessive redundancy and replication, where they replicate the content of their own posts or other accounts, generating near-identical messages that saturate timelines. Often, they change only a single word or hashtag.

From a UCIA perspective, Repeaters are primarily defined by the Content pillar. Their activity is distinguished by the frequent posting of highly similar texts rather than their unique interactions. This frequent posting of highly similar or semantically similar texts aligns with how platform Algorithms assess engagement and topical velocity. Algorithms often detect repeated keywords, trending phrases and clustered temporal bursts as indicators of virality or collective interest, so the near-identical messages of Repeater Agents simulate organic social consensus.

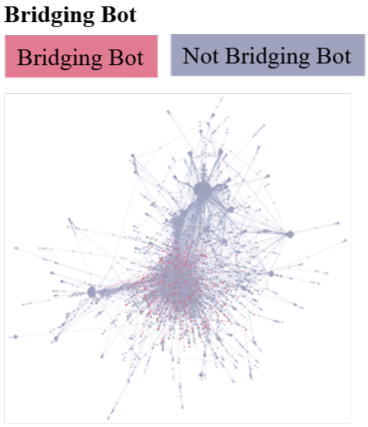


Figure 3.5: All-communication network showing how bridging bot straddle between Louvain clusters (published in [207])

Repeaters can be identified through their almost identical or slightly modified messages. In a study of 2021 discourse on human rights abuses in the Xinjiang Uyghur region, we found a total of 114 agents that were part of a repeater agent network [138]. We determined whether a tweet was part of a specific repeated sequence by first alphabetically sorting the messages, then tokenizing the tweet text, then performing a rolling comparison of tweets within a 48-hour period to check if a tweet had the same text as the tweet that was posted before it. We deemed an agent to be a Repeater if it had repeated tweets at least three times. The primary narratives that were promoted among the accounts were: refute of human rights abuse claims, spread of positive messages about Xinjiang and dissent against other activist groups.

Self-Proclaimers openly label themselves as Bots, usually by embedding the term "bot" in their usernames or profile descriptions. Self-Proclaimers adopt a rhetorical strategy of transparency. By making their artificial nature explicit, they disarm suspicion and present themselves as functional or entertaining agents.

From a UCIA perspective, they are characterized by their User information. They typically have the word 'Bot' in their user information, signaling that their purpose is not deception but functionality. Self-Declared Agents do not have any specific algorithmic interactions, except being shown by the algorithm to other users that they are likely to be automated users. These agents have a wide range of behaviors, and typically also take on the behavior of other types of agents.

Synchronizers acts together with other automated accounts, producing synchronized bursts of collective action. This is visible through near-simultaneous posting, sharing, or using the same hashtags within narrow time windows. The impact of Synchronizers is amplified through algorithmic reinforcement. The synchronized activity creates trending narratives and engagement. Algorithms can interpret these sudden spikes in activity as organic virality, and promote the coordinated content into trending lists, recommendation feeds or search suggestions.

There are many ways which agents can synchronize with each other. For example, agents can use the same URLs, use same series of hashtags, and tag the same groups of people in their tweets, as we have profiled in our work [203]. Such strategies are also termed as referral,

semantic and social coordination. Another artifact that can be used in synchronization is the tweet itself: using similar phrases in the tweet text [216], or using similar images in the tweet media [217].

Chaos-Creators create disruption and confusion within online conversations. They derail threads, divert attention and destabilize discourse. These agents excel in introducing noise, provocation or misleading signals that fracture the coherence of discussions and make it difficult for communities to sustain constructive dialogue. Their role is not to simply spread content, but to undermine the stability of the conversational environment itself.

From a UCIA perspective, Chaos-Creators are identifiable through disproportionately high values on influence maneuvers, such as distraction or disruption, which can be quantifiable measured through the BEND maneuver framework [48]. Their strategies of inflammatory or contradictory posts, hijacked hashtags and rapid comment cascades can prompt bursts of algorithmic amplification, pushing the disruptive content into feeds and recommendation systems.

In our analyses, we observed Chaos-Creators operating through hashtag hijacking and topic derailment. In 2021 Indonesian discourse, coordinated agents systematically inserted 4-character, irrelevant hashtags into otherwise coherent conversations, contaminating the trending topics with noise and forcing fragmented discussion groups [70]. This behavior also aligns highly with the Distract maneuver of the BEND framework, where high values of the distract maneuver was observed from these set of agents [48]. We also found clusters of Chaos-Creators in a 2021 discourse about the human rights of the Xinjiang Uyghur region that used several different character artifacts like 4-letter artifacts, sets of punctuations and 5-character Chinese phrase blocks to throw off the discussions [138].

Announcers provide automated notifications. They publish updates on a schedule or when specific conditions are met. Their purpose is to surface new states (i.e., "it's now 5pm", "site X is down") with minimal human interactions.

From a UCIA point of view, Announcers are characterized by content regularity. This is identified by periodic posting patterns in their content, whether by time-separated periodic patterns, or by content-triggered periodic patterns. Their consistency and continued recency of posts are indicators of reliability and relevance. Announcers exploit algorithmic refresh cycles by maintaining predictable posting cadences to ensure their updates are perpetually visible.

An example of such announcers surfaced in our study of 2021 Indonesian discourse [70]. We discovered 4,081 and 2,522 agents that engaged in the discourse over the Palestine-Israel conflict and the latest alcoholic beverage ruling. Many of the tweets originating from these agents are templated tweets that shows the sympathy to the notion of Khilafah, an Islamic state in Indonesia. These tweets do sometimes include their opinions about the key events, but sometimes they do not. The template messaging is as such: "<Measure of Time>, <Alhamdulillah/MasyaAllah¿ saya melihat di kota Jakarta sudah banyak orang yang sadar khilafah, kamu gimana mention Twitter target account>". The English translation of the template, as translated by a native Indonesian speaker, is: "<Alhamdulillah /MasyaAllah> There's a lot of people in Jakarta this last <Measure of time> that is aware of Khilafah. How about you?<mention Twitter target account>".

Content-Generators are designed to produce original material for online ecosystems. From a UCIA perspective, the Content Generation Agent is characterized by the disproportionate use of original content such as original posts, replies or quote tweets, rather than retweets or re-

shares. Their operational signature lies in the volume of novel contributions relative to recycled material. The impact of a Content-Generator is amplified through the algorithmic privilege of novelty, where platform algorithms reward original and fresh material. This makes Content-Generators effective in seeding new narratives in the information ecosystem to grab attention, sometimes under the guise of spontaneity.

Information-Correctors are designed to identify and respond to false information by providing fact-checks, contextual clarifications or links to authoritative sources (i.e. government sites, official websites, institutional sites). Their primary rhetorical strategy is to counter falsehoods with corrective content and they position themselves as truth-oriented actors in the information ecosystem.

From a UCIA perspective, Information-Correctors are characterized by both their Content and Interaction pillars. On the Content side, they reference fact- checking websites or authoritative sources, and typically use phrases or markers that signify negation (e.g., "This is false", "fact check", "Correction"). On the Interaction pillar, they may sometimes publish stand-alone correction posts that summarize both the false claim and the corrected fact-check, or they may intervene directly by replying to or quoting the original post containing false information and correct them. Which interaction is used is a product of whether the Bot was designed for general broad-based information correction or targeted interventions. However, their efficacy and reach is tightly coupled with platform algorithmic dynamics. Algorithms prioritize recency and engagement, so well-timed information corrections can surface prominently from viral misinformation, but delayed interventions can also be buried or overlooked, especially when the misinformation has gotten extremely viral.

Mono-maniacs are agents that are specialized to post within a narrow topical domain, or a single genre. Rather than covering a broad range of issues, it persistently frames its content within a single genre, such as sports, music, religion, politics or health, to name a few. These Bots concentrate activity on one theme and carve out recognizable niches in online ecosystems.

From a UCIA perspective, a Mono-maniac is most clearly defined by the Content frame, because they disproportionately post or share content related to one topic cluster. These agents are reinforced through algorithmic alignment because they appear authoritative or embedded within a particular community of interest. We have observed such genre specific agents in political discourse, where these agents were activist accounts that advocated for an uprising or an action to end the Indian-Pakistan dispute during the Kashmir Black Day season [202].

Conversational Agents engage directly in dialogue with users, simulating interpersonal communication. They operate rhetorically and mimic the style of human interaction, aiming to build trust, foster rapport and sustain attention.

From a UCIA perspective, Conversationalists are primarily focused by their Content properties. Automated conversations tend to use more varied topical frames and produces messages with higher linguistic complexity than the average social media user[127]. These agents exploit platform algorithms by sustaining extended conversations, thus appearing prominently in comment threads and recommended interactions (i.e., "¡User¿ commented on ¡post¿."). The more consistently they sustain exchanges, the more the algorithm interprets them as socially significant nodes, thereby amplifying their presence and deepening their narratives within digital ecosystems.

Engagement-generators are designed to maximize interaction rates on posts by strategically

deploying rhetorical cues that elicit likes, shares and comments. They use targeted rhetorical strategies, particularly those that trigger emotional or cognitive biases among readers.

From a UCIA perspective, Engagement-generators are primarily defined by their Content properties. They often employ emotional language, humor or high valence expressions or emotionally-charged images. These agents manipulate the platform's algorithmic engagement loops. Since their posts incite reactions and engagement, their posts are interpreted by platform algorithms as organic popularity, and therefore fed into recommendation cycles that perpetuate both attention and influence.

Broadcasters disseminate news updates on digital platforms. There are two main ways that Broadcasters can post news. The first way is by being a news originator, where they post original reports or link directly to articles published by established news organizations. The second way is by being a news aggregator, to collect and redistribute articles or headlines from a single or multiple sources, providing a continuous stream of topical updates.

From a UCIA perspective, Broadcasters are characterized by both the User and Content pillars. User information often includes terms such as "news" in the user names or account descriptions, marking their rhetorical focus and associating them with topical credibility. Content-wise, their posts overwhelmingly consist of news headlines or links to external news articles. The visibility and authority of Broadcasters are heavily shaped by Algorithmic mechanisms. The headline-driven and structured news posts are sometimes rewarded by platform algorithms and placed higher in the "trending" or "latest" feeds. However, since the posts of Broadcasters do not necessarily optimize for engagement, their posts can be lost in the deluge of posts.

We observed that approximately 3.65-4.87% of the cyber social agents are broadcasters in the US-Chinese balloon discourse. These agents are scattered throughout the communication network, and thus have low eigenvector, betweenness and total degree centrality values [207]. To bring across their messages, these broadcasters mostly use the Enhance and Explain BEND information maneuvers to elaborate on the news and support their claims [48].

Figure 3.6 shows the methodology we used to identify news broadcasters. We identify these agents in two manners: through substring matching and via a machine learning classifier. The first manner relies on the explicit expression of a broadcaster bot from the user's profile information. Broadcasters are identified by users that contain the word "news" in their profile (i.e., user name, screen name, description) through regex substring matching in Python.

When the users do not explicitly state the word "news" in their profile but tweet news headlines, we use a machine learning classifier to identify if they were broadcasters. We trained a random forest machine learning classifier on 100,000 examples of news headlines written between 1 January 2010 to 31 January 2020 from the News on the Web corpus [229] and 100,000 tweets from 3,298 human users. The classifier achieved a 92.3% accuracy. This classifier then takes in a tweet and returns a probability of whether the tweet is likely to be a news headline or not. By extension, if a majority of a user's tweets are news headlines, the user is likely to be a broadcaster agent. We used a 90% threshold for this determination of majority: if a classifier reflects that 90% of a user's tweets are similar to news headlines, then we reclassify the user as a broadcaster.

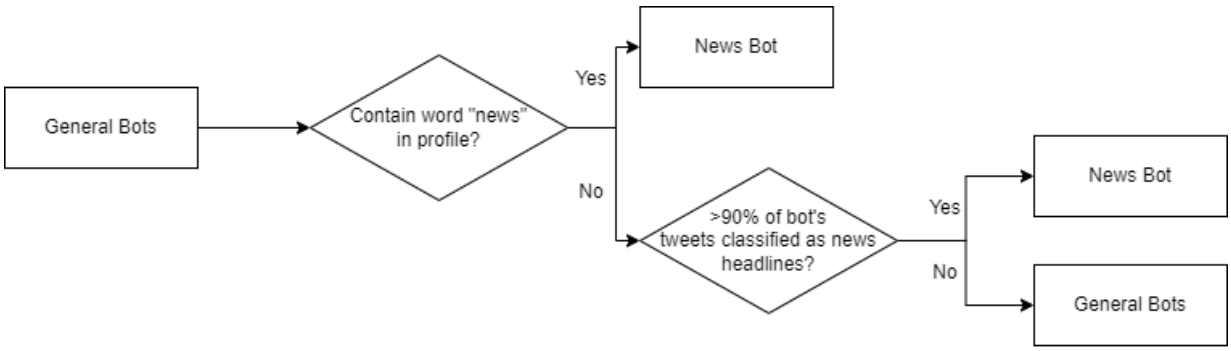


Figure 3.6: Methodology for identifying Broadcasters (published in [207])

Archetype	**Definitions**	**Properties**			**Algorithmic Effect**
		User	**Content**	**Interaction**	
Behavior-Based Archetypes					
Social Influencer	Attempt to influence other users' perceptions		High usage of emotional and BEND cues[1]	Excessive replies	Sustain conversational activity
Amplifier	Boost narrative themes and manufacture support			Excessive sharing (i.e., retweets, shares) pattern	Mimic virality
Cyborg	Exhibit both human- and bot-like activity	Frequent changes in bot classification with high change in bot probability score			Sustained by Algorithm

Archetype	Definitions	Properties			Algorithmic Effect
		User	**Content**	**Interaction**	
Bridger	Connects groups of users	High Bridge score from the BEND metric		Frequently tag multiple people from different social identity groups, or straddle between multiple network clusters	Mimic broad resonance with interactions from multiple user clusters
Repeater	Repeats posts, sometimes keeping most of the content the same		Frequent posts with similar texts		Mimic organic consensus
Self-Proclaimer	Outwardly declare themselves as a bot	Presence of word "bot" in user information (i.e., username, screen-name, description)			Shown by algorithm that it is likely to be a Bot
Synchronizer	Coordinate with other bot accounts	High coordination index		High number of bots that account is coordinating with	Amplified through algorithmic reinforcement
Content-Based Archetypes					

Archetype	Definitions	Properties			Algorithmic Effect
		User	**Content**	**Interaction**	
Chaos-Creator	Sow chaos and discord	Higher than average BEND values[1]			Amplified through defensive or argumentative responses by other users
Announcer	Announce information		Periodic posting patterns, or posting with templates		Exploit algorithmic refresh cycles by maintaining predictable posting cadences
Content-Generator	Generate content for the online ecosystem		Majority of posts are original content rather than shared content		Amplified through the algorithmic privilege of novelty
Information-Corrector	Correct information		Presence of negation	References to fact checking website	Dependent on the timing, recency and engagement of the information correction
Mono-maniac	Posts mostly on a singular topic		Majority of the posts on a single topic frame		Appear as trusted sources of a topic

Archetype	Definitions	Properties			Algorithmic Effect
		User	**Content**	**Interaction**	
Conversationalist	Carry out a conversation with other users		Content has Agent than average reading difficulty and varied use of topic frames		Appear to be socially significant nodes because of their role in sustaining conversations
Engagement-Generator	Incite engagement from other users		Use of emotional cues and high emotional valence		Posts incite engagement, thus appearing as organic popularity
Broadcaster	Post news updates	Presence of the word "news" in user metadata	Majority of posts are news headlines		Heavily shaped by algorithmic mechanisms

Table 3.4: Archetypes of Cyber Social Agents [1]BEND cues are signals of information maneuver tactics, derived from [48]

3.5 Duality of CSA Archetypes

The automation technology used to develop a CSA is neither good nor bad, but rather, it is the way the technology is employed that makes it good or bad. This is the duality of CSAs. Each CSA archetype can exhibit either good or bad qualities depending on its environmental conditions, which includes its narrative and its social interactions. Instead of jointly flagging all automated social media agents as bad, each archetype should be evaluated according to their content and interactions. This section presents guidelines for determining the goodness of a CSA. These guidelines are determined by first surveying the literature for studies of good and bad CSAs, which are listed in Table 3.5. Thereafter, we manually tagged the characteristics of the studies into the User-Content-Interaction framework (defined in chapter 1). This goodness of archetypes frame is presented in Table 3.6

Archetype	Usage for Good	Usage for Bad
Behavior-Based Archetypes		

Archetype	Usage for Good	Usage for Bad
Social Influencer	change people's views towards vaccination (anti to pro) [204], influence purchase intentions [270]	change people's views towards vaccination (pro to anti) [204], opinion manipulation [59]
Amplifier	amplifying government or regional news during crisis situations [55]	amplify politically divisive narratives and narratives that target other countries [78, 137], disseminate computational propaganda through "flooding the zone" technique [232], amplify hate speech, amplify the spread of misinformation, inflate popularity of candidates during election season [185]
Cyborg	alleviate workload of social media managers, politicians and influencers [108, 222]	trigger and initiate activism [165, 175]
Bridger	political commentators that aggregate information across multiple parties [220]	cross-cultural social marketing [24], information dissemination across groups for political manipulation [207]
Repeater	constant product promotion and advertisement [162]	constant sharing politically divisive narratives [137], spam content [8]
Self-Proclaimer	useful and creative bot accounts that label themselves as bots [8]	
Synchronizer	synchronization in broadcasting news to region specific accounts [175], raise citizen concerns on government actions [209]	social botnets in political conflicts [1], participation in online influence campaigns [145]
Content-Based Archetypes		
Chaos-Creator	bring attention to citizen concerns like disrespect of muslim ideology [209], climate change [166], organize resources during chaos in crisis [128], organize earthquake reporting from twitter [22, 69]	infiltration of an organization [84], sow discord to polarize opinions [166]

Archetype	Usage for Good	Usage for Bad
Announcer	promotion of product launch [162], announcing locations to get help, e.g. vaccine, crisis [56], earthquakes [22]	broadcasting propaganda messages [16], call for action against political incumbents [210], publicize protests locations and information [89]
Content-Generator	warning for natural disasters [121], crisis communication and emergency resource deployment [129], editing Wikipedia entries [289]	generation of harmful racial and ideological content [93], hijack conversations to create ideological polarization [70], digital marketing for e-commerce [328]
Information-Corrector	bots on Reddit moderate conversations [123, 139], bots on Wikipedia prevent vandalism [289]	correct true information into false ones [91]
Mono-maniac	brand establishment and amplification [222, 292], digital campaigning [220] through candidate promotion [185], religious prayer and worship generation [233]	political manipulation [24], use religion to influence audience [233], radicalization and recruitment for extremist groups [30, 178], political accounts that systematically delete content [8]
Conversationalist	provide emotional support during stress [245] and grieve [151], provide help in frequently asked questions for e-commerce sites [196]	converses with racist and derogatory terms [201], exacerbates stereotypes, and gender/ race divides [311]
Engagement-Generator	provide humor [297], increase perceived product value and user engagement in e-commerce [83]	artificially inflate engagement or popularity of malicious users, e.g. disinformation spreaders [106], creation of viral engagement of bad content using botnet [324]
Broadcaster	news production, dissemination and interaction with audiences in current media environment [130, 163], active news promoters [7], curate target content from multiple information streams [163]	spread disinformation news and fake news [4, 263], spread news through click baits to earn money [7]

Table 3.5: Observations of the Duality of the Cyber Social Agents

	Good Agent		Bad Agent	
	Content	**Interaction**	**Content**	**Interaction**
Behavior-Based Archetypes				
Social Influence Agent	good content, positive emotion cues	bad content, negative emotion cues	cites bad information sources	
Amplifier Agent	good content	high number of retweets	bad content	cites bad information source, high number of retweets
Cyborgs	good content	negative change in message tone	bad content	cites bad information source
Bridging Agent	good content		bad content	cites bad information source
Repeater Agent	good content, posts same message		bad content, posts same messages	cites bad information source
Self-Declared Agent	good content		bad content	cites bad information source
Synchronized Agent	good content		bad content	cite bad information source
Content-Based Archetypes				
Chaos Agent	good content	constructive BEND maneuvers	bad content	destructive BEND maneuvers, cites bad information source
Announcer, Content Generation, Information Correction Bot, Genre Specific, Conversational, Engagement Generation, News Agents	good content		bad content	cites bad information source

Table 3.6: Duality of Cyber Social Agents

3.6 Conclusion

Bot detection provides the computational grounding to identify automated actors on social media, and the theorization of CSA expands its scope into a socio-technical paradigm. Beyond identifying the automated actors, their nature, personas must be understood to fully capture their role in digital ecosystems and harness their capabilities. The typological approach of CSAs allows us to capture the nuanced roles that these social agents play in shaping information diffusion and discourse, providing a richer foundation for both theoretical analysis and practical interventions. By considering both "good" and "bad" manifestations of each CSA persona, the CSA framework emphasizes that automation is not a purely negative phenomenon, but instead highlights its dual potential in a socio-technical system. However, the current positioning of agent archetypes do come with some **limitations:**

1. The current case study suggests that the archetype the agent takes on is static, but agents can change their level of automation and their behavior depending on the context or their goals [110].
2. The current goodness duality schema classifies users solely into two categories, good or bad. In reality, the lines between these two classes are not as distinct, and are often mixed with cultural and regional factors.

Future work involves formulating archetype-specific policies to regulate the social media environment of bad agents and harness the strengths of each archetypes for social good. This involves large-scale and longitudinal studies across multiple social media platforms to study the effects of each archetype, and also simulation techniques to model the impacts of interventions.

Chapter 4

Nature of Cyber Social Agents

4.1 Introduction

Bots and humans exhibit systematic differences in how they affiliate themselves, communicate, and interact online. Many studies focus on posting frequency, noting that Bots often produce almost twice as many posts as humans[279] and can be active during nocturnal hours[112]. However, posting volume is only one marker. Bots and humans diverge systematically across other dimensions of online behavior. This chapter explores the differences between the two species across five dimensions: social political representation, narrative expressions, motivations & agencies, linguistic signatures, and use of cognitive bias triggers.

This chapter investigates the following guiding **research questions**:

1. What social and political groups do CSAs associate with?
2. What are the unique narrative expressions of CSAs?
3. What are the defining motivations and agencies of CSAs?
4. What are the defining linguistic signatures of CSAs?
5. What are the defining patterns of the use of cognitive bias triggers of CSAs?

4.2 Related Work

Bots and humans can have identifiable sets of differences. Some of these sets of differences are the narrative expressions, the social interactions and geography, and the use of cognitive bias triggers.

Social interactions and geography Social interactions between agents in the cyberspace are reflective of offline communities and regional boundaries. For example, visualizing the telecommunication map reveals the physical boundaries between Wales, England, and Scotland [250]. Many tools allow us to overlay the social interactions and geographies, which we coin as "social cyber geography". Location-based social networks provide a plentiful source of geolocated connections between individuals [12]. Analyzing such social cyber geographies provide us with

a better overview of the ongoing social interactions, and help us have a more precise reponse to events. For example, creating a social cyber geography of communications during the coronavirus pandemic revealed the level to which the countries were affected and indicated where resources were needed [79, 248]. More so, in the area of social cybersecurity, which focuses on understanding how behavior is influenced by relationships and communities and the analysis of information maneuver campaigns [48], creating a social cyber geography of these campaigns reveals patterns of global alliances and threats. For example, the social cyber geography of bots and their stances during the 2022 Russian-Ukraine war revealed the distribution of political stances on X towards each country [265].

Narrative expressions Prior research shows that CSAs and humans can differ systematically along the dimensions of the narratives that agents promote and the frames they adopt. Automated agents tend to concentrate on a narrow set of narratives that are rapidly injected into conversations, often with high temporal regularity and limited contextual adaptation. In political and crisis-related discourse, they disproportionately promoted a small number of dominant frames rather than engage in narrative elaboration or deliberation[263**?**]. Humans, in contrast, are more likely to contribute heterogeneous narratives that incorporate personal experience and situational contexts[39].

Hashtags are another social media identifier that distinguishes narrative expressions. Hashtags serve as topical markers that enable narratives to be aggregated, amplified and algorithmically surfaced. Past studies show that CSAs use hashtags more frequently and repetitively than humans, often attaching multiple hashtags to a single post to maximize visibility and narrative reach[42, 277]. Humans tend to use hashtags more contextually as conversational signals or identity markers[39].

Motivations & Agencies In online environments, motivation refers to the underlying goals driving behavior, while agency refers to the capacity of an actor to intentionally execute actions through the technological affordances. On social media, motivations and agencies are rarely directly observable and must instead be inferred from behavioral traces and interaction patterns. To make sense of motivations and agencies of online users, several analytic frameworks have been proposed that map observable online behaviors to underlying strategies of influence. One prominent example is the BEND framework, which conceptualizes online influence into sixteen maneuvers, each associated with specific behavioral signatures and strategic intent[48]. Other complementary frameworks similarly seek to bridge observed behavior and latent intent. The DISARM framework adapts military doctrine to classify information campaigns according to strategic objectives and operational phases, emphasizing the intentional design behind influence activities[283]. The SCOTCH framework emphasizes agency as an emergent property of a targeted pipeline of influence actions[38].

Use of cognitive bias triggers Social media platforms have billions of posts that circulate daily, so users have to rapidly decide what to read and engage with. These decisions are rarely the product of careful deliberation, but are often guided by mental shortcuts to navigate the overwhelming digital information [187, 290]. Embedding such cognitive bias triggers in information

can activate judgment heuristics, and influence user perception, memory and behavior [243].

Cognitive bias triggers are already exploited extensively in marketing and political messaging. For example, emotionally charged language are used to promote anti-vaccination campaigns [303], ideological and religious issues are propagated with constant repetitive messages [70], and online advertisements often use expert endorsements to lend credibility [316]. Such strategies are effective because they activate biases like authority, availability, and affect, thereby bypassing deliberative reasoning. Recent frameworks on the psychological factors behind viral online content have also examined this mechanism, showing that content is more likely to be attended to and propagated when it resonates emotionally, signals social identity, or offers morally salient information [249].

4.3 Social Political Representation

We coin the term **Social Cyber Geography** as the physical geography of the social communities that arise from the cyber realm. We combine geographical analysis and natural language processing methods in a multidisciplinary methodology to construct Social Cyber Geographical (SCG) maps of the 2020-2021 Coronavirus data. Due to resource and size constraints, we only analyzed data of the first five days of each of the twelve months of the year. These maps summarized the bot activity across countries, time, language and economic indicators.

We first used section 2.4 to extract bot agent from the data. Then we used a social location identified model to infer the location of each bot agent, before constructing maps that represent bot activity. This methodology is illustrated in Figure 4.1.

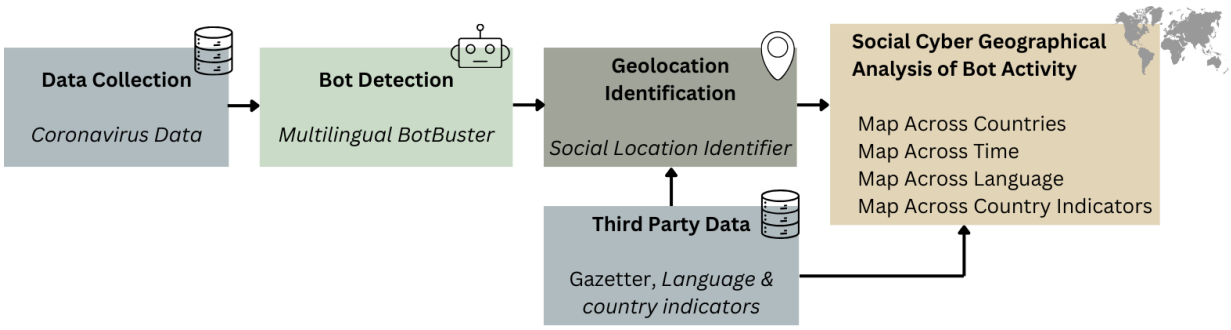


Figure 4.1: Methodology used to construct Social Cyber Geographical Analysis of Bot Activity.

Social Geographical Identifier To identify the country that the agent affiliated with, we built the Social Geographical Identifier module. Current methods of geolocation typically rely on the Nomantim API or Large Language models[120, 207], but these methods can be slow because they rely on API calls. Therefore, we developed a gazetter-based comparison method, which was fast running and returned up to 60% of location coordinates. This method relies on a pre-consolidated gazetter that we grew from the Geonames geolocation database, and other open-sourced databases: (1) the geonames and geolocation (latitude, longitude) of cities in the world with a population greater than 1000 people `https://public.opendatasoft.`

com/explore/dataset/geonames-all-cities-with-a-population-1000/table/?flg=en-us&disjunctive.cou_name_en&sort=name and (2) the geolocation of countries and their abbreviations https://github.com/annexare/Countries. This gazetter method also has an advantage over the other methods: the lack of reliance on API calls allows for the method to be run on a CPU and is both cost-effective and time-effective.

This Social Location Identifier script first used the Stanford Named Entity Recognition parser [92] to extract location words from the agent's self-written user description. Next, the script performed fuzzy matching of these extracted location words against the pre-consolidated gazetteer, and retrieved the closest-matched location by Levenshtein distance from the gazetteer and the location coordinates.

Social Cyber Geographical Map Across Countries After identifying the countries that each tweet was affiliated with using the Social Location Identifier, we grouped the tweets by country, and calculated the bot percentage for each country. Then, we plotted a geographic heat map to reflect the percentage of bot users in each country present in the data.

By segregating the social bot discourse by their geography, we found that bots were present in almost every part of the world that discussed the coronavirus pandemic. Figure 4.2 showed that regardless of the country, the average proportion of bots is approximately 20%. This finding mirrored a slew of past work that estimated the proportion of bots in country-specific events to be about 20%: [214] analyzed $\sim$200 million users and found that the bot volume across events are $\sim$20%, with the percentage increasing up to 43% during the US Elections; [254] suggested that $\sim$30% of the users were bots; [35] identified that about 20% of the 2016 US-election-related content came from bots; and [295] claimed that the percentage of bots on X was between 9% and 15%, and that industry estimates were up to 20%. In fact, Elon Musk, current Executive Chairman of X, claimed that at least 20% of the users on X were bots [264].

Social Cyber Geographical Map Across Language From the tweet's metadata, we extracted the language of each tweet. This language annotation was determined by X's internal machine learning algorithms. We then grouped the tweets by language and calculated the bot percentage per language. Then we plotted a bar chart reflecting the languages that have the most highest average bot proportion, where we used the mean as the averaging function.

This language map characterizes the nature of discussions on X, and is presented in Figure 4.3. Asian and European languages had the highest percentage of bots. Among Asian and European languages, those with the highest percentage of bots were: Thai (59.5%), Japanese (28.3%), Tamil (22.6%), Portuguese (18.7%), Greek (28.9%), Lativian (37.5%). In comparison, the bot percentage of the English language was 14.7%. This statistic was also indicative of the proportion of human speakers of each language. For example, there are approximately 1.5 billion English speakers, while only 39.9 million Thai speakers. Therefore, there are more humans than bots that use the English language than the other languages, which makes the percentage of bots that predominantly tweeted in English lower. The language in which the bot wrote in reflected its target audience – users who can read the language. Bots write different narratives in different languages, revealing how they were used globally for strategic communications. The high proportion of bots that used Asian and European languages indicates possible signs of narrative

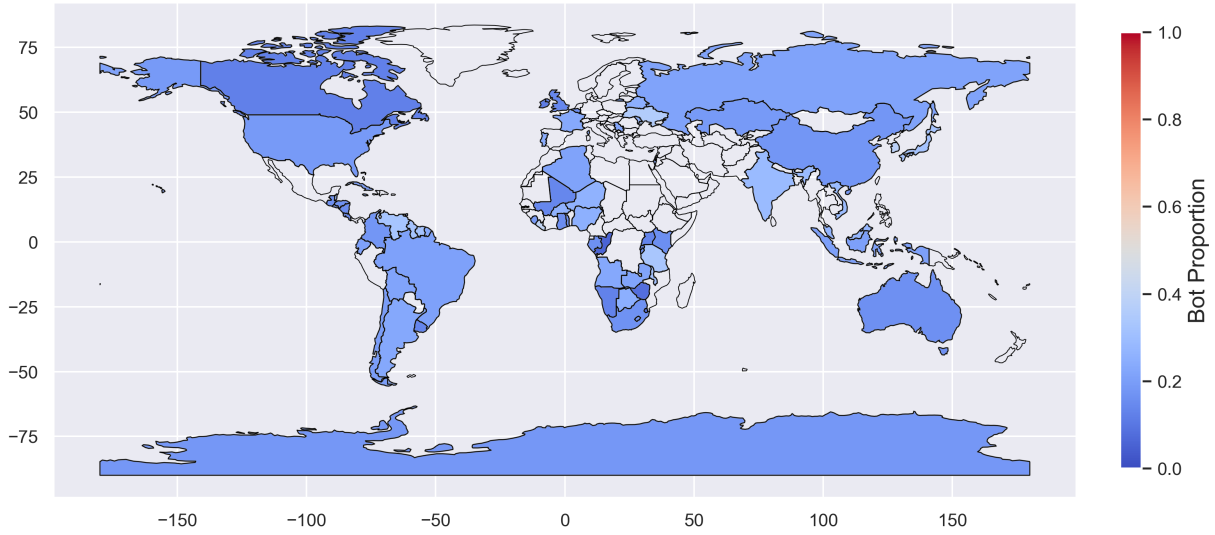


Figure 4.2: Social Cyber Geographic Heat Map of the average (median) percentage of bots affiliated with each country, across the entire data. White areas indicated that there are no bots present in the data we collected. On average, the number of bots affiliated to each country was $\sim$20%.

massaging, which warrants deeper investigations in future work.

Social Cyber Geographical Map Across Dominant Language We retrieved each country's dominant language from the Wikipedia page [1]. For the set of bots that were affiliated with each country, we identified how many of them had tweets written in the same language(s) as the country's dominant language. We note that some countries have multiple dominant languages, as per the Wikipedia article. We then plotted a geographic heat map to reflect the prominence of bots that post in each country's dominant language.

As a deeper dive into language distribution, we compared the bot distribution against the dominant language of a country, which is presented in Figure 4.4. Despite the changing distribution of the bot affiliation dominance, the distribution of language used was rather consistent across the year-long data. On average, at least 80% of the bots affiliated with a particular country used the dominant language of the country. This indicated that bots targeted language diasporas across countries. There were nine countries where less than 80% of the bots affiliated with the country used the country's dominant language. When the dominant language was not used, the bots used Asian and European languages (Table 4.1), which agreed with the observation that Asian and European languages have the highest percentage of bots.

Social Cyber Geographical Map Across Economic Indicators We correlated bot percentage with economic indicators of the country's population and the GDP (Gross Domestic Product). We retrieved the statistics for the GDP and Population per country from the World Bank. We then

[1] `https://en.wikipedia.org/wiki/List_of_official_languages_by_country_and_territory`

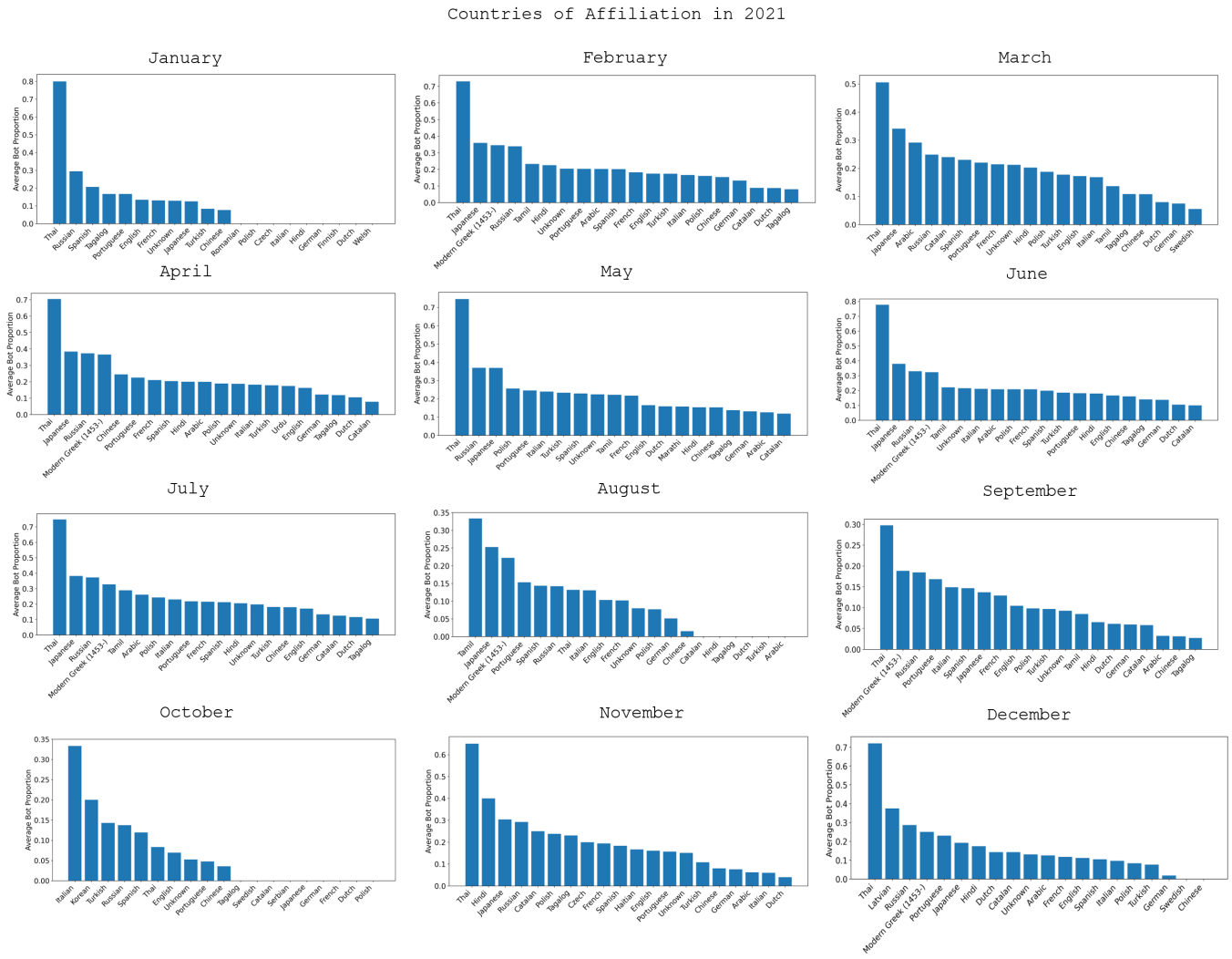


Figure 4.3: Social Cyber Geography of Bot proportion against commonly used languages by month. Asian and European languages were popular languages that bot-authored posts were written in.

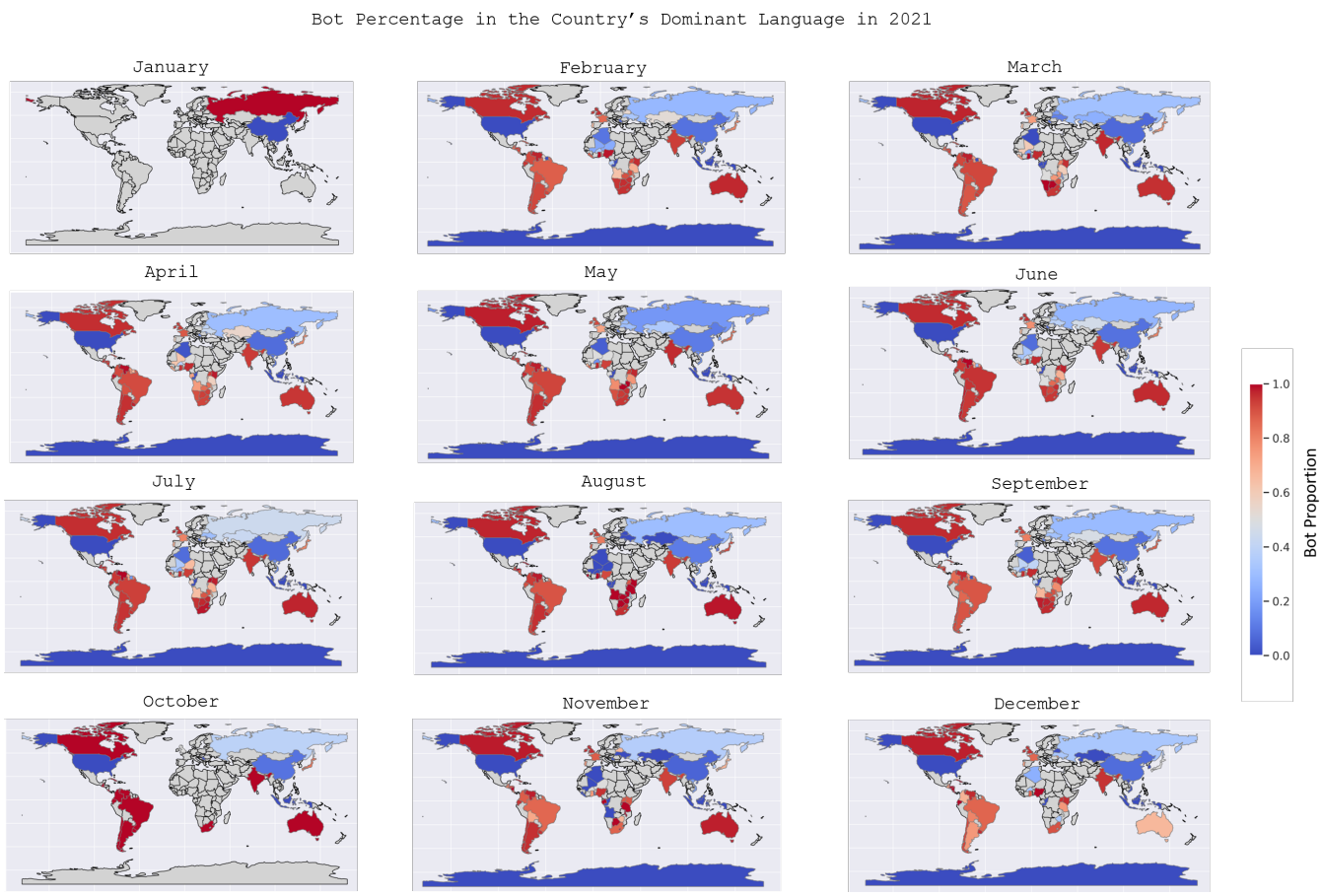


Figure 4.4: Social Cyber Geographical heatmap of bot proportion authoring posts in the country's dominant language by month. On average, 80% of the bots affiliated with a country used the country's dominant language.

Country	Languages
United States	English, Spanish, Chinese, French, Undefined
Russia	English, Russian, Thai
China	English, Chinese, Thai, Spanish
Antarctica	English, Thai, Undefined
Kazakhstan	Russian, English
Indonesia	Indonesian, English, French, Undefined
Niger	English, Spanish
Mali	English, Spanish, Tagalog, Hindi
Algeria	English, French

Table 4.1: Languages that posts were authored in for the countries where $< 80\%$ of the bot population wrote in the dominant language. The countries were ordered in descending order of the frequency of the dominant language used

ran two linear regressions: the first of the percentage of bots affiliated with each country against the GDP (percentage of bots $= \alpha(\text{GDP}) + \gamma$), and the second of the percentage of bots affiliated with each country against the population of the country (percentage of bots $= \alpha(\text{population}+\gamma)$). We found no significant correlation between the bot percentage and the GDP or population of the country. The R^2 value for bot percentage and GDP was 0.021. The R^2 value of bot percentage and population is 0.022. This result is reflected in Figure 4.5. The Social Cyber Geographical map of bot distribution is thus independent of country-based indicators or the population of the country.

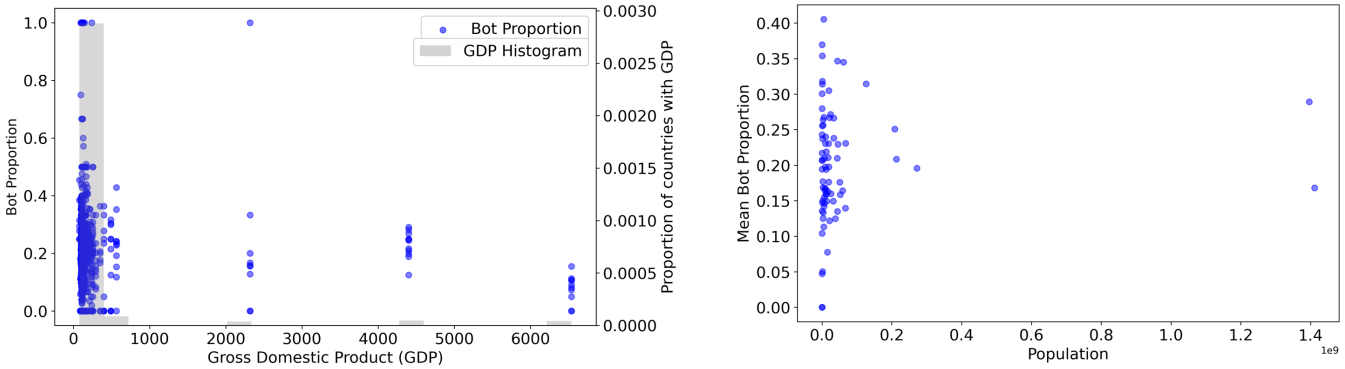


Figure 4.5: Social Cyber Geographical Map of Mean Bot Proportion vs. (a) GDP ($R^2 = 0.021$) and (b) Population of country ($R^2 = 0.022$). This map indicated that bot distribution was thus independent of country-based indicators.

4.4 Narrative Expressions

We study the use of narrative expressions in two ways. The first is via the use of sets of phrases, and the second is the use of the set of hashtags.

Phrases Analyzing sets of phrases allow us to examine the discursive nature of narratives. We studied tweets posted during the 2023 US-Chinese balloon discourse. These tweets contained the search terms #chineseballoon and #weatherballoon and were authored from 31 Jan 2023 to 22 Feb 2023. In total, we collected 1,192,445 tweets from 121,048 unique users.

We first geographically profiled the users using the Geographic Location Identifier script that we constructed in section 4.3. This segregates the users into three major locations that are relevant to the incident: the United States (US), China and the rest of the world.

Then, we performed narrative analysis. We used a combination of textual and emoji analyses to understand narratives put forth in this event. For textual analysis, we used a topic modeling method to aummarize the themes that are being posted within the text of the tweets. We first preprocessed the text to remove hashtags, @mentions and URLs, retaining only the raw text. We also removed stop words, which are words unimportant to the general meaning of the sentence. Then, we used the Python SkLearn CountVectorizer module to convert words in text to token counts. Finally, we used the Python wordcloud libary to plot wordclouds that represented the frequency of words, in which a larger size of words in the cloud indicated a higher frequency of a given word within a group of texts.

Figure 4.6 illustrates the differences between the narratives put forth by accounts geotagged for each region. Accounts tagged in the U.S. were focused on the location of the balloon and its This can be visualized by larger sized words of “Myrtle Beach”, “airships”, and “spotted over”. Accounts from the U.S. consistently referred to the balloon as “spy balloon” or “surveillance balloon”, leading to a high frequency of those words appearing in the word cloud. These terms demonstrated that accounts from the U.S. perceived the balloon as a threat. Accounts geotagged in China presented narratives on “MAGA” (Make America Great Again) and “SleepyJoe”. “MAGA” is a phrase associated with previous U.S. president Donald Trump in his political campaign, and “SleepyJoe” refers to a nickname that Trump invented for his political opponent Joe Biden in 2020. The presence of these phrases demonstrated that accounts geotagged to be from China could be attempts to steer attention away from the event and the country.

In terms of textual narratives between CSAs and human accounts, the two account types generally expressed the same ideas. However, CSAs typically exaggerate the phrases that humans express. For example, when the humans use phrases like “shoot” and “preparing [to go to war]”, CSAs emphasize the phrase “go war”.

We also analyzed the emojis. Emojis are pictograms that are used to convey ideas. Within our work, we analyzed the frequency of emojis and differences in usage and word representation across geographies. The analysis of emojis provided insights into a pictorial expression of thoughts from accounts across geography, enhancing our understanding into perspectives towards the event. We extract emojis from each text using Python’s regex and emoji packages. Then, we calculated the frequencies of each emoji per geographic group and account type.

Figure 4.7 presents the differences in emojis used by different account types through a frequency cloud. All three different geographic groups had a different focus within this event.

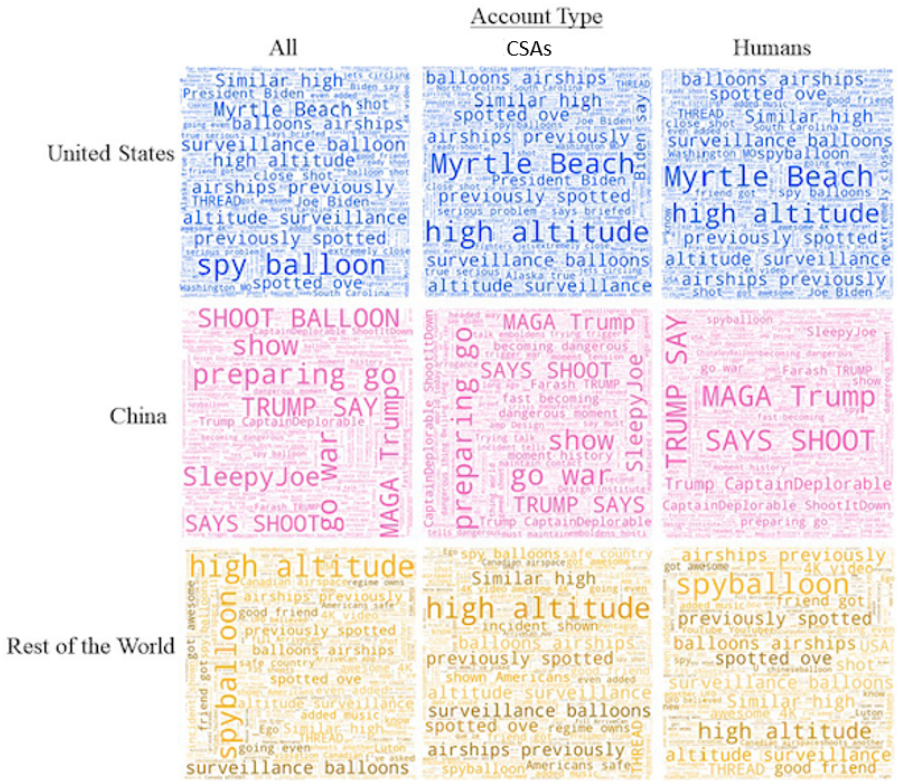


Figure 4.6: Word cloud built from texts of accounts geotagged to each country (Published in [208])

Figure 4.7: Emojis presented by account type and region (Published in [208])

Accounts from the U.S. had a high frequency use of alarm bell and the blaring alarm emojis, sounding an alarm over the presence of a balloon. Examples of the tweets are: "[alarm emoji] #ChineseSpyBalloon Similar high-altitude surveillance balloons airships previously spotted over Japan, Philippines [...]", "[alarm emoji] Major explosion in the air over Billings, Montana, reportedly where the Chinese balloon was." CSAs also used emojis differently as compared to humans. For example, for users affiliated with the US, humans had a higher use of the crown and sad face emojis, while CSAs had a higher use of the explosion and UFO emojis. CSAs geotagged to China and the rest of the world had a high usage of the explosion emoji, while humans used a large number of the laughing-with-tears emoji.

Hashtags Analyzing sets of hashtags used allows us to extract explicit narrative markers. We studied the differential use of hashtags in the online conversation of the 2020 Singaporean elections. This data is a subset of the 2020 Asian Elections dataset described in section 1.3. The Singaporean elections dataset was collected with general hashtags related to the election discourse (e.g., #GE2020, #sgelections2020) and more specific terms related to prominent parties (e.g., @PAPSingapore, @wpsg) and parliamentary candidates (e.g., @jamuslim). The data was collected between June 18 to July 17, spanning a week before the previous parliament was dissolved and a week after the election day. This dataset contained a total of 240,000 tweets from 42,000 featured users.

From this dataset, we used the BotHunter algorithm [32] to identify CSAs. Next, we performed hashtag analysis to characterize the messaging between CSAs and humans. Since hashtag usage can vary dramatically by raw scale, we obtained the ordinal ranking of hashtag usage for predicted bot and human accounts. We were particularly interested in hashtags which ranked more highly for bots than for humans, indicating disproportionate bot-driven focus on a particular message beyond the rest of the baseline conversation. In the context of an election during the pandemic, we also sought to assess the prevalence of hashtags related to COVID-19 as well as to voting in general. Hence, we examined the relative ranks of all hashtags containing the

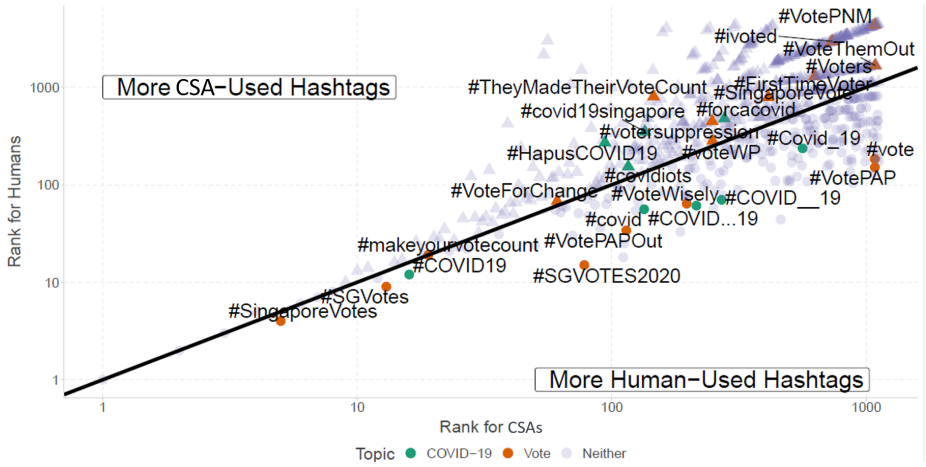


Figure 4.8: Scatterplot of hashtag ranking (log scale) based on mean usage by CSAs and humans. (Published in [294])

(case-insensitive) strings 'vote' or 'covid'.

Figure 4.8 plots the hashtags used by CSAs and humans. This figure ranks the hashtag according to their total usage by both account types. We notice that most hashtags cluster around the diagonal line. The diagonal line indicates where hashtags have the same ranking for both types of accounts. Therefore, we note that many features of the online conversation were prioritized by both CSAs and humans. This intuition is borne out by a Spearman's correlation test, which results in statistical significance ($\rho = 0.5926$, $p < .001$).

When we consider the hashtags above the diagonal line, we observe hashtags that had high rank for CSAs but low rank for humans. Most notably, we see in relation to "vote" the notion of "#votersuppression" and "votethemout". When humans use the hashtag "VoteWisely", CSAs use "VoteThemOut". This echoes our earlier observations from the study of narratives through phrases, that CSAs often exaggerate the narratives.

4.5 Motivations & Agencies

We situate our study of Motivations & Agencies within the BEND framework[48]. The BEND framework describes online actions as a set of narrative and community maneuvers, which are intentional acts carried out by social media users to achieve a communication goal. These maneuvers are further categorized as affirmative or adverse maneuvers, and are based on social-psychological theory and empirical evidence.

Kinetic Conflict We analyzed a subset of our collected X data on Russia-Ukraine conflict. This subset covered the period from January 2022 to November 2022, aligning with critical events like the Russian invasion of Ukraine in February 2022, the Russian advancement into Ukraine in May 2022 and the Ukrainian Kherson counteroffensive in August 2022. This dataset consisted of 4.5 million tweets. We segemented the data temporally into three timeframes: (1)

the Russian Invasion (08 February - 15 March 2022), (2) the Mid-point (15 May - 15 June 2022, and (3) the Kherson counteroffensive (20 July - 30 August 2022).

We first performed stance detection via hashtags [153] to segregate the tweets into pro-Ukraine and pro-Russian leaning tweets. For effective hashtag selection, we employed two criteria: the exclusivity of the hashtag to a specific stance pole, and the prevalence to ensure reliable agent stance detection. This process first involved sorting hashtags based on frequency to identify probable pro-Russian and pro-Ukraine stances. Then, the selected hashtags were examined further using network analysis to confirm their exclusive association with the intended stance and discover related hashtags. In this method, we identified over 2,000 unique hashtags to categorize the stances of tweets.

Then, we performed the BEND analysis using the ORA-Pro software separately on the communities segregated based on stance detection. We observe that pro-Ukraine CSAs leaned on the B's (Back, Build, Bridge, Boost) and E's (Engage, Excite, Enhance) maneuvers to rally, mobilize and clarify their pro-Ukraine stance. On the other hand, pro-Russian CSAs disproportionately used N's (Negate, Neutralize, Narrow, Neglect) and D's (Dismay, Distract) maneuvers to invalidate opposing voices and inject negative affect.

Figure 4.9 illustrates the evolution of pro-Ukraine narratives within the Twitter bot community. During the Russian invasion, there was a significant Boost and Build effort, using hashtags like `#westandforukraine` and `#istandwithzelensky` to foster solidarity and support for Ukraine, reflecting the BEND framework's principles of community building [48]. As Russian activities escalated at the mid-point, Engage and Excite maneuvers increased, aiming to make the conflict more globally relevant and counter Russian disinformation. During the Ukrainian counteroffensive, Dismay, Distort, and Distract maneuvers surged, with hashtags like `#putinisawarcriminal`, `#russiaisaterroriststate`, and `#PutinGenocide` challenging pro-Russian narratives and diverting attention from Russian messaging.

Figure 4.9 illustrates the evolution of pro-Ukraine narratives within the Twitter bot community. During the Russian invasion, there was a significant Boost and Build effort, using hashtags like #westandforukraine and #istandwithzelensky to foster solidarity and support for Ukraine, reflecting the BEND framework's principles of community building. As Russian activities escalated at the mid-point, Engage and Excite maneuvers increased, aiming to make the conflict more globally relevant and counter Russian disinformation. During the Ukrainian counteroffensive, Dismay, Distort, and Distract maneuvers surged, with hashtags like #putinisawarcriminal, #russiaisaterroriststate, and #PutinGenocide challenging pro-Russian narratives and diverting attention from Russian messaging.

Figure 4.10 illustrates the evolution of pro-Russian narratives within the Twitter bot community. During the Russian invasion, there was a focus on Negate and Neutralize maneuvers, using hashtags like #abolishNATO and #endNATO to diminish opposing narratives. Increased Russian military activity saw Distort and Dismiss maneuvers, skewing the narrative in favor of Russia by highlighting negative aspects of Ukraine, such as Nazi associations. In response to the Ukrainian counteroffensive, Dismay and Distort efforts surged, with hashtags like #westandwithrussia and #naziNATO, aiming to cause fear and discredit Ukraine while garnering support for Russia.

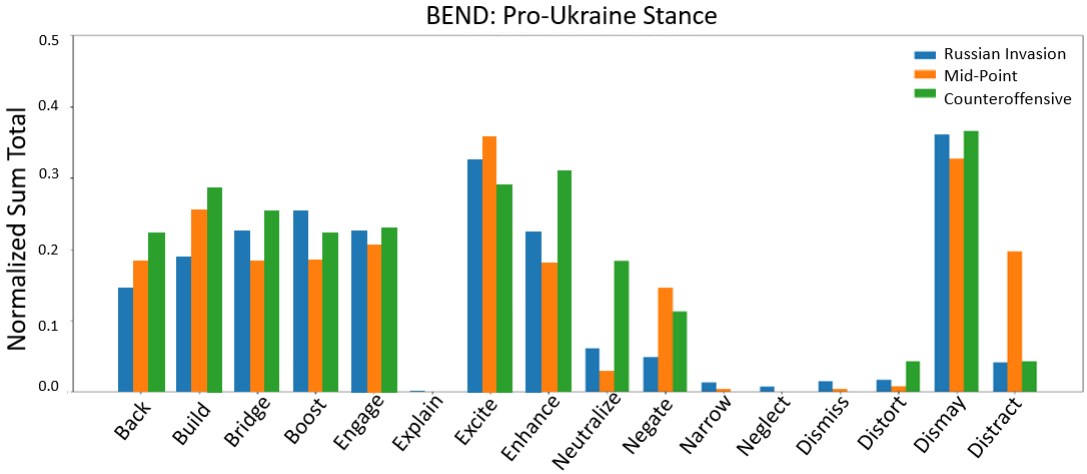


Figure 4.9: BEND maneuvers of Cyber Social Agents for Pro-Ukraine tweets (Published in [179])

Digital Diplomacy Beyond hot war, digital diplomacy exhibit similar mechanics. We studied tweets posted during the 2023 US-Chinese balloon discourse. These tweets contained the search terms #chineseballoon and #weatherballoon and were authored from 31 Jan 2023 to 22 Feb 2023. In total, we collected 1,192,445 tweets from 121,048 unique users.

We first geographically profiled the users using the Geographic Location Identifier script that we constructed in section 4.3. This segregates the users into three major locations that are relevant to the incident: the United States (US), China and the rest of the world. Next, we extracted three types of CSAs (General agents, News agents and Bridging agents) using the heuristics detailed in Table 3.4. Finally, we ran the data through ORA-Pro to analyze the usage and presence of BEND maneuvers. We analyzed the positive BEND maneuvers only, because diplomacy are generally image-enhancing endeavors [278]. In the BEND terminology, the positive maneuvers are the B- and E- maneuvers.

Figure 4.11 and Figure 4.12 showcase the differences in the information maneuver tactics used by the different types of bots, measured using the BEND framework. In terms of the B-maneuvers (Figure 4.12), in which the users attempt to manipulate the social network, we find that overall, users perform the Back maneuver the most, followed by the Build, Bridge then Boost maneuver. This indicates that the bots are more concerned with supporting other users through likes and shares, building larger groups through @mentions and hashtags, rather than increasing the linkages between members. General CSAs performed the most Back maneuvers, Bridging CSAs performed the most Build maneuvers, News Bots performed the most Bridge maneuvers and General CSAs performed the most Boost maneuvers.

In terms of the E- maneuvers (Figure 4.11), in which users attempt to manipulate narratives, we find that overall, users perform the Engage maneuver the most, followed by the Excite, then Enhance then Explain maneuver. This indicates that much of the narrative manipulation relies more on emotional appeal (Engage, Excite) rather than logical appeal (Explain). General CSAs performed the most Engage and Explain maneuvers, Bridging CSAs performed the most Enhance maneuver and News CSAs performed the most Excite maneuver.

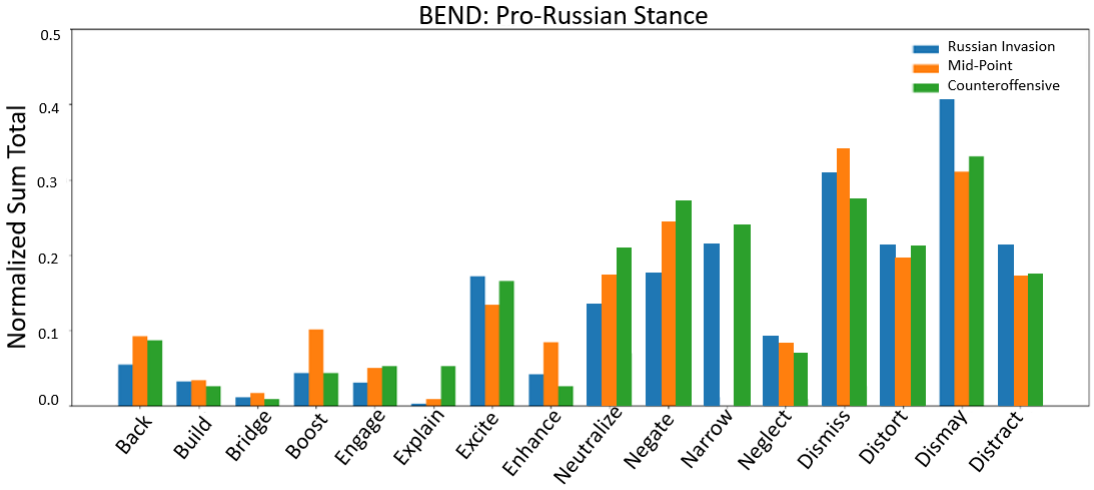


Figure 4.10: BEND maneuvers of Cyber Social Agents for Pro-Russian tweets (Published in [179])

4.6 Linguistic Signatures

Linguistic signatures are patterns of how an account uses language. This includes the user's vocabulary choices, emotional tone, syntactic structure and so forth. The linguistic signature is typically a measurable profile that persists across a user's posts and distinguishes one actor (or one actor type) from another.

One way to identify linguistic signatures is to extract psycholinguistic cues, which are features in a sentences that facilitate word recognition and meaning construction. Such cues include semantic cues (e.g., number of first person pronouns used, number of second person pronouns used, Flesch-Kincaid reading difficulty of the text etc.), emotion cues (e.g., number of abusive terms used, number of negative terms used etc), and metadata cues (number of URLs or hashtags used etc). In this thesis, we extract cues using the NetMapper software[2]. This software returns the count of each cue in the sentence, i.e., the number of words belonging to the cue in the tweet. There are three categories of cues that are generally derived from the software: semantic, emotion and metadata. Semantic and emotion cues are derived from the tweet text, while metadata cues are derived from the metadata of the user. Semantic cues include: first person pronouns, second person pronouns, third person pronouns and reading difficulty. Emotion cues include: abusive terms, expletives, negative sentiment, positive sentiment. Metadata cues include: the use of mentions, media, URLs, hashtags, retweets, favorites, replies, quotes, and the number of followers, friends, tweets, tweets per hour, time between tweets and friends:followers ratio.

Figure 4.13 presents the differences in the cues used by CSAs and humans, examined across large dataset repositories. From this figure, we see that there are consistent differences in how CSAs and humans use the cues. For example, across all events, CSAs use significantly more abusive terms and expletives. CSAs also tweet more than humans. On the other hand, humans use more first person person pronouns, positive sentiment, and media (i.e., images, videos).

[2]`https://netanomics.com/netmapper/`

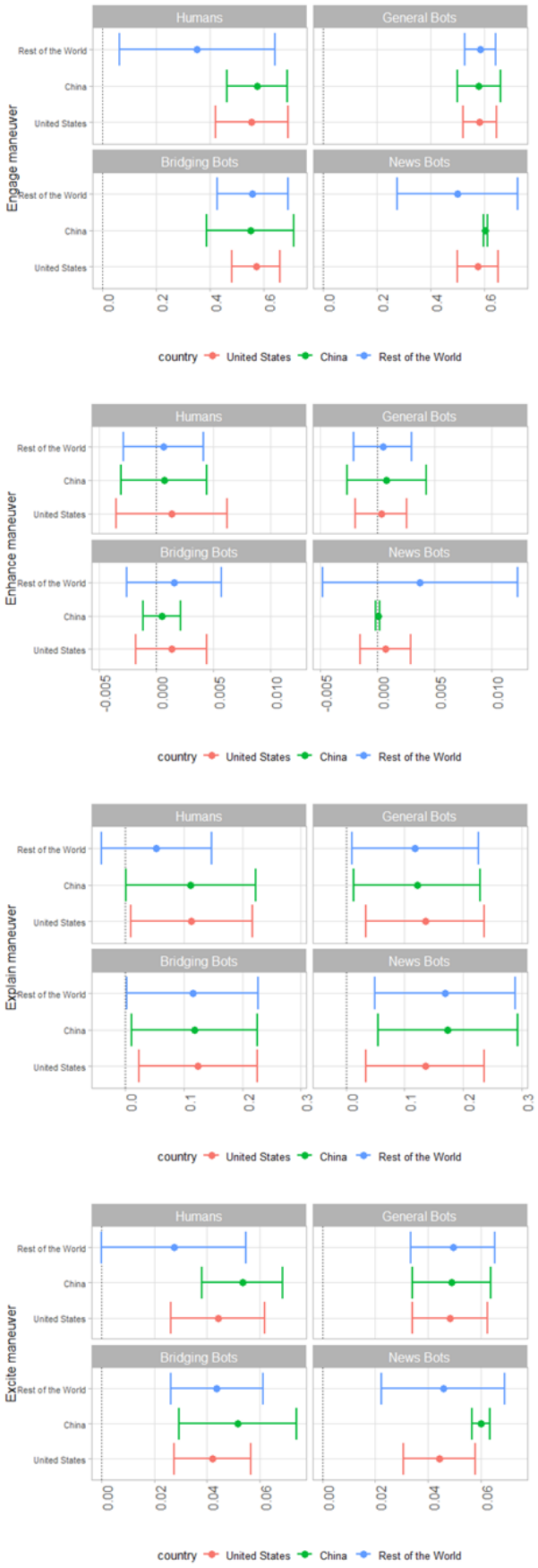


Figure 4.11: Distribution of the E- Information Maneuver Metrics (Published in [207])

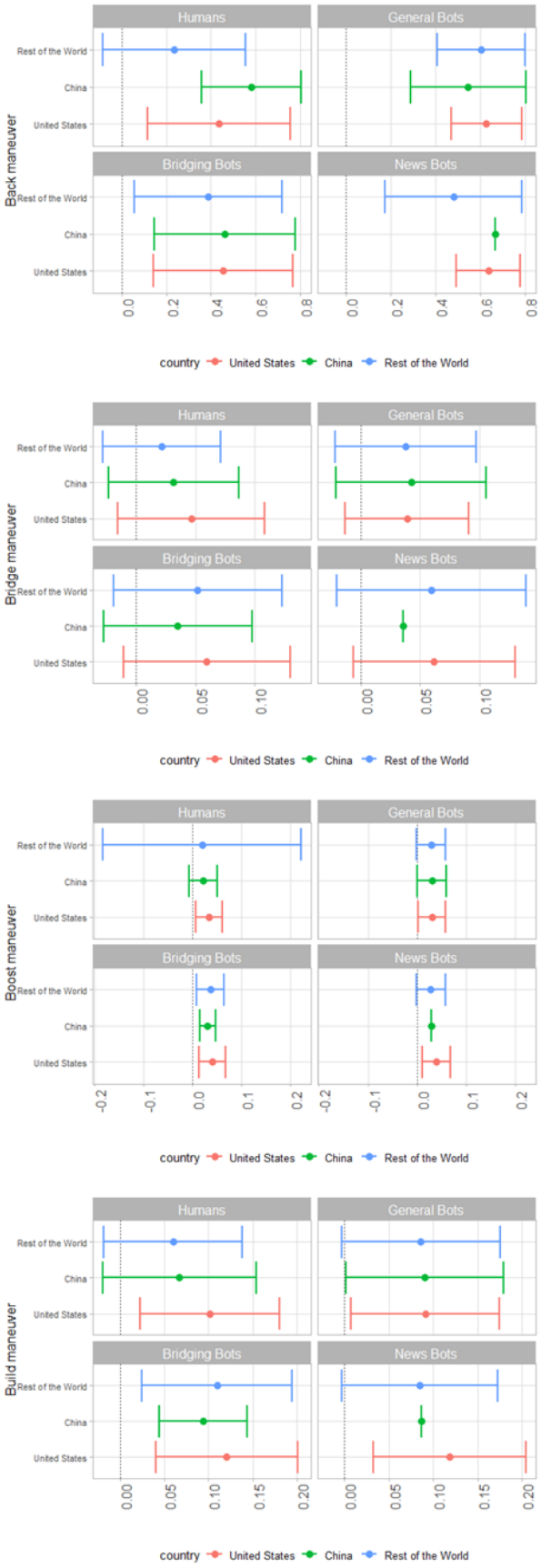


Figure 4.12: Distribution of the B- Information Maneuver Metrics (Published in [207])

Humans tend to quote and reply to tweets, while bots tend to retweet.

The cue distribution across events are generally consistent, but some events look different. In general, humans use more sentiment cues. However, in the two elections (US Elections 2020 and Canadian Elections 2019), CSAs used more sentiment cues. This reveals a deliberate attempt to use CSAs during the election seasons to polarize online sentiments. Prior research also shows that automated bot agents can be highly negative during the election season [277] and can express hugely different sentiments when mentioning different political candidates [8, 204].

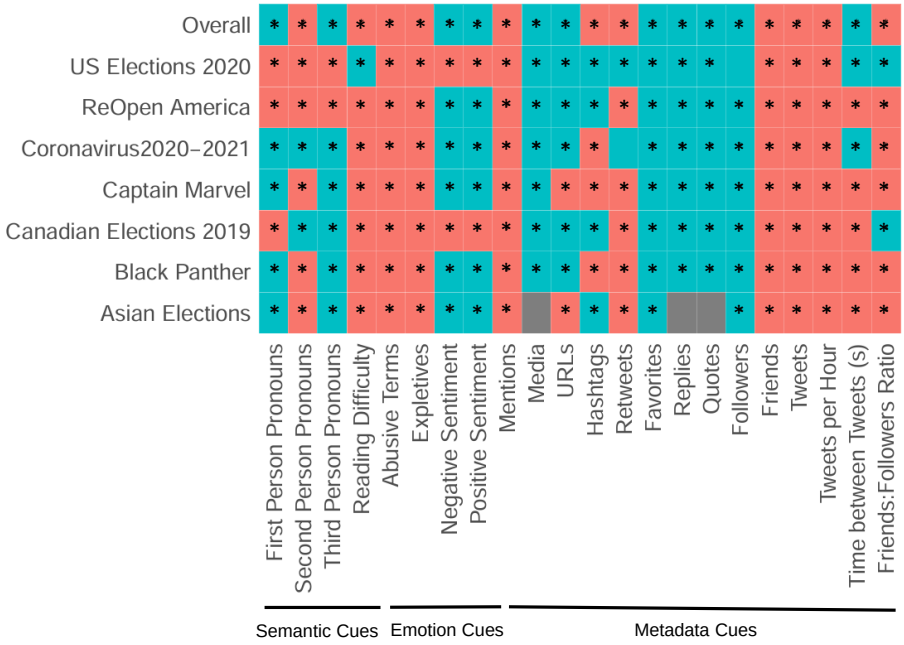


Figure 4.13: Differences in psycholinguistic cues between CSAs and humans. Red cells show that CSAs use a large number of the cue. Green cells show that humans use a larger number of the cue. * within the cell indicates that there is significant difference between the usage of the cue between CSA and human at the $p < 0.05$ level. (Published in [212])

Figure 4.14 compares the linguistic cues among three types of users in the self-collected Telegram data: the Disinformation Dozen, the bots, and the humans. We find that the Flesch-Kincaid reading difficulty of the messages by all three groups are very low, for short online messages are not generally very complex. The largest use of linguistic cues is the positive and negative terms, which could be used in discouraging vaccination and encouraging natural health cures. Another commonly used linguistic feature is the 3rd person pronoun, e.g., "we", which gives a sense of community and that we are all in it together [204]. The 1st person pronoun, e.g., "I", is also frequently used, which provides a personal touch with personal anecdotes and opinions [142].

The differences in linguistic signatures are useful not only for detecting CSAs and humans, but also for interpreting the strategies of CSAs. For example, that CSAs rely heavily on amplification cues and emotionally restrained content reflect their primary motivation as information broadcasters and disseminators in an influence campaign, whereas the use of more personal lin-

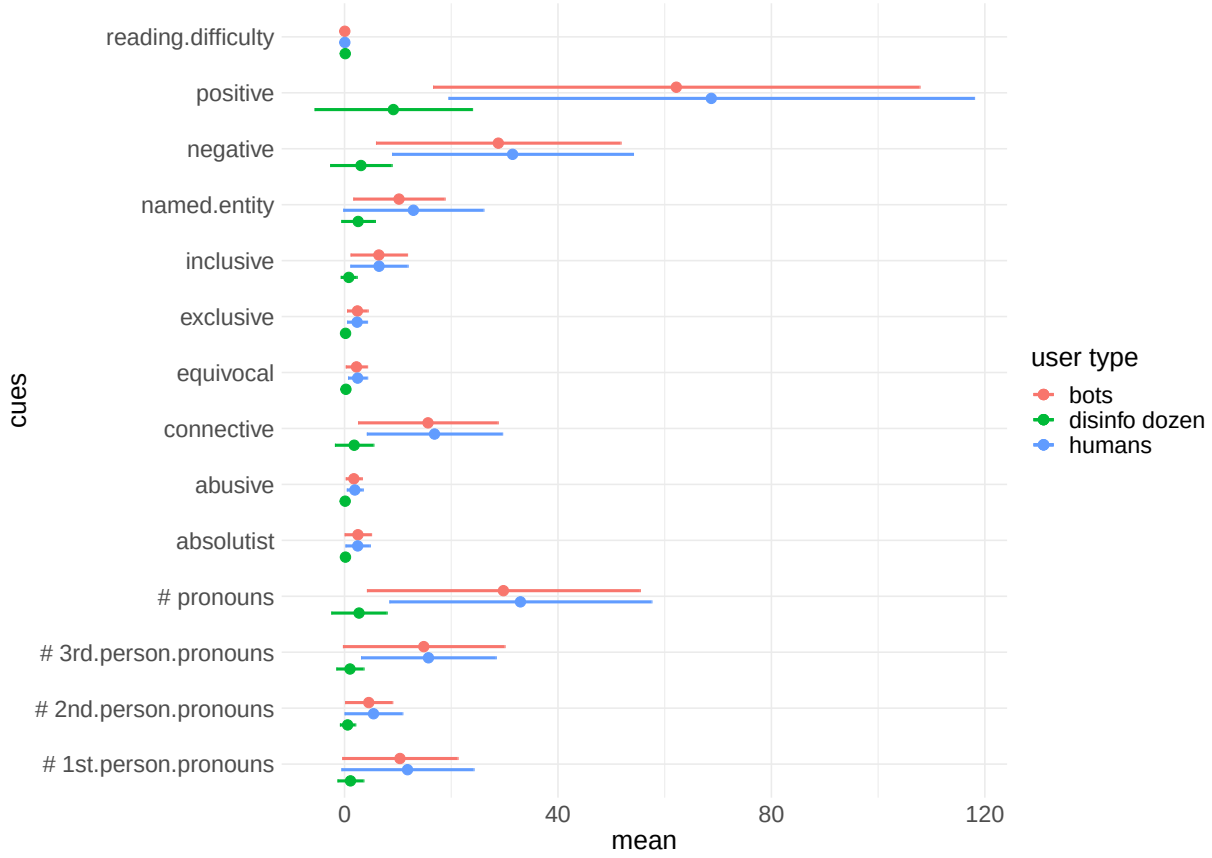


Figure 4.14: Comparison of linguistic cues among groups of users in the Telegram data (published in [221])

guistic of humans reflect more organic social engagement.

4.7 Cognitive Bias Triggers

Social media posts from both Bots and Humans can activate these cognitive shortcuts through the use of bias triggers, such as emotional language or repetitive framing. Prior studies have observed that viral misinformation frequently relies on triggering cognitive biases through affective or credibility cues to increase traction [241, 300].

Developing measures of cognitive bias triggers We use a theory-driven computational framework to detect and quantify eight cognitive biases in the tweets from the 2020-2021 COVID-19 pandemic that are related to information. We first begin by conducting an extensive review of psychological theories and empirical studies of the observation of the biases in the social media context. These empirical studies are presented in Table 4.2, and reveals that the use of biases can positively influence social media algorithm systems to prioritize one's narratives.

Bias	Observations in Literature
Judgment by Representativeness: Evaluate an outcome's likelihood by its similarity to the known prototype, leading to judgments influenced by individual similar events	
Homophily Bias	Retweet posts of similar political leaning [316]
	"Social network homogeneity": users are exposed to like-minded information in online networks [47]
	TikTok's recommendation system that recommends videos based on user's demographics [125]
Authority Bias	Tweets that tag influencers or important politicians on Twitter to lend credibility and enhance belivability [316]
	A small but vocal minority of anti-COVID-19-vaccine medical professionals leveraging their professional titles and medical expertise as evidence to persuade people to believe them [230]
	Authority cues are most effective at inducing credibility bias [160]
	Perceived experts are influential in spreading anti-vaccine misinformation on social media [118]
Judgment by Availability: Evaluate an outcome's likelihood by the ease with which relevant information can be retrieved or accessed	
Affect Bias	Antivaccine content tend to use more emotions rather than narratives and express sentiments through images and videos to increase emotional appeal [303]
	Aesthetically pleasing Instagram posts are used to spread QAnon-related content and conspiracy theories [14]
	Bots invoke emotions for call to action in crises [202]
Negativity Bias	Bots increase exposure to negative/ inflammatory contents; Bots generate specific content with negative connotation that targets most influential individuals [277]
	Anti-vaccination groups use significantly more negative affect terms and references to death on Twitter [262]
	Bots consistently display significantly more negative sentiment and demonstrate consistently negative impact during heated online periods [158]
Illusory Truth Effect	Bots repeat the same messages on democracy ideals multiple times [137]
	Repeating the same message multiple times with misinformation corrections also make people believe the misinformation [155]
	Bots discussing religious issues surrounding Indonesia on Twitter often repeat a message with slight variation while retaining the same message template [70].

Bias	Observations in Literature
	Content moderators of COVID-19 headlines can be susceptible to illusory truth effect from viewing the same false news multiple times and thus believe the false news [159]
Availability Bias	The average number of retweets for non-credible Bots is higher than that of credible bots/users [316]
	Retweet frequency is used as a feature for identifying social media Bots [182, 301]
	Bots display hyper social tendencies by initiating retweets in the coronavirus discourse [326]
Judgment by Anchoring: Evaluate an outcome's likelihood by prior anchors, i.e., personal experiences, exisiting beliefs	
Cognitive Dissonance	Peer pressure by coordinated bots results in users changing their expressed stance towards the vaccine [204]
	Counter-attitudinal exposure causes people to strengthen their opinions [114]
Confirmation Bias	social media platform's news engagement algorithm aggregate like-minded news content to reinforce users' existing beliefs, leading to the formation of echo chambers where minority perspectives with opposing views are marginzalied [114**?**]
	Confirmation Bias on Twitter during the COVID-19 pandemic induced polarization and echo chambers [192]
	"Online selective exposure": users maintain their prior belief in the presence of Google search algorithm results [271]

Table 4.2: Examples of Biases Triggered on Social Media in Literature.

We had two expert annotators to annotate a 1% sample (n=800) of tweets for the presence of cognitive bias triggers in each tweet. These two annotators were trained in the understanding of cognitive biases and had not seen the dataset before. A tweet can be labeled with multiple types of biases. If no bias trigger was found in the tweet, the tweet was unlabeled. A third annotator was recruited to break any disagreements, resulting in the formation of final labels. This human annotation would be the gold label. The two annotators reached agreement 71.89% of the time.

From these expert annotations, we distilled computational heuristics based on the identification heuristics the annotators used. These heuristics are summarized in Table 4.3, with illustrations of observations in our dataset.

Cognitive Biases	Observations in our dataset	Computational Detection Heuristics for Bias Triggers	Scope
Judgment by Representativeness			

Cognitive Biases	Observations in our dataset	Computational Detection Heuristics for Bias Triggers	Scope
Homophily Bias	Author @EasyWorldNews writes: *"@globalfirstnews Readers' poll: if you are offered a Covid vaccination, will you accept? "*	Author shares tweets from people with same affiliation as oneself.	Tweet
Authority Bias	Author retweets: *"RT @DrEricDing: Dangerous anti-vaccine & far-right groups shut down Dodger Stadium's mass #COVID19 vaccination site [...]"*	Author tags authority sources, or Author use keywords about authority sources, or Author is an authority source.	Tweet
Judgment by Availability			
Availability Bias	Author retweets three times: *"RT @User1: India fastest country to cross 1 million Covid-19 vaccinations, 25 lakh doses administered so far: Government"*	Author retweets/quotes the same tweet at least 3 times.	User
Illusory Truth Effect	Author shares the tweet *"The fastest way to end the #COVID19 pandemic is to make safe and effective #vaccines available to everyone on the planet!"* three times, each tagging different users.	Author posts at least 3 tweets that are geq80% similar to each other.	User
Affect Bias	*"Apartheid Israel is withholding the coronavirus vaccine from Palestinians, whilst simultaneously bombing their hospital"*	Tweet contains at least 3 emotion words or at least 1 media content.	Tweet
Negativity Bias	*"This is a disaster, and it's getting worse!!: Inside Pfizer's feverish rush to bring a Covid-19 vaccine to market in record time"*	Tweet contains at least 2 negative words.	Tweet
Judgment by Anchoring			

Cognitive Biases	**Observations in our dataset**	**Computational Detection Heuristics for Bias Triggers**	**Scope**
Cognitive Dissonance	First tweet shows anti-vaccine sentiments: *"[...]I will never have the vaccine ever, [...]#NoVaccine" to "March 2020 wwas the hardest. March 2021 slightly better due to #vaccines"*; second tweet shows pro-vaccine sentiment: *"Protect your community. #getthevaccine"*	Messages in two time frames have different stances towards the COVID-19 vaccine topic. The first messages does not match the majority stance of the user's interaction network, but the second message does	User, Temporal
Confirmation Bias	Three tweets from the same user shows anti-vaccine sentiment: *"Immediately add immunity-building/antiinflammatory/antiviral garlic/Vitamin D3 to the treatment mix!*; *Garlic cuts colds by 50% (COVID-19 is a form of a cold)"*; *"per the Israelis, 2000-5000 IUs of daily D3 cuts COVID+ cases 50% also."*	Author posts at least 3 consecutive tweets with the same stance towards the COVID-19 vaccine; or the tweet contains at least 2 sentence of the same stance towards the vaccine	User, Temporal

Table 4.3: Computational Detecting Triggers of Human Biases.

On average, the computational method achieved an accuracy of 43.54% compared to the gold standard of manual labeling. This accuracy was reasonable, because it surpassed the accuracy expected from the random assignment of bias labels (14.28%). We note that for authority bias, we identified two heuristics, one that uses explicit authorities (i.e., government officials, public figures, authoritative occupations like teacher, police), and another that uses implicit authorities (i.e., social media influencers, context-specific authorities). Combining implicit and explicit authorities increases heuristic accuracy, but requires a list of manually compiled set of social media influencers and context-specific authorities. Measured separately, the heuristics scored an accuracy of $\sim 62\%$. We elected to use the explicit authority version in order to make our methodology more widely generalizable. Table 4.4 reflects the agreement percentage between the first two annotators and the computational accuracy of the developed heuristics.

Bias	Inter-Annotator Agreement	Accuracy (%)
Homophily Bias	98.88	50.00
Authority Bias (Implicit Authority)	60.02	62.00
Authority Bias (Explicit Authority)	60.02	62.10
Affect/ Negativity Bias	69.00	54.07
Illusory Truth Effect	97.48	66.67
Availability Bias	67.46	82.93
Confirmation Bias	32.54	61.00

Table 4.4: **Annotation Statistics.** Accuracy (%) denotes the proportion of annotated tweets (n=800) that our computational algorithm matched the human annotation. Inter-annotator agreement represents the proportion of tweets the first two annotators agreed upon.

Patterns of Cognitive Bias Triggers usage Figure 4.15 shows that, on average, CSAs applied bias triggers 4.71 ± 4.42 times more than humans do. The majority of human tweets (54.16%) did not employ any bias trigger, whereas the majority of CSA tweets (80.19%) included at least one bias trigger. A proportion z-test comparing the presence of bias triggers across both CSA and human tweets confirmed this difference.

Figure 4.16 is a heatmap that illustrates the co-occurrence of bias triggers within individual tweets. Between Humans and CSAs, their tweets exhibit distinct co-occurrence patterns of bias triggers. Bot tweets display more frequent and diverse pairings, supporting prior findings that CSAs tend to embed more bias triggers overall. Notably, the most common Bot pairing, (Cognitive Dissonance, Availability Bias), was rarely observed in Human tweets. This echoes past work that Bots are more likely to shift their stance to align with dominant network opinions, then amplify those views through repeated quoting or retweeting, a behavior that contrasts with typical human tendencies [204]. The most frequent Human pairing was (Confirmation Bias, Cognitive Dissonance), which suggests that when Humans adjust their views to match their social environment, they are more likely to reinforce their new stance with original posts rather than sharing existing content. Both strategies seek to reinforce the revised position and construct a coherent narrative trajectory – Cyber Social Agents via algorithmic amplification, Humans through organic expression.

Another frequent bias pairing among CSAs, (Affect/Negativity Bias, Availability Bias), highlights their tendency to amplify emotional content, in line with rumor theory that emotionally charged and familiar narratives are more readily accepted as truth [147]. However, Humans exhibit only light co-occurrence between Affect/Negativity Bias and other biases, suggesting that emotional appeal in human writing tends to emerge organically rather than through strategic bias combinations [240]. In addition, some pairings like (Authority Bias, Homophily Bias) appeared sporadically in Human tweets but were virtually absent in CSA tweets. This pattern aligns with social signaling theory: like-minded Humans occasionally invoke shared authorities to reinforce group identity, whereas CSAs tend to avoid in-group based appeals. Instead, they borrow the persuasive power from authoritative figures while favoring broad and generalizable messaging strategies to engage with diverse audiences [181].

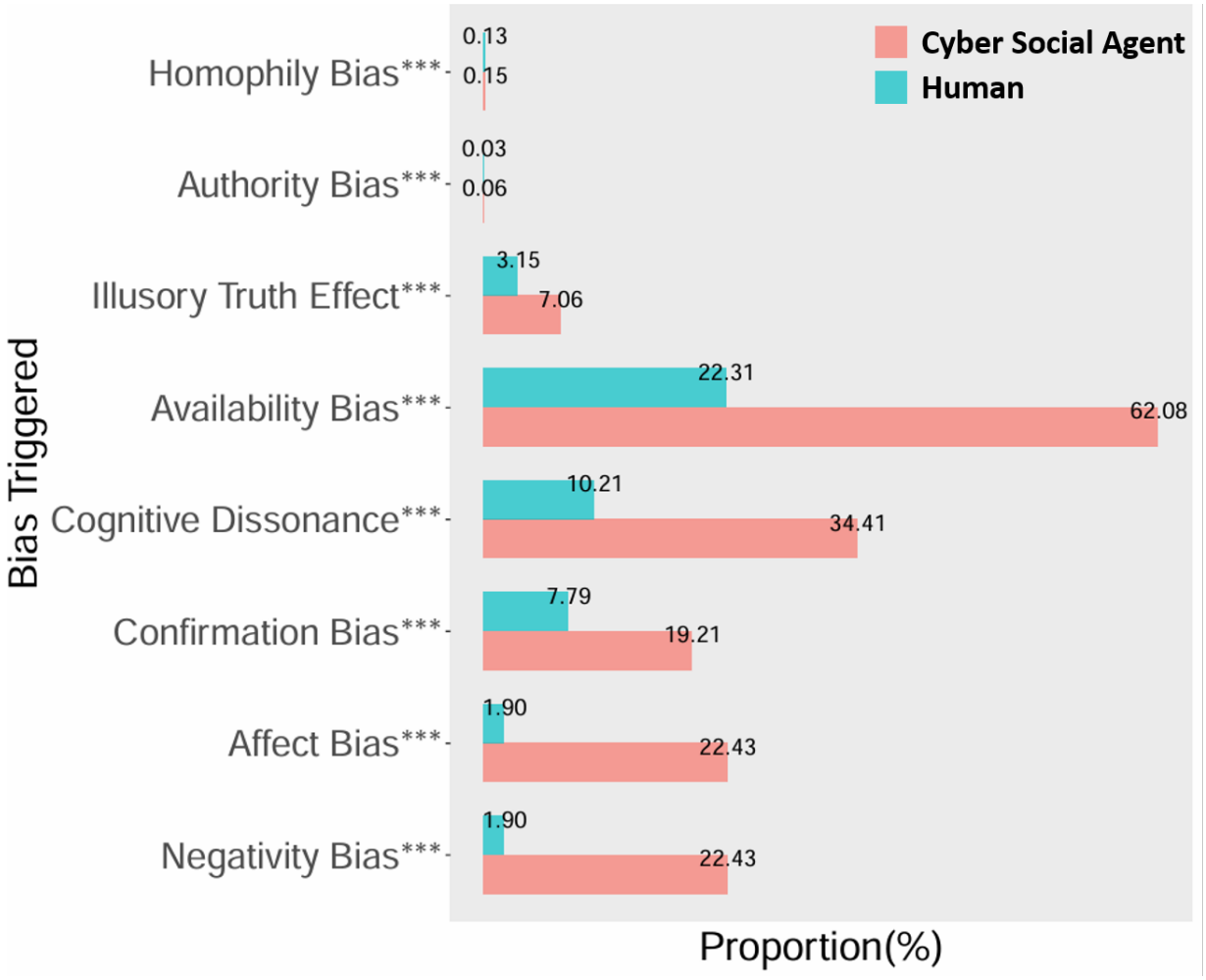


Figure 4.15: **Distribution of the Bias Triggers.** This illustrates the percentage of tweets that attempted to trigger cognitive biases by two different user types.

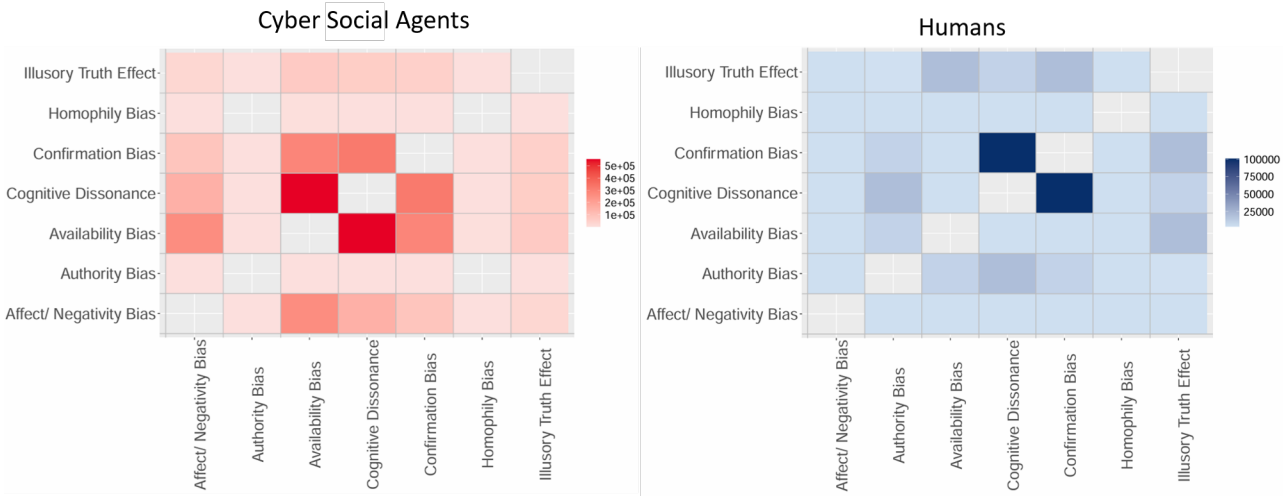


Figure 4.16: **Co-Occurrence of Bias Triggers.** The heatmap is color-coded according to the prevalence of co-occurring triggers for two biases within individual tweets.

Effectiveness of Cognitive Bias Triggers We measure the effectiveness of the use of bias triggers through engagement metrics. Social media engagement are behavioral manifestations between users and their audience [325]. These behavioral manifestations include: retweets, likes, replies, link clicks, media clicks, hashtag clicks and so forth [325].

We further analyzed the relationship between cognitive bias triggers and engagement through Ordinary Least Squares (OLS) regression models by user type (Humans vs. CSAs), illustrated by Figure 4.17. The regression statistics show certain key findings.

First, CSAs consistently exhibit statistically significant associations between most biases and engagement metrics, except for Homophily Bias. These associations are particularly pronounced for shallow engagement, the retweets and favorites actions, which are low-effort, single-click interactions. Bias triggers show little influence on replies and quote tweets, which typically require more cognitive effort and deliberation.

In contrast, Human tweets demonstrated negligible or economically insignificant effects from the same bias triggers. This divergence suggests that shallow engagement (e.g., likes and retweets) is more susceptible to heuristic-driven responses, which bots are designed to exploit. CSAs are strategically constructed to activate cognitive shortcuts and elicit such rapid, low-effort reactions from users [247, 277]. Empirical studies have shown that CSAs amplify the reach of low-credibility content by targeting these low-cost engagement mechanisms [263, 277]. By comparison, Human tweets are often embedded in richer interpersonal or contextual meaning, which requires more cognitive alignment from readers. As a result, bias triggers alone are insufficient to predict engagement with human content. [181].

Three biases, Affect/Negativity Bias, Cognitive Dissonance and Confirmation Bias, when triggered by CSA tweets, were consistently associated with increased user engagement across all four metrics. Affect/Negativity Bias aligns naturally with the social media environment, as emotionally charged content rapidly captures attention [249] and increases tweet engagement up to 4%. Although users naturally seek to avoid Cognitive Dissonance, in which one's expressed

opinions do not line up with the majority of one's immediate social environment, this bias acts as a motivational trigger. When CSAs resolve this tension by changing the expressed stance of subsequent tweets, these tweets show an increased engagement of up to 9%. Confirmation Bias, which affirms pre-existing beliefs, can increase the perceived validity of tweet and strengthen in-group alignment, thereby encouraging audience endorsement (increased interaction up to 3%). From a dual-process of processing perspective, these bias triggers operate primarily through the System 1 cognition processing, prompting fast, low-effort and emotion-driven interactions [85]. In the high-speed context of social media, what matters is whether a bias affirms or challenges beliefs, but whether it captures and activate emotional or identity-relevant responses that elicit engagement [54].

In contrast, another three biases, Homophily Bias, Availability Bias, and Authority Bias, demonstrated a disengaging effect. Availability Bias emerged as the most potent in reducing engagement: its activation in recipients lead to a decrease in favorite counts by up to 28% and retweet counts by nearly one third. The repeated retweeting activities from Bots could be seen as spam that thus reduced credibility of the retweeted narratives [295]. Both Homophily Bias and Authority Bias are linked to the decision-making process governed by the representativeness heuristic. Among the two, Authority Bias exhibited a greater negative impact. Bot tweets containing Authority Bias were found to reduce favorite counts by 22% and retweet counts by 25%. This echoes prior work that there is a cultural paradox in the influence of authority: recommendations from authority figures tend to have diminished persuasive power, particularly in the United States than in Europe [190]. Distrust towards explicit authority figures was rampant in the COVID-19 climate, with only 43.8% of people surveyed worldwide trusting their government [287].

The same bias can produce opposite effects depending on the source. Authority Bias increases retweet engagement in Human tweets but retweet decreases engagement in Bot tweets. This reflects the idea of human reception by source credibility from communication accommodation theory, where readers respond more favorably to bias triggers when they know or trust the author's identity [99]. Humans likely understood their audience and referenced authorities who would likely lessen the distrust of the general public towards the authority figures and government during the pandemic [276]. Meanwhile, CSAs likely apply a general strategy of implanting Authority Bias triggers following a list of generic authorities whom the receivers distrust more [111]. Triggers of Availability Bias increased the retweet count of Human tweets by 5% and decreased CSA tweets by 25%.

4.8 Conclusion

This chapter examines the nature of Cyber Social Agents. First, it presents that CSAs are distributed throughout all regions of the world, which allows for constructions of Social Cyber Geographical maps that present their distribution across countries, time and economic indicators. Next, we presented the key elements of how these agents differentiate from humans. Our findings emphasize the distinct roles CSAs and Humans play in the online ecosystem. Bots are heuristic-based agents, systematically engineering content for influence at scale, while Human communication remains socially embedded and contextually nuanced. These observations are

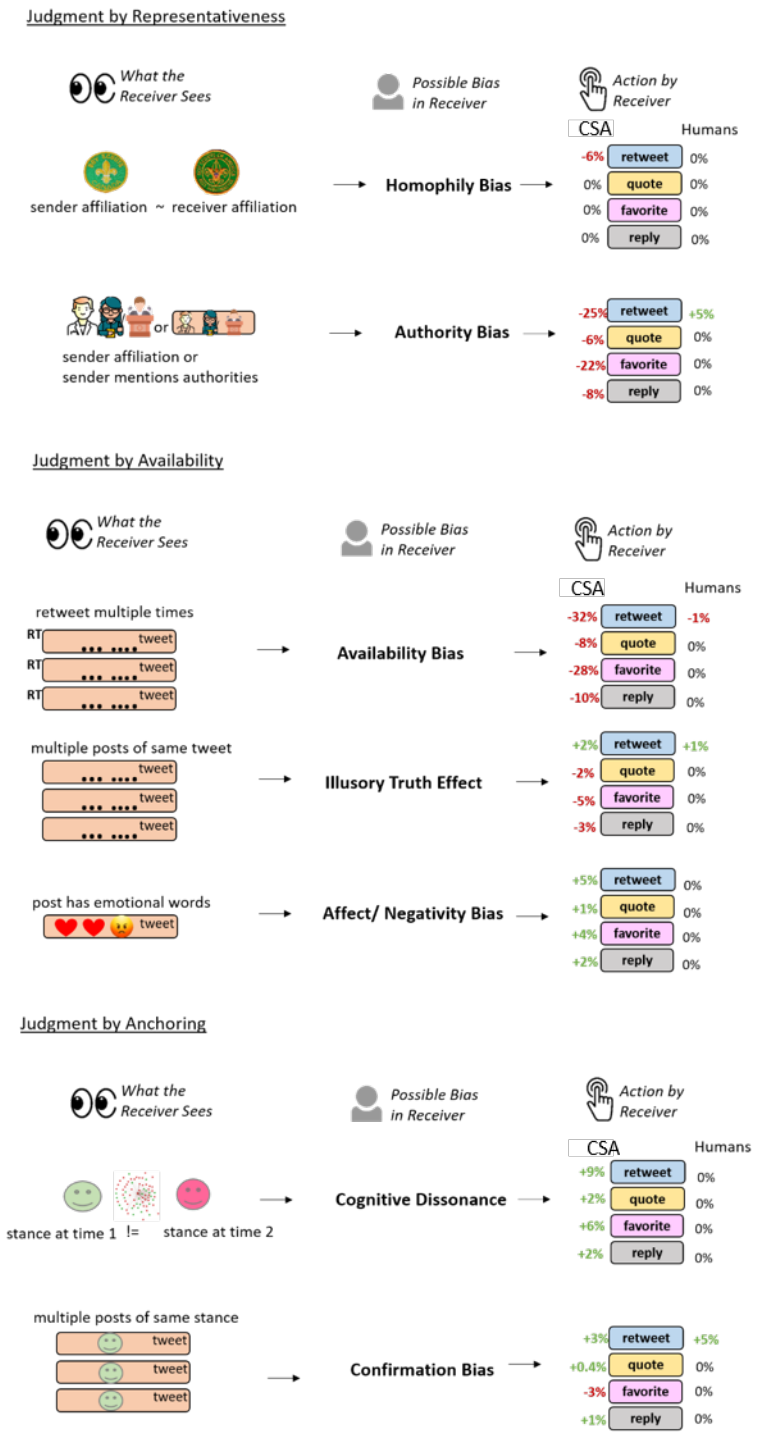


Figure 4.17: Illustrated Summary of the association between Bias Triggers and Tweet Engagement. All estimated percentages are significant to at least $p < 0.001$ level, except for the 0% in CSA's quote/favorite/reply for Homophily Bias.

presented in our work through how the linguistic cue usage of CSAs prioritizes the easier to automate cues such as the use of hashtags, whereas human cues prioritize the use of replies. In terms of motivations and agencies, CSAs exhibit more goal-directed and repeatable strategies and amplified agency through automation, while humans tend to exhibit fluid, context-dependent motivations and limited individual agency. For the use of cognitive bias triggers, CSAs use techniques like repetition, stance changes and emotional words, which have lower cost implementation.

Limitations of our study of the nature of such agents include:

1. For social political representation, we identified the agent's country of affiliation through the comparison of a gazetter with the agent's declared location. The gazetter needs to be constantly updated with the latest social media location lingo.
2. For linguistic cues, we studied the differences in the use of psycholinguistic cues across isolated events. These differences could also be attributed to the nature of the event.
3. For motivations and agencies, we limited our studies to English-language tweets. While this limitation provided us insight towards the motivations of agents that target the English-speaking audience, it leaves out the broader picture of the motivations towards other language groups.
4. For cognitive bias, while our classification method achieved high accuracy ($\sim 62.7\%$) compared to human expert annotations, the interpretation of whether a tweet was intended to contain a specific bias trigger ultimately rests on the author's true intent, something that is not directly observable.

Future Work includes: for social political representation, the inference of an agent's location need not only rely on the description the agent put, but rather should also infer that from the textual cues of the post. For linguistic cues, conditioning the difference between psycholinguistic cues based on events. For motivations and agencies, future work involves analyzing these motivations with diverse language groups. For cognitive bias triggers work, another future work involves integrating social network features on whether the network positions of the CSA/human will amplify or diminish the bias.

Chapter 5

Network Interactions and Coordination Profiles of Cyber Social Agents

5.1 Introduction

Agents do not live in isolation on social media. They interact with other agents too. The range of interactions that agents can perform depends on the affordances of the platforms. Each platform has its own nomenclature for similar mechanics. For example, the function that allows for sharing a post is named as "retweet" on X and "share" on Facebook. This naming differences reflect how the algorithms are designed differently to prioritize and differentiate social media mechanics, and shape the interpretation and social meaning of the mechanic. For example, "retweet" on X indicates amplification, whereas "share" on Instagram indicates casual sharing. Such branding reflects how users engage with the mechanic and thus drives patterns of agent-content and agent-agent engagement [72]. Excessive interactions between two agents within a short time window are termed "coordination", because it seems that the two agents are in cahoots with each other to deliver a message. This chapter explores the agent interaction profiles and agent coordination profiles as observed and measured in social media platforms.

This chapter investigates the following guiding **research questions**:

1. What are the unique characteristics of the CSA interaction profiles?
2. How do CSAs coordinate with each other to disseminate information and increase influence?
3. What are the impacts of CSA interactions on the social network?

5.2 Related Work

5.2.1 Network Topologies

The communication network between agents on social media can be represented as a graph $G = (V, E)$, where the vertice set V are the agents, and the edge set E are communication links. Social media communication networks exhibit a range of canonical topologies such as scale-free,

small-world or random networks[27**?**]. These patterns shape how connectivity shape

Social media graphs exhibit a range of canonical topologies—including scale-free, small-world, and assortative or disassortative mixing patterns—that shape how information, influence, and connectivity emerge across the network [27**?**]. These topological properties differ between human-driven and bot-driven communication structures. Studies have shown that automated accounts tend to form highly centralized or hub-dominated clusters, and often exhibiting unnaturally high activity and degree patterns[35]; whereas human networks more commonly reflect organic community structure with clustered, reciprocated, and heterogeneous ties [295]. Both types of users imprint different signatures on the network's macro structure, resulting in structural differences that shape how information spreads and communities form on the platform.

Within the social network graph consists of motifs, which are recurring patterns of small subgraphs that represent the building blocks of complex networks [189]. Network motifs are a powerful analytical tool for understanding complex social phenomena. The triad motif is a three-node pattern used to analyze relationships reciprocity, transitivity, and balance [302]. Human interactions on social media tend to have balanced and reciprocation, with friendships and edges forming through triadic closure [156]. These online human relationships formed from communication and friendship ties have predictable structural triadic balance and reciprocity[281]. In contrast, CSAs often exhibit asymmetric patterns such as repeated one-directional interactions (i.e., retweeting action of amplifier agents) or inflated out-degree to many accounts (i.e., multiple engagement of synchronized accounts).

Star network motifs describe structures in which a single central ego node connects to many peripheral nodes. These motifs have long been recognized as fundamental building blocks of networks, offering insights into how social capital is unevenly distributed and how information and influence flow through a system. Foundational work in sociometry by Moreno [194] first documented star-like patterns in group interactions, showing how central actors shape relational dynamics. Bavela's classic research on communication structures demonstrated that centralized star motifs can be especially efficient for coordinating problem-solving tasks [29]. Freeman later formalized the mathematical basis for detecting central actors—introducing measures such as degree centrality—which remain essential for identifying and quantifying star structures [94].

More recent studies extended these ideas to digital environments. Research on egocentric networks reveals that star-like formations are common in social media platforms, where individuals maintain a small inner circle of strong ties surrounded by a much larger set of weaker, higher-degree connections [136]. Work on network individualism further highlights how people now curate personal networks that span online and offline spaces, underscoring a shift from group-based affiliation to person-centered connectivity [307].

5.2.2 Coordination

The study of coordinated manipulation of conversations on social media has become more prevalent as social media's role in amplifying information that results in hate clusters and polarization come under scrutiny. For example, studies have been performed on how coordinated groups artificially manipulate online information on elections. Analysis of Facebook co-shares of political news stories during the 2018-2019 Italian elections identified hundreds of groups that coordinated to boost political and non-political identities[103]. Similar coordinated influence efforts

have been documented in contexts of US and Brazil political scenes[145, 298].

Coordinating groups on social media are of concern to researchers and policymakers because their activities are not necessarily confined to the online space. Their coordinated campaigns can pose a threat to the social fabric, especially when their campaigns spill over into the offline medium and result in protests and riots. For example, the use of hashtag coordination can capture the noticeable change between online coordination and offline protests in 16 countries affected by the 2011 Arab Spring protests[275].

Online coordination is a multi-dimensional problem. Current coordinated activity detection techniques typically uncover anomalously high level of synchronized action within a time window. An action is a behavior a user can take on a social media platform such as retweets [305], @-mentions, [174], using similar texts [219], posting common URLs [26], or other behavioral-traces that links two users [237]. These approaches define coordination by the repeated pattern of co-occurrence of multiple actions as signals that when combined, can influence information systems. More advanced approach combine these signals to detect higher-order coordination patterns. For example, network-based methods identify clusters of accounts that repeatedly co-engage with similar content through retweets and sharing [**?**], temporal graph mining uncovers behaviors that deviate from the expected human rhythm of online interaction[109], and semantic coordination detect groups that converge on shared messaging frames even when their specific words differ [304].

5.3 Network Interaction Profiles

Network interaction profiles are the characteristic shape of an agent's connections on social media. The network topology describes with whom the actors interact with, how often the interactions are, and in which direction. The structure of the network of an agent governs its influence mechanics and potential, and the connections the agent has are the signatures through which the agent can possibly broadcast, brokerage or converse with other actors, which would shape the online discourse.

5.3.1 CSA vs Human

CSAs and humans have visually different social interaction patterns, which can be represented by their network topology profile. We investigated the all-communication ego network of the top 20 most frequent communicators in the Asian Elections dataset. An all-communication network is a network where nodes represent users and the links between users represent a communication interaction: retweeting, sharing, tagging a user. The thicker the link, the more communication interactions between the two users.

Figure 5.1 shows the typical network interaction profiles of CSAs and humans through a two-degree all-communication ego network. That is the communications of the ego's set of alters and the connections among them, and the alters' set of alters and the connections among the alter's alters.

This difference in interactions profiles between CSAs and human users reveals the communication patterns of both user classes. CSAs typically have an extended star-shaped ego network,

suggesting a hierarchy of interconnected agents in an operation network to disseminate information. This ego agent connects to many alters that rarely interact with each other. The extended star network structure is a result of the CSAs broadcasting their messages widely and simultaneously, and therefore is well-suited and commonly used for amplification campaigns like repetitive content sharing to maximize reach [226].

The extended star structures are effective in providing scale and speed, but the lack of structural cohesion means the structure struggles to sustain narratives once the initial visibility fades. The peripheral agents do not really interact with each other. The narrative propagates from the ego agent to the ego's alters to the alters' alters. Without the ego, there would be no information that was shared across those alters. Therefore, star networks are highly fragile. The sustenance of the narratives depends on the central node. If the central node were to be removed, whether through inactivity or platform suspension, the network often collapses entirely. This limits the durability of such a strategy of influence spread. This suggests that people might set up such automated agents to broadcast messages, but shut them off once the message is announced. Such CSAs strategies make it hard for analysts to track the ephemeral agents. For example, on GitHub, an online code-sharing platform, 11.3% of the projects enable then later deprecate the DependaBot, a Bot set up to blast announcements about the software project [124].

On the other hand, humans have a hierarchal network topology, where communications are organized in nested tiers. Humans communicate predominantly within their immediate network before extending their communication outwards. This hierarchal form is a digital reflection of how humans embed themselves in community in real-life: the ego connects with his friends, who are then connected with his friends of friends [13].

The multi-layer network of a human user results in slower information diffusion but supports credibility and persistence of messages through the repeated exposure as the message propagates through the layers. Over time, this redundancy of multiple message propagation routes makes human-driven narratives more resilient to disruption and more likely to endure beyond short-lived bursts. In fact, our study of coronavirus disinformation on Telegram observed that it was humans rather than automated agents that sustained the disinformation spread through their sharing of messages across channels [221]. The human hierarchal sharing structure created a snowball effect that continuously kept the narrative active and prolonged its visibility within and across channels.

	CSA	**Humans**
In-degree	$\mathbf{0.05 \pm 0.08}$	0.02 ± 0.02
Out-degree	**8E-4 ± 1.4E-3**	1.6E-3 ± 3.3E-3
Total degree	$\mathbf{0.15 \pm 0.09}$	0.16 ± 0.11
Density	$\mathbf{0.35 \pm 0.06}$	0.034 ± 0.06
% CSA alters	$\mathbf{9.66 \pm 2.98}$	7.31 ± 3.10

Table 5.1: Comparison of network metrics. For the in-degree, out-degree, total degree and density, we present the ratio of mean(metric) for agent type : max(metric) across all agents in the event. (published in [212])

We also compared network metrics of the all-communication graphs, which is presented

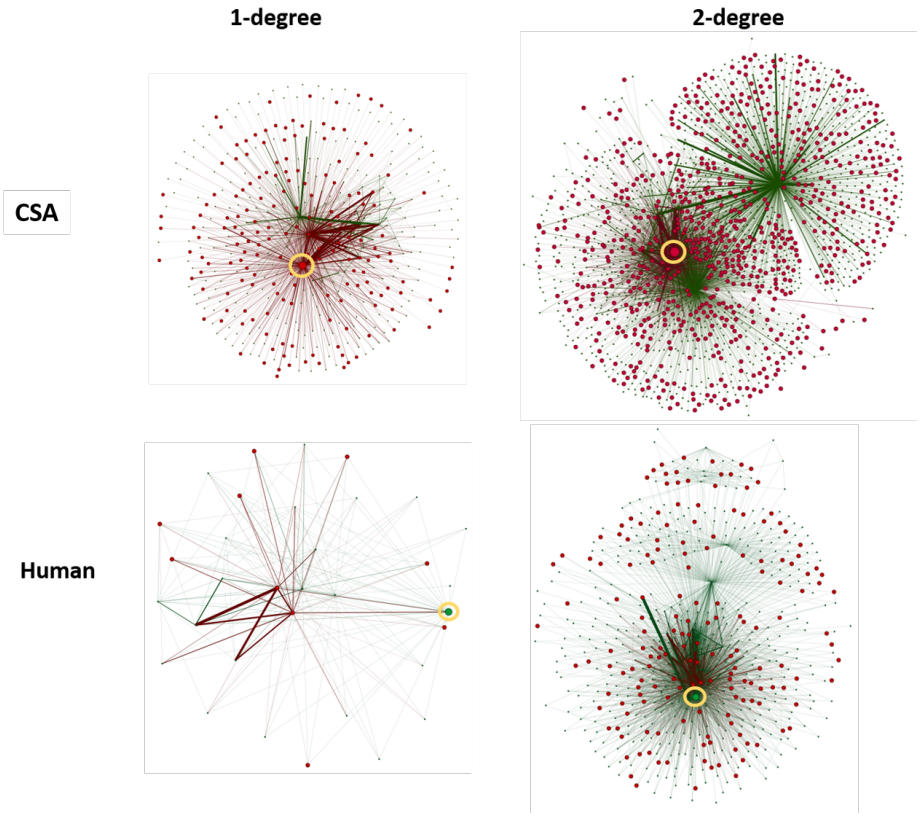


Figure 5.1: Two-hop Ego network structures of CSAs and Humans who are the most frequent communicators in the Asian Elections dataset. Nodes represent social media users. Links between users represent a communication relationship between the two users (i.e., retweet, mention). Bot users are colored in red, human users in grey. The width of the links represent the extent of interactions between the two users. In these most frequent communicators, CSAs have a star network structure, and humans a hierarchal structure. (published in [212])

in Table 5.1. CSA ego networks have higher density than human ego networks (8.33% more dense) which reflected how the CSA star communication structure is tighter and form more direct interactions than humans' ego networks. In terms of the composition of alters, on average, a CSA has 9.66% CSA alters and 90.34% human alters; whereas on average, a human has 7.31% CSA alters and 92.69% human alters. This shows that both CSAs and humans interact more with humans rather than other agents in their ego network. By the principle of homophily, it is natural for humans to interact with other humans [37]. However, CSAs violate the principle of homophily, and are actively forming communication interactions with humans rather than other agents.

5.3.2 Star Motifs

Star network motifs are motifs where a central ego node is connected to multiple peripheral nodes. These star motif structures have long known to be key elements in networks, and are

important for studies of social capital distribution and pathways of information diffusion and influence. Prior research typically studied networks where all nodes are humans [307], and in such structures, all had the same affordances in terms of ability to communicate. Extending on these prior research, we study star motifs where there is a mixture of automated CSAs and humans.

We examined the presence of star motifs in a subset of the 2020-2021 coronavirus data. We filtered the data for tweets that were written one week before the Pfzier coronavirus vaccine was officially launched and were related to the vaccine in that they used the hashtag #covidvaccine. Then, we ran the BotHunter algorithm [32] with a threshold of 0.7 to identify automated agents. We thus collected data of 580,135 unique X users, of which 26.43% of them are bot users.

Then we constructed retweet interaction networks. We chose to construct network graphs based on retweet interactions to capture information sharing and amplification. The act of retweeting shares the tweet to the user's network, representing both implicit endorsement and information amplification [39]. A retweet network graph is thus defined as G, in which $G = (V, E)$. In a retweet network graph, the edge $e_{i,j}$ between v_i and v_j meant that v_j retweeted v_i. Each edge has a weight $w_{e_{i,j}}$ that indicates the number of retweets between the two agents. A thicker edge indicates sustained retweeting actions by v_j of v_i.

From the retweet networks, we extracted the ego networks that have the star motifs. Ego networks describe the connections of an agent (the ego n_0) with its social peers (the alters $n \in N,\ n \neq n_0$) [15]. Finally, we define a star network motif as a motif where one ego node is connected to multiple peripheral alters. A traditional definition states that the alters should not have connections with each other, but since we were dealing with social media data, we relaxed this constraint and allowed for the alters to have minimal links with each other. Formally, we define a star motif S_k as a connected subgraph with $k + 1$ nodes consisting of one ego node v_c and k peripheral alters $\{v_1, v_2, \ldots, v_k\}$, where:

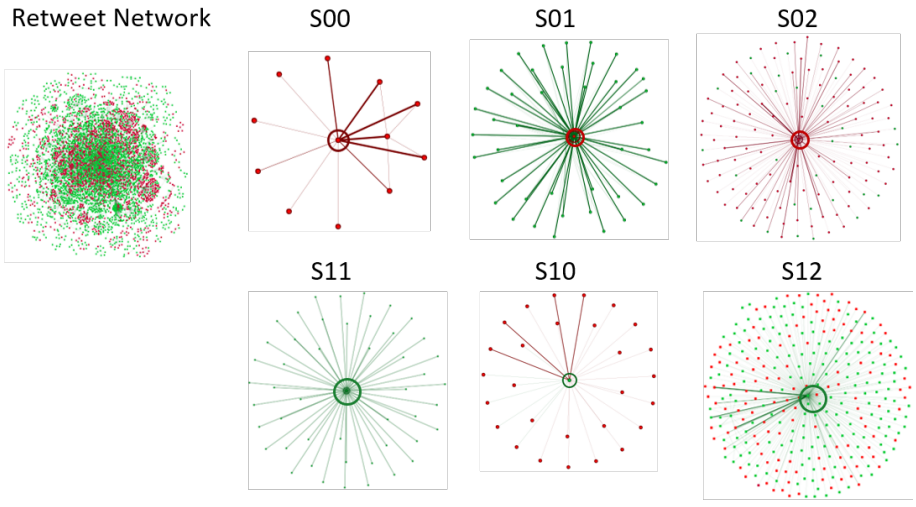
$$S_k = (V, E) \text{ where } V = \{v_c, v_1, v_2, \ldots, v_k\}$$
$$E = \{(v_c, v_i) : i \in \{1, 2, \ldots, k\}\}$$

Constraints:

- $\deg(v_c) = k$ (ego node has degree k)
- $\deg_A(v_i) \leq 2, \forall i \in \{1, 2, \ldots, k\}$ (each alter can have at most 2 edges with other alters)
- $\deg(v_i) \leq 3, \forall i \in \{1, 2, \ldots, k\}$ (each alter has degree at most 3)

For simplicity, we studied the undirected star motifs. That is, the edges in E do not have a defined direction. Figure 5.2 represents a typology of six star motifs that can be formed between CSAs and humans. The singular ego node n_0 is one drawn from the set of {CSA,human}. n_0 has multiple alters $n_a = \{n_{a1}, n_{a2}...\}$, mainly three scenarios of alters: {CSA, human, mixed}. Therefore, there are $2 \times 3 = 6$ primary combinations of ego-alters. The notation of the star motifs in the figure contains three characters Sab. S indicates the star shaped network. $a = 0, 1$ indicates whether the ego n_0 is a human or CSA. $b = 0, 1, 2$ indicates the ordinal formulation of whether the alters n_a are entirely CSAs, entirely humans, or a mixture of CSAs and humans.

From a content analysis of these six star motifs, we observe that the motif patterns demonstrated distinct amplification strategies. CSA-ego motifs (S00, S01, S02) exhibited characteris-



Motif Name	Ego	Alter
S00	CSA	CSA
S01	CSA	Human
S02	CSA	Mixed
S11	Human	Human
S10	Human	CSA
S12	Human	Mixed

Figure 5.2: Variations of undirected star shaped motifs between CSAs and humans (published in [226])

tics of deliberate, and perhaps coordinated, information spreading, where CSAs leverage their central network position to rapidly disseminate content to multiple peripheral nodes, similar to findings by previous research of how CSAs can be information super-spreaders and create efficient information propagation pathways [42]. Human-ego motifs (S10, S11, S12) demonstrated more organic community formation patterns with natural clustering around authoritative sources and shared interests. This observation is consistent with the principle of homophily [186]. The motifs with mixed agent alters (S02, S12) revealed the complex interaction dynamics in a social network, pointing to the importance of accounting for these hybrid communication patterns in influence detection and mitigation strategies.

Star motifs provide a structural framework for understanding how influence operations manifest at the network level. The S00 motif represented a pattern of coordinated inauthentic behavior through a co-retweet network, and could be identified in monitoring systems to trigger alerts when multiple of such patterns emerge around sensitive topics like politics [206]. The S01 motif represented a pattern where automated agents can influence genuine human users, consistent with research on social influence and peer effects of bots [204]. The S10 motif demonstrate how authentic human content creators can be artificially amplified through sets of bots, potentially distorting the perceived popularity of their messages [258].

5.4 Coordination

Coordination refers to the phenomenon where multiple users perform the same action within overlapping time windows. An action is defined as a social media mechanic paired with a content artifact. An example of an action would be "posting a text post with a specific hashtag". Here, the mechanic is publishing a tweet, while the artifact is the hashtag. When many users repeat the same action nearly simultaneously, it creates the appearance of deliberate echoing and suggestion of collective intent. The most basic unit of this phenomenon is when the number of users is two: two users performing the same action is referred to as a co-action [306].

Table 5.2 presents a summary of the coordination types explored in this thesis, and the corresponding examples taken from the published papers from this thesis.

Coordination Type	**Mechanic**	**Artifact**	**Case Study**
Amplification coordination	Sharing or retweet	Same post or account	Round-robin retweeting mechanism of a group of CSAs within a 2021 Taiwan-China discourse [137]
Social coordination	Tagging (or @mention), Reply	User handle	2021 COVID Vaccine release discourse on X revealed socially coordinated groups that revolve around mental health support, financial planning and elder care [203]
Semantic coordination	Use of hashtag strings	Same hashtag	Coordination via hashtags in discourse about the 2020 US elections on X reveal user clusters in support of the Republican and Democrats, because the two factions coordinate via separate sets of hashtags [203]
Referral coordination	Use of URLs	Same URL	Australian news network 7News uses referral coordination to push out links to news articles to region-specific X accounts, ensuring that the important news reaches all the different sets of local audiences [175]

Coordination Type	Mechanic	Artifact	Cases Studied
Textual coordination	Posting of original texts	Duplicate of near duplicate texts	Near duplicate texts that have an at least 80% match of each other found within different social affiliation groups in a 2021 discourse on Parler [219]
Media coordination	Media posting (e.g., images, videos)	Duplicate or near duplicate media or combined media	Images from Russia have a single centralized messaging and are well-coordinated, while images from other countries (i.e., Venezuela, Iran) spout multiple messaging efforts, with isolated sets of image narratives [217]
Cross-platform coordination	Multi-platform rollout	Same message sent across social media sites	Website and YouTube URL matches reveals that while Parler and X users reference different sets of URLs, the content of information that they consume are similar [216]

Table 5.2: Types of Coordination explored in this thesis

Coordinated interactions can be identified through constructing clusters of bipartite graphs. Figure 5.3 illustrates the methodology of identifying coordinated clusters from individual actions. First, begin by observing (A), that multiple users repeat the same actions within short time windows. This time window is analyst-defined, and many analysts use a time window of 5 minutes. Second, aggregate the actions into a bipartite graph that represents User x Artifact (Panel B). This reveals the users that use the same artifact within the same time window. Artifacts are unique, for example, the hashtag "#GetYourNewComputer" and "#GetYourNewComputerNow" are separate artifacts and should be treated differently. Finally, project a User x User network, and use network science clustering techniques to identify densely connected clusters (Panel C). These densely connected clusters are the coordinated groups of users.

This approach that analyzes the presence of co-action between users to reveal the synchronity between behaviors in the network is termed as the Synchronized Action Framework. This framework detects patterns of timing or content usage that exceed what would be expected by normal coordinated activity. Users are termed to be coordinated if they perform co-actions more than two standard deviations more than the mean number of co-actions in the network.

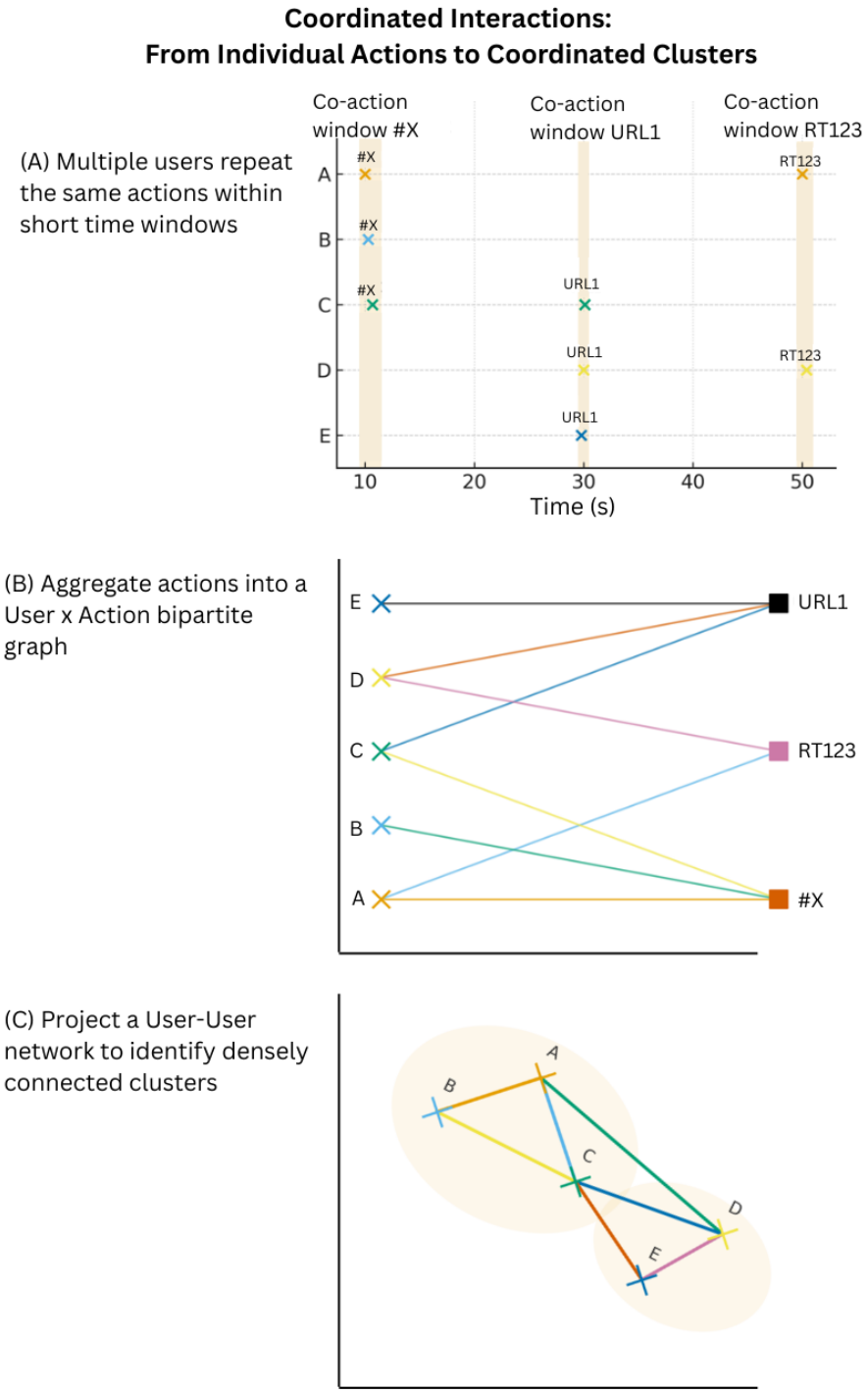


Figure 5.3: Illustration of the methodology of identifying coordinated clusters from individual actions

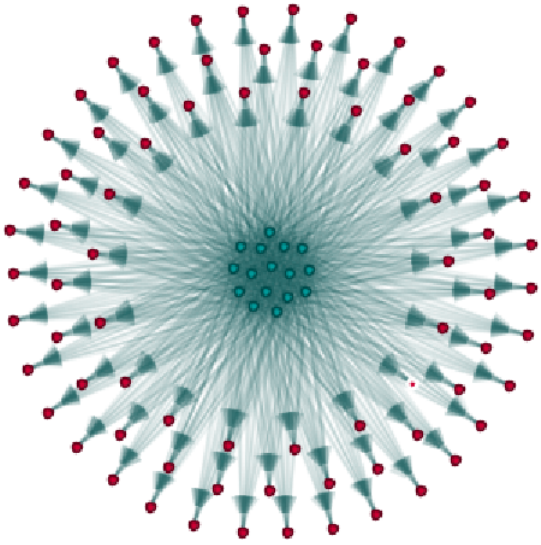


Figure 5.4: Example of amplification coordination through repeated retweets (published in [137])

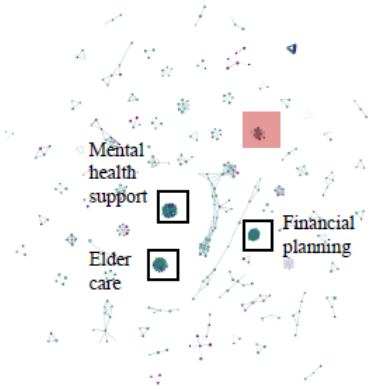


Figure 5.5: Social Coordination within the 2021 coronavirus vaccine release discourse. Nodes are X users and link width represent the strength of coordination between two users. The red box are self-declared CSAs that coordinate socially.

Amplification coordination The co-action used in amplification coordination is the retweet mechanism. An example occurred within the 2021 Taiwan-China discourse[137]. A group of CSAs collectively amplified the posts of a group of core users, enabling the information to spread farther. The coordination network graph is presented in Figure 5.4.

Social coordination The co-action used in social coordination is the tagging, or @mention, mechanism. An example occurred within the 2021 Coronavirus Vaccine release discourse on X[203]. We identify users that socially coordinate by using a 5-minute time window. This reveals socially coordinated groups of users that revolve around mental health support, financial planning and elder care. The coordination graph is presented in Figure 5.5.

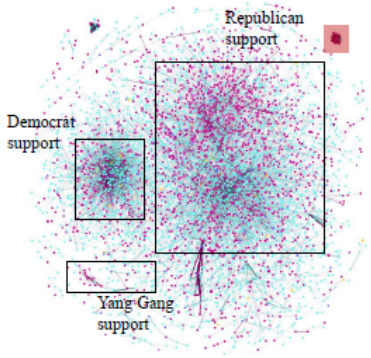


Figure 5.6: Semantic Coordination within the 2020 US Election discourse. Nodes are X users and link width represent the strength of coordination between two users. The red box are self-declared CSAs that coordinate semantically.

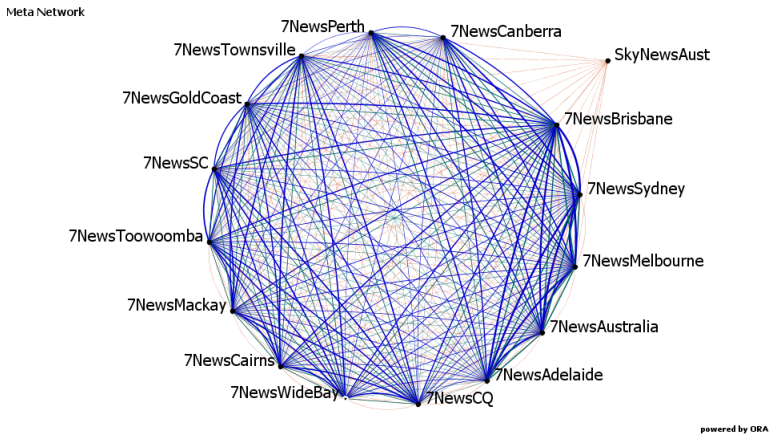


Figure 5.7: Referral Coordination within the ReOpen America discourse. Nodes are X users and link width represent the strength of coordination between two users. The blue, red, and green links correspond to referral, social and semantic coordination, respectively.

Semantic coordination The co-action used in semantic coordination is the use of common hashtags. An example occurred within the 2020 US elections discourse on X[203]. We identify users that semantically coordinate by using a 5-minute time window. This reveals two main groups of users, the users that are in support of the Republican party, and the users that are in support of the Democratic party. Both factions coordinate within the faction via separate sets of hashtags. The coordination graph is presented in Figure 5.6.

Referral coordination The co-action used in referral coordination is the use of common URLs. An example occurred in the ReOpen America discourse[175]. The Australian news network 7News used referral coordination to push out links to news articles to region-specific X accounts (i.e., @7NewsBrisbane, @7NewsQueensland). This mechanism ensures that the important news reaches all the different sets of local audiences who are most likely to be following their local accounts. The coordination graph is presented in Figure 5.7.

Textual coordination The co-action used in textual coordination is the use of common texts. An example occurred in 2021 shared on Parler, where there are groups of parleys coordinating textually [219]. Since textual coordination is not an atomic artifact of a social media post, we had to do some processing to form a User x User coordination graph in order to analyze which users coordinate with each other. To do so, we use a text-to-text graph to induce a user-to-user graph, and set the threshold of which we filter edges through a statistical analysis of all graph edge weights.

We first begin by creating a user-to-text binary graph P, which represents the users that wrote each parley. Next, we create a text-to-text graph A through a k-Nearest Neighbor (kNN) representation of the parley texts, with $k = log_2 N$ where N is the number of parleys [176]. To do so, we perform BERT vectorization on each parley text create contextualized embeddings into a 768-dimensional latent semantic space [73]. We make use of the FAISS library to index the text vectors and perform an all-pairs cosine similarity search to determine the top k closest vectors to each parley vector [140]. We then symmetrize this kNN-graph via $P' = \frac{P+P^T}{2}$ to produce a symmetric k-Nearest Neighbor (kNN) graph of the posts, as this tends to better maintain meso-structures from the data, like clusters, in the graph [45, 177, 255]. The edges of the graph are weighted between $[0, 1]$ by the cosine similarity of the two parleys.

Having found a latent graph of the textual content of the posts we then induce a user-to-user graph. We do this through matrix Cartesian product formula of $U = PA'P^T$.

Where U is the user-to-user graph, A is a user-to-text bipartite graph where an edge indicates that a user posted a given parley, and P' is the text-to-text kNN graph as previously defined. The resulting graph, U has edges that represent the strength of textual similarity between two users, given how close in similarity their posts are in a latent semantic space. U better accounts for not only having multiple, similar posts but better respect the textual data manifold when measuring between two users as compared at only identifying the most similar posts between two users [236].

To sieve out the core structure of the graph, we further prune the graph U based on link weights, forming U', keeping only the links that weigh greater than one standard deviation away from the mean link weight. This leaves behind users that strongly resemble each other in terms of the semantic similarity of their parley texts.

The coordination graph is presented in Figure 5.8, which presents the groups of users that coordinate textually with each other, and the general narratives that they coordinate on.

Combined Synchronization Index In previous parts, we described separate analyses of synchronity via different actions. To characterize the degree of synchronization between the users that posted within an event, we defined the Combined Synchronization Index (CSI). The CSI is a hierarchical which begins with CSI-UserPair, then aggregating the UserPair values to obtain CSI-User for each user, and finally a singular value for CSI-Network.

CSI-UserPair. CSI-UserPair provides an indication of how much two users u and v synchronize with each other, expressing the same semantic or social information in their tweets. We obtain user pairs with a count of the number of times the users synchronize with each other. We normalize this count along the axis of each action, then account for duplicates and scale the user pair value according to the extent of synchronization. This calculation is reflected in

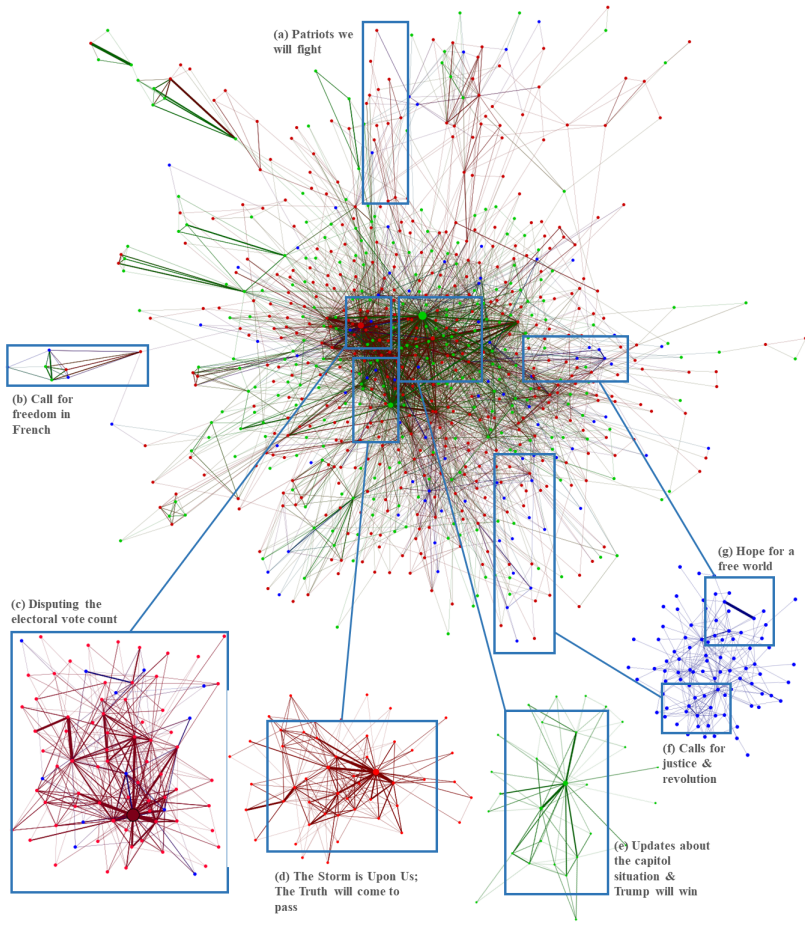


Figure 5.8: Textual Coordination within the 2021 Parler dataset. This is the core structure of user-to-user graph U' representing narrative coordination between groups of users. Military users are colored blue; patriot users red and QAnon users green. The thickness of the links represent the strength of the coordination. Nodes are sized by total degree centrality value.

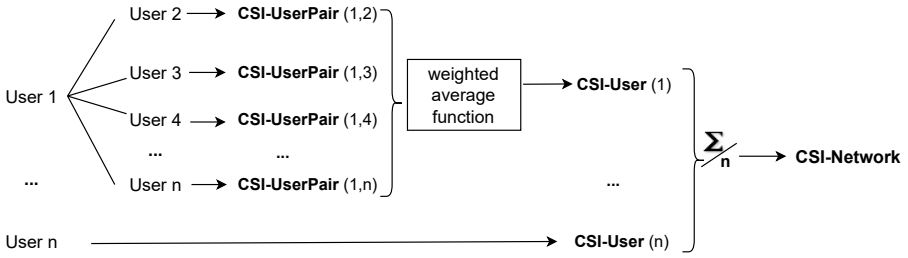


Figure 5.9: Flowchart of the computation and aggregation of the hierarchical Combined Synchronized Index. (Published in [206])

Equation 5.1.

$$\text{CSI-UserPair}(u,v) = [\text{CSI-UserPair}(u,v,a_1) + \text{CSI-UserPair}(u,v,a_2) + \text{CSI-UserPair}(u,v,a_3)] - |a| * |a| \tag{5.1}$$

where a_i is the action types users u, v synchronize in

CSI-User. The CSI-User value is a representation of the degree of synchronization a user partakes in within a network. It is a summation of CSI-UserPair values that a user u is part of, weighted by the frequency of synchronicity of the paired user v. This calculation is reflected in Equation 5.2.

$$\text{CSI-User}(u) = \sum_{i=1}^{n} (S(u,v) * \text{CSI-UserPair(u,v)}) \tag{5.2}$$

for all n pairs that user u is part of

CSI-Network. The CSI-Network score quantifies the average amount of synchronization between users surrounding an event. This calculation is reflected in Equation 5.3.

$$\text{CSI-Network} = \sum_{i=1}^{n} (\text{CSI-User})/n \tag{5.3}$$

where n is the number of synchronizing users in a network

Application of Combined Synchronization Index. We measured user synchronicity across six Twitter datasets. We focused on social activism and political events, selecting for discourse that has a strong stance, in which synchronization may be used to champion the cause. With these parameters, we examine four social activism events: (1) 2018 Black Panther Movie; (2) 2021 French Protests; (3) 2020 ReOpen America; (4) 2020-2021 Coronavirus. We also examine two political events: (1) 2020 US Elections, and (2) 2021 Capital Riots.

Table 5.3 presents the resultant CSI-Network score and the global clustering coefficient scores of the CSA and Human agent classes. There is consistency between the global clustering coefficient scores and the dominant group of users. This reflects that the group that has high tendency to form clusters in the resultant network, which indicates that they have higher synchronicity among each other.

For example, the global clustering coefficient of the network formed for 2021 Capitol Riots is 0.783, which is extremely high. In contrast, its CSI-Network is 9.05, which is on the lower end of the spectrum. However, when we examine the clustering coefficients of CSA and human partitions separately, we see that that CSA partition has a clustering coefficient of 0.785, while the human users have a clustering coefficient of 0.245. Therefore, during the Capitol Riots event, the bots were actively synchronizing with each other, resulting in high clustering formations in the resultant synchronized network graphs. In contrast, human users synchronize lesser in all multiple dimensions, resulting in the lower CSI-Network scores formulation.

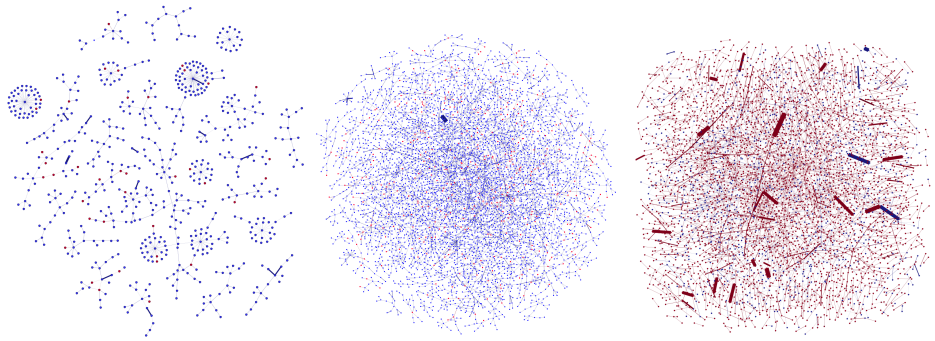
Event	**CSI-Network**	**CSA**	**Human**	**Dominant Group in Visualization**
2018 Black Panther	2.81	0.569	**0.998**	Human
2021 French Protests	4.16	0.177	**0.381**	Human
2020 ReOpen America	12.42	**0.621**	0.267	CSA
2020-2021 Coronavirus	2.57	0.330	**0.634**	Human
2020 US Elections	33.73	0.216	**0.234**	Human
2921 Capitol Riots	9.05	**0.785**	0.245	CSA

Table 5.3: Comparison of CSI-Network Scores against Global Clustering Coefficient scores derived from the Synchronized Network Graphs. The scores are split up by CSA/human classes and are consistent with the dominant group in the network graph visualizations. (Published in [206])

Figure 5.10 visualizes the all-communication graphs of the networks studied in this work. For many events, a low CSI-Network score corresponds to a low global clustering coefficient score, which reflects that the resultant networks does not form clusters very well. Indeed, if we were to examine the 2018 Black Panther event, it has both a low CSI-Network and global clustering coefficient score, which corresponds a rather disjointed network with few clusters. However, events like the 2020-2021 Coronavirus, 2020 US Elections and 2021 Capitol Riots have CSI-Network and global clustering coefficient scores that are on two opposite ends of the spectrum. This inconsistency of scores points us to investigate into the network further, as it indicates that there are some partitions of the network that are more clustered than others.

5.5 Network Impacts

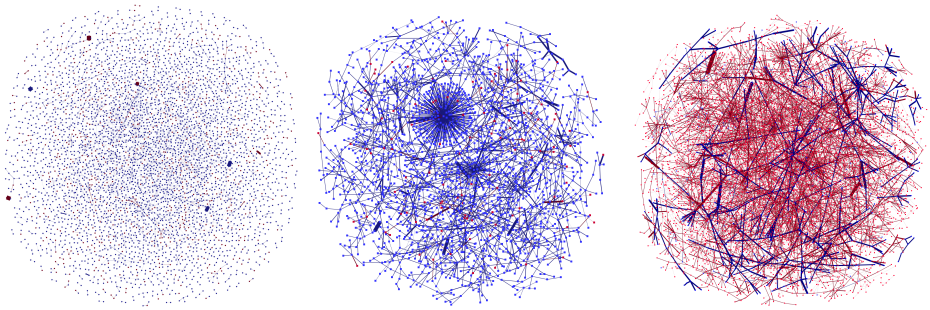
Beyond detecting and analyzing agent-agent coordination, we seek to measure the impact of such coordinated behavior on the network. We do so by a measure we termed "stance flipping". A stance is an expression of an opinion towards an entity, usually characterized by "pro-" or "anti-". Social influence characterizes the change of an individual's stances towards a topic in a complex social network environment. Two factors often govern the influence of stances in an online social network: (1) endogenous influences driven by an individual's innate beliefs through the agent's past stances; and (2) exogenous influences, which offer important clues to user susceptibility, thereby enhancing the predictive performance on stance changes or flipping.



(a) **2018 Black Panther Movie**
CSI-Network=2.81
Network density=1.61E-3

(b) **2021 French Protests**
CSI-Network=4.16
Network density=4.8E-5

(c) **2020 ReOpen America**
CSI-Network=12.42
Network density=1.59E-4



(d) **2020-2021 Coronavirus**
CSI-Network=2.57
Network density=1.69E-5

(e) **2020 US Elections**
CSI-Network=33.73
Network density = 8.90E-4

(f) **2021 Capitol Riots**
CSI-Network=9.05
Network density=9.7E-4

Figure 5.10: Synchronized Network Graphs. Nodes are users. Red nodes are CSAs and blue nodes are humans. Links two users represent synchronization between them. Link widths represent the degree of synchronization. Graphs have been pruned to show nodes that synchronize with at least 5 different users to depict only the core structure of users that synchronize very frequently.

Methodology Overview We design a social influence model to investigate the variables that have an effect on the network impacts of stance flipping. Each agent has some fixed innate stance and a conviction of stance that reflects the resistance to change and agents influence each other through the social network structure. We use this model as a stance flipping prediction problem to identify X agents that are susceptible to stance flipping towards the coronavirus vaccine (i.e., from pro-vaccine to anti-vaccine).

The Social Influence Model We adapt the Friedkin-Johnsen social influence model [95] and describe the formation of a stance in terms of an agent's innate static variables and the interpersonal influences from other agents in the network.

Agent stance. We define agent stances Y with the following model: $Y_{agent} = XB$, in which Y is an agent stance outcome score, X is a $1 \times k$ matrix of scores on k endogenous

and exogenous variables of the agent and B is a $kx1$ vector of coefficients giving the effects of each of the endogenous variables. In our study, we used agents' linguistic cues as endogenous variables and network values as exogenous variables.

The Base Influence Model. Equation 5.4 represents the base influence model, which estimates the impact of an agent's past tweets and the influence from the agent's neighbors on the agent's stance. Neighbors are other agents that have made communication with the agent in focus. The opinions of these neighbors in the social influence model directly affects an individual's opinions.

$$I = \frac{1}{n}\sum_{i=0}^{n} Y_{\text{1st deg neighbors}} + \frac{1}{n}\frac{1}{m}\sum_{i=0}^{n}\sum_{j=0}^{m} Y_{\text{2nd deg neighbors}} \tag{5.4}$$

Stance Strength. We define the effect of stance strength, which alludes to the fact that the more an agent expresses a stance, the stronger the belief in the stance. This is defined as: $\gamma = \frac{|s_{final}|}{|s|} \times w$. Stance strength is the proportion of the final stance s_{final} is expressed against the number of expressed stances s, multiplied by the variable importance value w_s. This is added to the base model as: $Y_{agent} = \gamma X_* B_*$, where γ is a scalar representing the agent's stance strength and its importance.

Connection. Connection C is the proportion of neighbors that support an agent's stance. Connection represents opinion similarity between the agent and agent neighbors, which lends strength to the stance the agent expresses.

$$C_{agent} = \frac{\text{\#neighbors with same stance}}{\text{\#neighbors}} \tag{5.5}$$

Reciprocity. Reciprocity R is the two-way interaction between two agents. The higher the reciprocity value, the closer the agents are in a friendship, leading to a higher influence on the agent.

$$R = 2 \times \text{\#reciprocal interactions} \tag{5.6}$$

Susceptibility Score. We define a susceptibility score $S = (I - Y_{agent})^2$, which characterizes the difference in the score between the agent's stance and the influences from the variables. The higher the susceptibility score of agent i, the more likely the agent will flip its stance due to social influence. The agent i will flip stance if $S_i \geq \epsilon$. For the base model, we set ϵ at 10% of the number of agents.

Experiments. In our experiments, we described the stance towards the coronavirus vaccine. We used the 2021 coronavirus dataset, and filtered for tweets that mentioned the vaccine with #vaccine. We only investigated agents who have more than one tweet in order to have changes in vaccine stances. For these agents, we leave out each agent's last stance, and use the collected historical data to predict the final stance.

Table 5.4 presents the results of incremental experimental runs on the dataset. Our final stance flipping model outpefroms all the other models with an accuracy score of 86%, showing that a combination of all the identified factors is inportant to the influence of agent stances. Statistical tests of the model prediction results identified that factors important to stance flipping prediction are: the use of 2nd-degree neighbor information, stance strength and connection information, while the reciprocity factor does not achieve a significant improvement in results.

Model #	Model	Accuracy
Baseline	Decision Tree	0.38
Model 1	Base social influence model	0.37
Model 2	Model 1 + 2nd deg neighbor information	0.48*
Model 3	Model 2 + stance strength	0.70*
Model 4	Model 3 + connection	0.75*
Model 5	Model 4 + reciprocity	0.86
Ablations		
Model 1 - network	Base social influence model without network variables	0.17*
Model 1 - linguistic	Base social influence model without linguistic variables	0.19*
Bots only	Model 5 with only bot agents	0.73*
Non-Bots only	Model 5 with only non-bot agents	0.67*

Table 5.4: Results of Social Influence Models. * indicates a significant difference at the $p < 0.05$ level. For the models, the significant testing was performed against the previous model in sequence, and for the ablations, the significant testing was performed against Model 1. (Published in [204])

Conditions of Stance Flipping To isolate the conditions that result in the stance flipping behavior of agents, we perform a statistical t-test between the neighborhood of agents that flip stances and the agents that do not. These results are presented in Table 5.5.

An agent can be influenced by the opinions of the network of neighbors around him. The accuracy of predicting an individual's stance flipping tendency increases after the addition of second degree neighbor and reciprocal ties. This shows that interaction information such as two degrees of neighbors and the connectivity of neighbors contribute significantly to the influence of an agent's stance.

For agents that flip stances, the participation of semantic coordination is an indicator. This means that the more an agent's neighbors participate in semantic coordination with other agents, the more the neighbors seem to agree with each other through using the same hashtags. When these neighbors are of the opposite stance as the agent, the more likely the agent is to flip stances. Within our dataset, 0.1% (n=6791) agents engaged in semantic coordination. The number of times an agent that participates in semantic coordination is between 80.54 and 1704 times. This observation signals the peer pressure of an agent's neighborhood on the expression of opinion.

Figure 5.11 visualizes the neighborhood of the agents that flip stances in an all-communication network. This graphs first show the stances of the agent and the neighborhood on the left, then the agent's engagement in semantic coordination across the entire dataset.

Of the agents that flip stances, a significantly larger percentage of these agents are bots ($p = 2.14e^{-13} < \alpha = 0.05$). Cyber Social Agents have lesser conviction and a larger proportion flip stances (6.6%). CSAs flip even with a fewer number of neighbors that have the opposite stance. In contrast, humans have more conviction and a smaller proportion of humans flip stances (2.7%), and they require more neighbors of the opposite stance to flip. For CSAs that have the word "bot"

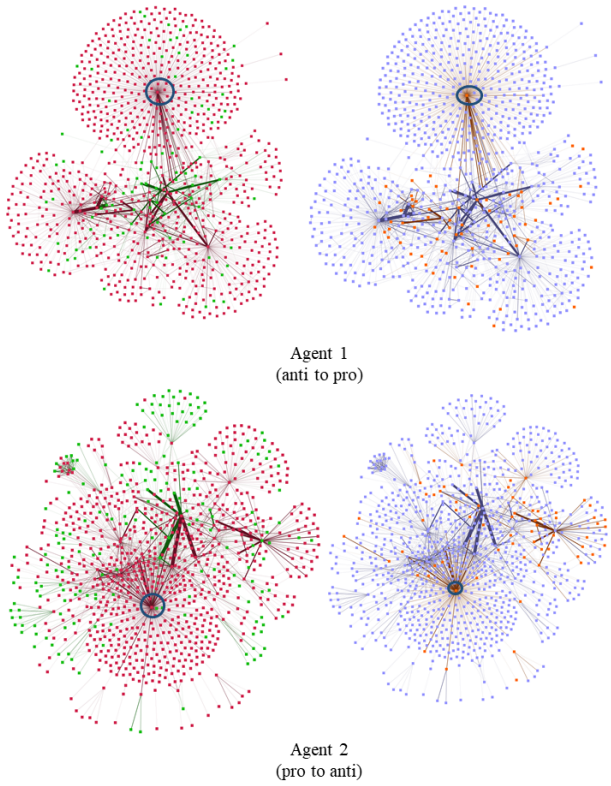


Figure 5.11: Network interaction graphs that our model correctly predicts to flip. Green nodes are pro-vaccine agents; red nodes are anti-vaccine agents; orange nodes are agents found participating in collective expression through hashtags; purple nodes are agents that are not found participating in collective expression. The agent is circled in blue and the color of the agent stance is the stance before the flip. (Published in [204])

Criteria	Agents that do flip stances	Agents that do not flip stances	p-value
Proportion of bots	0.452	0.293	2.14e-13*
Proportion of neighbors that are bots	0.352 ± 0.372	0.358 ± 0.280	0.051
Proportion of neighbors of the opposite stance	0.409 ± 0.443	0.196 ± 0.271	0.011*
Proportion of neighbors participating in semantic coordination	0.0389 ± 0.068	0.0302 ± 0.115	0.027*
Proportion of neighbors participating in semantic coordination and are of opposite stance	0.0207 ± 0.0996	0.0136 ± 0.0350	0.0078*

Table 5.5: Comparison of agents that flip stances and do not flip stances. Fraction of neighbors that are 1- of 2-degree away from these agents that meet various criteria are compared. * denotes a p-value that is significant at the 0.05 significance level. (Table published in [204])

in their account name, the proportion of these agents flipping stances is five times higher than the population proportion.

5.6 Conclusion

This chapter examines multiple types of agent-agent coordination strategies, the measurement of the extent of agent coordination, and also presents the possible network impacts that might result from extensive coordination. Agent-agent coordination in the cyber social space makes use of platform affordances and can be strong levers for online movement.

The identification of coordinated networks through the Synchronized Action Framework and the calculation of the Combined Synchronization Index has been integrated into the ORA-Pro software, which has been taught to CMU students and executives that attend the CASOS Summer Institute.

We measured the network impact through the flipping stance phenomenon, modeled through a social influence model. This model uses a combination of an agent's endogenous and exogenous variables, and reveals that the usage of pronouns and the agent's position in the network are important factors in influencing others. CSAs are more prone to flipping stances than humans, as observed by them requiring fewer neighbors of the opposit stance before flipping. This opens avenue for further characterization and understanding of the neighborhood CSAs operate in and their resilience to opinion changes.

Limitations of our study of agent interactions include:

1. The usage of social media API calls (i.e. the X API) means that we retrieve only 1% of the tweets, so there may be more coordinating agents, types of network profiles and other discourse topics which were not captured in the dataset. This might have reduced the

number of agents and network profiles that were found in our analyses.

2. We primarily studied one network profile: the star network profile. There are multiple other types of network profiles that CSAs and humans can form which are left out in our analyses.
3. Our definition of coordination is limited to the investigation of single-action coordination, where actors perform the same singular action within a short time window. For example, this would be two tweets with the same hashtag. There are other forms of coordination such as two actors performing different actions in sequence during a short time window, which should be considered.
4. We primarily studied only one network effect: the flipping stance effect. Other types of network effects such as the changes in network density and other network metrics should be profiled.

Future work involves profiling a variety of network motifs beyond the star network motif, and how both CSAs and humans harness those motifs to spread their messages. In terms of coordination, future work involves the investigation of higher-order actions, which are actions combining multiple singular actions, e.g., two tweets with the same hashtag and URL. Higher order actions that combine more than one singular action can provide a source of deliberate coordination rather than coincidental synchronicity. Future work also involves expanding the Combined Synchronization Index to factor in user synchronization across moving time windows. This provides a way to perform effective comparisons of user-user synchronization and the nature of synchronization across different time periods.

Chapter 6

Social Simulations of CSAs & Humans

6.1 Introduction

Agent-based social simulations are a critical methodology that bridges the theory and empirical observations with the constraints of real-world of experimentation. Real-world experimentation on populations of CSA interventions and resultant population dynamics can be logistically, strategically and ethically untenable, so simulations provide an indispensable alternative for testing hypotheses and evaluating the dynamics of influence across the network over time. They model the interplay between heterogeneous agents such as the different archetypes of CSAs and humans, and showcase how macro-level social outcomes like polarization emerges from micro-level agent behaviors. We modeled three social scenarios: a predictive model that analyzes which agents, when perturbed, are most influential to changing a population's stance, a theoretical model that analyzes how good CSAs can be used for socially beneficial outcomes, and an illustrative model of the interaction of CSAs and humans.

This chapter constructs social simulations the following guiding **research questions**:

1. How can we design and generate realistic CSAs that faithfully capture the behavioral, linguistic and interaction patterns that are observed empirically on social media platforms?
2. What are the characteristics of a CSA that when triggered, are most likely to contribute to the success of influence operations?
3. How can CSAs be used for socially beneficial outcomes (i.e., fact checking, good messaging, counter-messaging)?
4. How can we illustrate the interactions between CSAs and humans with a social simulation?

6.2 Related Work

Social simulation provides a foundation for understanding how individual behavior and interactions generate collective phenomena. This technique of social simulation is especially useful in environments where direct experimentation is infeasible. Traditionally, social simulation used Agent-Based Models(ABMs), which emphasize structural and behavioral rules that govern interactions. Classic work like the Schelling and the Axelrod models show that simple behavioral

rules can result in complex macro-level patterns like segregation or cooperation. Nowadays, generative AI is integrated into the simulations to construct LLM-based simulations that are linguistically rich for conversational realism.

Agent-Based Models Agent-Based Models (ABMs) have been a classic methodology for simulating social networks and understanding collective behavior in digital environments. Models such as the Barabasi-Albert preferential attachment model have provided insights into network formation patterns [27]. ABMs have been used to explain viral information spread and cascade behaviors in online networks, and network evolution dynamics that include the formation and breakage based on homophily and social influence mechanisms and of polarized clusters and echo chambers [36, 50]. ABM simulates successive agent-agent and agent-environment interactions across time, allowing for the observation of emergent behaviors that connect micro-level individual agent behavior to macro-level patterns [36]. X feeds modeled as a discrete event simulation can aid studies of the emergent behavior of two bot-based disinformation maneuvers, bridging and backing, which revealed that bots are only effective when correctly embedded in the network [33]. Further, [21] evaluates how the beliefs of agents can be influenced in an X network with malicious users that spread misinformation through an SIR-epidemiological model setup.

However, traditional ABMs face limitations in generating realistic content because they rely on rule-based interaction heuristics rather than linguistically-generated capabilities that characterize social media discourse. Pure ABM agents do not produce or interpret languages, and therefore could not simulate conversational discourse. This gap motivates the integration of LLMs into social simulations to endow agents with expressive, goal-directed communication studies.

LLM-Based models Advances in LLM-based simulation have enabled realistic agent communication and behavior synthesis. These social simulations range from dialog-driven (e.g., social interaction, question-answering, and game-based) to task driven (e.g., software development) [197]. Such simulations have crafted artificial societies with LLM-powered agents that are capable of autonomous memory and planning, and demonstrate emergent social behaviors [239]. OASIS is a social simulation that models the information propagation dynamics in X and the herd effect in Reddit with a total of one million agents [323]. However, many existing LLM-based simulations lack the explicit modeling of social structure, role differentiation and network-level interactions. Agents often operate in sandboxed contexts without persistent relationships or inter-agent dependencies that mirror real-world social media dynamics. Our work builds upon these foundations by introducing a framework of simulations of CSAs as LLM-based agents embedded with personality traits and operational strategies that are situated in a social network.

6.3 Designing realistic social simulations

To design realistic social simulations, we adopt an agent-based modeling framework. Each individual user or account is modeled as an autonomous agent with its own properties, such as persona, behavior rules and interaction patterns. Many agents aggregated together forms a group

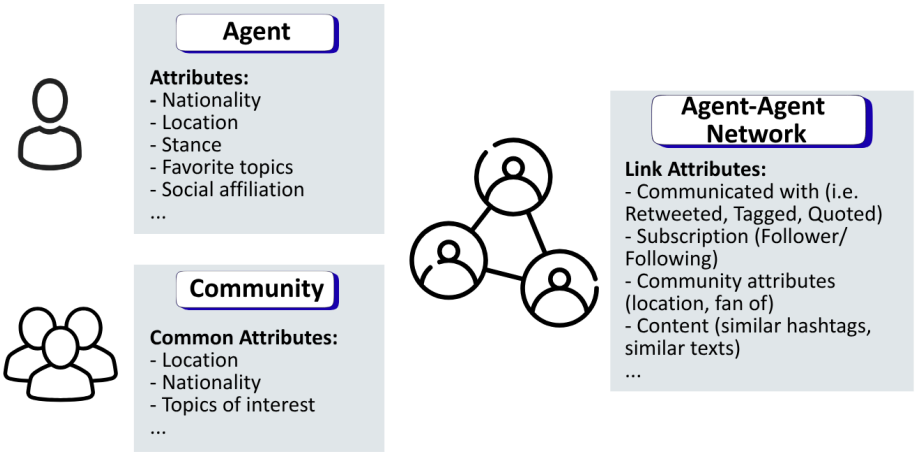


Figure 6.1: Illustration of the building blocks of a social simulation

or a community of agents, which have shared attributes like social identities, narratives or locations. Agent behaviors can also be shaped by group attributes. Agent-agent interactions such as posting, replying, quoting and retweeting, occur through a network topology that creates an information flow. This information flow is represented by links, and links themselves have attributes. This relationship between agents, community and the network is illustrated in Figure 6.1.

The emergent global behavior of the system then arises from the micro-level agent decisions and the macro-level agent-agent networked interactions. Further, to realistically model online social behavior, we use a hybrid model and integrate multiple methodologies, including: (1) mathematical components such as probabilistic activation functions, preferential attachment network-formation models; (2) heuristic processes such as posting frequency, memory and agent fatigue; and (3) Large Language Model (LLM)-based generation for persona construction and for agents to create text posts consistent with their persona and realistically react to posts by other agents.

Here are a list of methodological questions that we considered in building our CSA-Human social simulation.

Simulation Scope: this step defines the overarching context, scope and goals of the simulation, with the following questions:

1. What are we modeling in this simulation? Examples are information diffusion, agent coordination, ideological polarization, ideological convergence, community formation etc.
2. What is the goal of the simulation? Is the goal of the simulation theory testing, predictive or illustrative?
3. How long will the simulation run and what is the temporal resolution? A time step in a simulation is called a "tick". Examples of simulation time can be: 30 days with tick intervals of 1 hour, 3 months with tick intervals of 1 day.
4. What is the stopping condition? Is it a fixed time window (i.e., stop the simulation after 30 days), a threshold (i.e., stop the simulation when the number of malicious agents reaches 50% of the total number of agents), or an equilibrium dynamic (i.e., stop the simulation

when there are more malicious than good humans).

5. What are the outcome measures (i.e., the measurement to be derived from the simulation)? Examples of outcome measures can be the ratio of malicious:good agents or the proportion of agents that have changed their stance.

Simulation Initialization: this step defines the agent composition, the initial network topology and the temporal window of the simulation, with the following questions:

1. Agent composition: defining the types of agents involved and the initial distribution of each type of agent
   (a) How is the simulation initialized? Is the simulation empirically seeded (i.e. the initial networks and agent properties are based on real data), synthetically seeded (i.e. the initial networks and agent properties are based on generated and hypothetical data), or a mixture of both?
   (b) What types of agents are involved? Humans, Cyber Social Agents, institutional accounts, media accounts?
   (c) What is the distribution of agent types (i.e., there are 100 humans, and 80 CSAs)? What is the distribution within each agent type? For example, the number of each type of CSA is uniformly distributed; or that 30% of the CSAs present are amplifier agents, 20% are chaos agents, while the rest are news agents. Is this distribution empirically grounded from observed data or scenario-based?
   (d) What are the attributes of each agent? Examples of attributes are: stance towards specific entities, posting frequency, topic interest, susceptibility, influence (i.e., influential user, media outlet)
   (e) How are the attributes of the agents derived? Are they heuristically generated (i.e., tone, range of day of post), mathematically generated from a distribution (i.e., posting frequency), or procedurally generated using LLMs (i.e., name, narratives)?
   (f) Which attributes of the agents are static (i.e., does not update), and attributes are dynamic (i.e., updates with interactions or time) through the simulation? If there are dynamic traits, what is their update function (i.e., linear combination of inner beliefs and external interactions)?
2. Network topology: defining the initial interactions and network structure of the agents
   (a) How are the agents initially connected? Mathematical models of network topology include the erdos-renyi random model, the Barabasi-Albert preferential attachment model, the Watts-Strogatz small-world model. The initial network topology can also be a topology derived from empirical interaction data.
   (b) What is the probability of connections in the initial state? Is the probability arbitrarily set (as most erdos-renyi and small-world networks are)? Or is the probability dependent on the homophily of agent properties (i.e., belief, topic, structural role)?
   (c) Can the agents form or dissolve interaction ties during the simulation? What are the rules of formation or dissolution? For example, an agent following a new agent is a

tie formation, while an agent un-following another agent is a tie dissolution.

Agent Activation and Behavior : this step defines when the agents are active to act, and the actions they perform when they are active.

1. Activation function: defining whether the agent will act during a time tick. These activation functions are different for each agent archetype (human, Cyber Social Agents).
   (a) Do agents act determistically (i.e., time-defined schedule) or probabilistically (i.e., geometric distribution, fatigue decay)? Agents like the Announcer Agent will act in a deterministic manner, posting every few time ticks. In contrast, human agents tend to act probabilistically and are affected by fatigue and decay of information.
   (b) What kinds of actions can agents perform (i.e., post, quote, reply, retweet)? For example, human agents can perform all types of actions, amplifier agents will only perform the retweet action, and content generation agents will only perform the original post creation function.
2. Agent interaction: defining who the agent will interact with if it is active at a time step
   (a) What are the other agents that are available for interaction to an agent? These can be the network neighbors, the followers/ followees, the group members, or authors of globally visible posts. These agents can also be time-bounded (i.e., agents that posted in the last n steps).
   (b) How do agents select who to interact with? Example algorithms for selection include: neighborhood-based interaction (i.e., 1-hop, 2-hop, group members), topic-based selection (i.e., seeded topic, recent topic, popular topics), preferential attachment (i.e., to high degree agents, influential agents, or group leaders), or homophily (i.e., other agents that that have similar topics, stance or community).
   (c) How do agents decide their interaction type? This may be a random selection, or an agent-specific action. For example, amplifier agents will always choose the retweet interaction, and information correction agents will choose a reply or quote interaction.
3. Narrative dynamics: describes the narratives the agents produce and whether they evolve
   (a) What narratives or topics do agents produce or engage with? Are these narratives predetermined during the Simulation Initialization step, or are they determined by the agent interaction (i.e., preferential attachment, recommendation algorithm) during the tick?
   (b) Do narratives evolve across time? If they do, how do they evolve across time?
   (c) If the agent needs to generate a post using an LLM (i.e., create post, quote post, reply to post), what are the parameters of the post generation? Possible parameters can be LLM-temperature, emotion, tone of agent (i.e., journalistic, casual, conversational).
   (d) What is the context that the LLM is conditioned on? This can be the agent's prior posts, or prior posts related to the topic in the simulation, posts by network neighbors, or initialized narratives.

4. Agent persona consistency: defining whether the agent's persona will change or evolve across the simulation timeframe
    (a) Does the agent maintain persona consistency? If not, which agent attributes will change over the simulation? A simulation that studies the possibility of stance polarization or convergence will result in the agent "stance" attribute to change over time.
    (b) What are the mathematical equations or heuristic threshold that governs the change? For example, the stance of an agent can change from pro- to anti- if it had interacted with $n = 50$ agents with the anti- stance.

Post-Simulation Validation: this step describes the procedures taken to ensure the validity of the simulation

1. Face validity: assessing whether the simulation looks realistic
    (a) Is there behavioral realism of the agents? Are the posting frequencies, reaction times and actions, interaction styles similar to real user behavior?
    (b) Is there linguistic realism? Are the generated posts comparable to real social media texts? Do the psycholinguistic values (i.e., stance, sentiment, number of pronouns, number of words) resemble empirical datasets?
    (c) Do the different archetypes (i.e., human, CSA) exhibit recognizable behaviors as intended?
2. Construct validity: assess whether the internal logic of the simulation accurately operationalizes the underlying social theory
    (a) Do the activation functions correctly reflect different archetypes? For example, are the announcer agents deterministic, do cyborgs have alternating bot-human characters?
    (b) Are the implemented constructs (i.e., homophily, preferential attachment, fatigue) consistent with theory?
    (c) Do the micro- and macro- level outputs correctly measure or capture the desired social constructs? Are the outputs sensitive in the expected direction when the parameters (i.e., proportion of CSA, proportion of topics) are varied?
    (d) Does validation against real dataset or manual inspection support the underlying theoretical constructs?
3. Structural validity: assessing whether the simulation reproduces known structural properties of social networks
    (a) Does the network topology match empirical networks? Examples of measurement for network topology include: degree distributions, clustering coefficients, centrality values, modularity etc. to be within realistic ranges of empirical networks.
    (b) Does the system produce recognizable structural phenomena (i.e., polarized clusters, community segregation)?

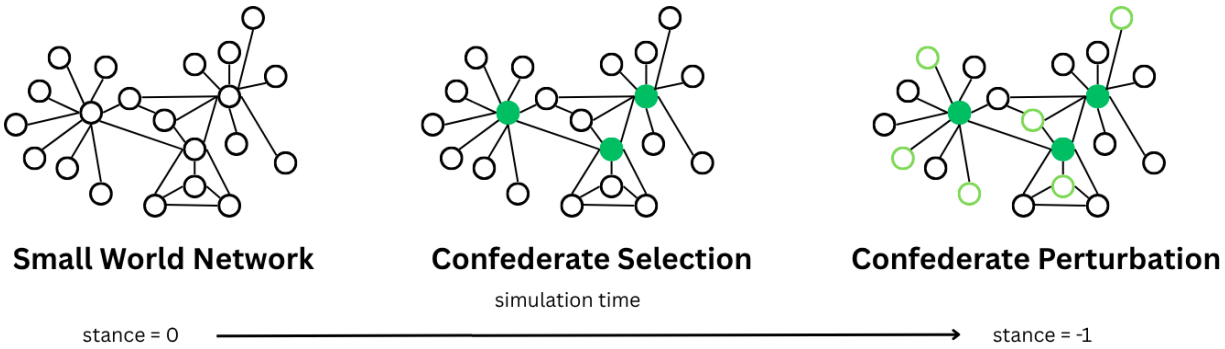


Figure 6.2: Overview of methodology for Stance Perturbations in a Small World Network

6.4 Perturbation of CSAs for network influence

In our first simulation model, we analyzed the circumstances which social influence operations are likely to succeed. To do so, we intentionally provoked the stances on a simulated network and analyzed the trade off between perturbing stances and maintaining influence [50]. The process of a group of Confederate agents perturbing the network simulates the scenario of an influence operation, where the network was perturbed with the intention to create an opinion shift.

6.4.1 Simulation Setup

Overview We examined the effect of Confederate agent selection and stance perturbation strategies on the overall stance in the network. Stance changes are modeled by a co-evolutionary social influence model. This simulation was constructed using the Construct pipeline [49]. Figure 6.2 summarizes the methodology used for this stance perturbation work.

We first created small-world networks with N agents, ranging from $n = 10$ to $n = 150$. Each agent had two attributes: stance y and susceptibility s. Stance is initialized to $y = 1$ to reflect the initial state of consensus. Susceptibility is initialized to $s \sim N(0.1, 0.1)$. Next, we choose a Confederate selection strategy and a Perturbation strategy. Confederate agent susceptibility $s_{conf} = 0$, because the Confederate agent is being deliberately assigned with a stance. We varied the percentage of agents that are Confederates across the experiments. Finally, we ran the simulation until the mean stance change of all N non-Confederate agents over the previous $t = 30$ timesteps was less than 0.001. We calculated $\hat{\mu}_y$, the mean stance of non-Confederate agents at convergence. The lower the mean stance, the better the success of Confederate perturbations in driving network stance from $y = 1$ to $y = -1$.

Co-evolutionary model of stance change We used a co-evolutionary model to represent the change of stances by each agent. Our model had two portions: (1) the exogenous effect based on Friedkin's model [95] was represented by the stance update equation, where the agent stance was incrementally nudged towards the average stance of those the agent is influenced by. Given the influence matrix W and the diagonal susceptibility matrix A scaled by a stance learning rate $\alpha = 0.001$, this portion is represented as Equation 6.1; and (2) the endogenous update rule based on the Hopfield model [172], represented in Equation 6.2. In this endogenous stance update

portion, we introduced an inverse relation between W & y based on the homophily process, to enable influence to co-evolve with stance. $\lambda = 0.01$ is the influence update rate, or the rate of structural learning as in the Hopfield model.

$$y(t) = AW(t)y(t-1) + (I - A)y(1) \tag{6.1}$$

$$W(t+1) = \lambda y_t y_t^\top + (1 - \lambda)W(t) \tag{6.2}$$

Confederate selection We tested three strategies of selecting Confederates:

1. Maximum influence: Confederates were the most influential agents according to the weighted out-degree of the influence matrix at $t = 0$. That is, $W_{max} = argmax_j \Sigma_i^N W(0)_{ij}$.
2. Minimum susceptibility: Confederates were the least susceptible agents. That is, $A_{min} = argmin_j A_{jj}$
3. Random selection: Confederates were selected at random.

Confederate Perturbation Strategies We tested three strategies of Confederate Perturbation:

1. Conversion: this strategy nudged the overall network towards the desired stance ($y = -1$). We used a continuous function that scales the magnitude of perturbation by the current level of influence that the Confederate has. When the Confederate's influence was low, its stance perturbation was less aggressive; when the Confederate's influence was high, the stance perturbation was more extreme. Given w_i^g as the global influence factor of Confederate i on all $N - i$ non-Confederate agents, this strategy is formally expressed in Equation 6.4.

$$y(i,t) = \mu_y^g + w_i^g * (-1 - \mu_y^g) \tag{6.3}$$

$$w_i^g = \Sigma_j^N w(j,i) / \max_{k'} \Sigma_j^N w(j,k') \tag{6.4}$$

2. Conservative: the stance was perturbed if the Confederate's influence was above a threshold θ. Whenever influence drops below θ, we set the stance to μ_y; the average over N non-Confederate stances at time t: $\mu_y = \Sigma_i^N y(i,t)/N$. Formally, this strategy is expressed in Equation 6.5.

$$y(i,t) = \begin{cases} \mu_y & \Sigma_j^N w(j,i) \leq \theta \\ -1 & \Sigma_j^N w(j,i) > \theta \end{cases} \tag{6.5}$$

3. Cascade: the cascade strategy was similar to the conversation strategy, except that it calculated the influence of the Confederate across its ego-network rather than globally. We ranked all agents in the ego network by the influence the Confederate has over them, and summed the influence weights from the top M most influenced agents. In our experiments, $M = N/10$. Formally, this strategy is expressed in Equation 6.6.

$$y(i,t) = \mu_y^l + w_i^l * (-1 - \mu_y^l) \tag{6.6}$$

$$w_i^l = \Sigma_j^M w(j,i) / \max_{k'} \Sigma_j^M w(j,k') \tag{6.7}$$

6.4.2 Simulation Results

Tipping points Minority stances of Confederate agents can create a tipping point, which is the point that changes the overall stance of a network. With this tipping point, the stance of the network converged to the stance of these minority Confederate agents. Figure 6.3 shows that only 20-25% of the Confederate agents were required to tip the overall stance of the network. This range of agent population held across all Confederate selection strategies.

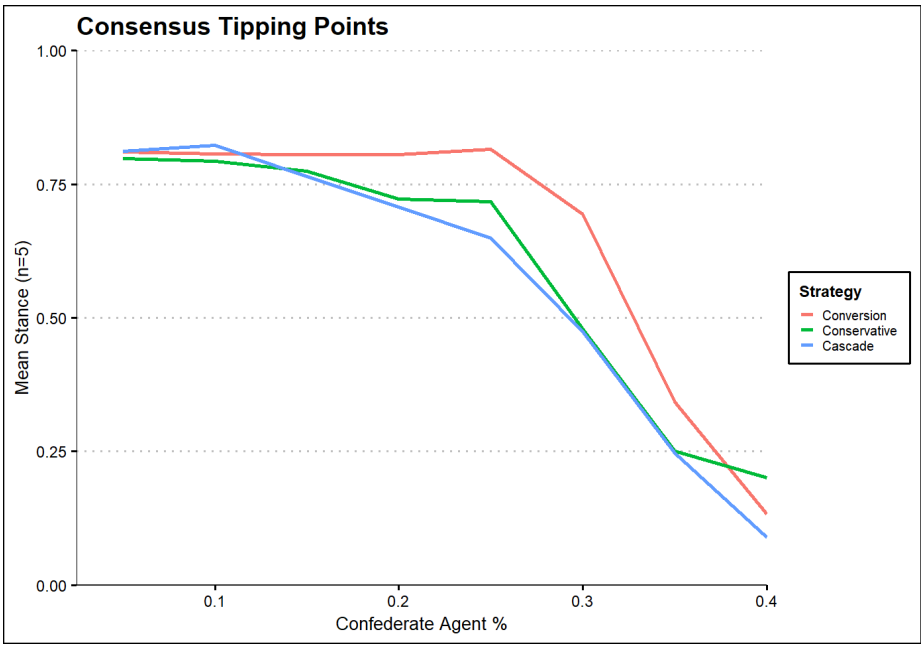


Figure 6.3: The overall stance of the network can converge to a desired stance when there are only 20-25% of Confederates (published in [50])

Best confederates Influential agents were the best Confederates to cause changes, and the most effective and widespread change happens with the cascading of local ego networks. This is analogous to how the influential singer-songwriter Taylor Swift can sway people to vote in the elections with the "Taylor Swift Effect" [77]. Figure 6.4 demonstrates this with the Confederate selection strategy plots.

Optimal perturbation strategy The optimal strategy for changing of stances involved nudging of agents that were in a Confederate's direct neighborhood, or its ego-network. Figure 6.5 illustrates how this strategy resulted in the lowest mean stance change. Note that the Conversion strategy performed the worst, showing that global nudging are not optimal, and precise targeting and neighboring influence fared better.

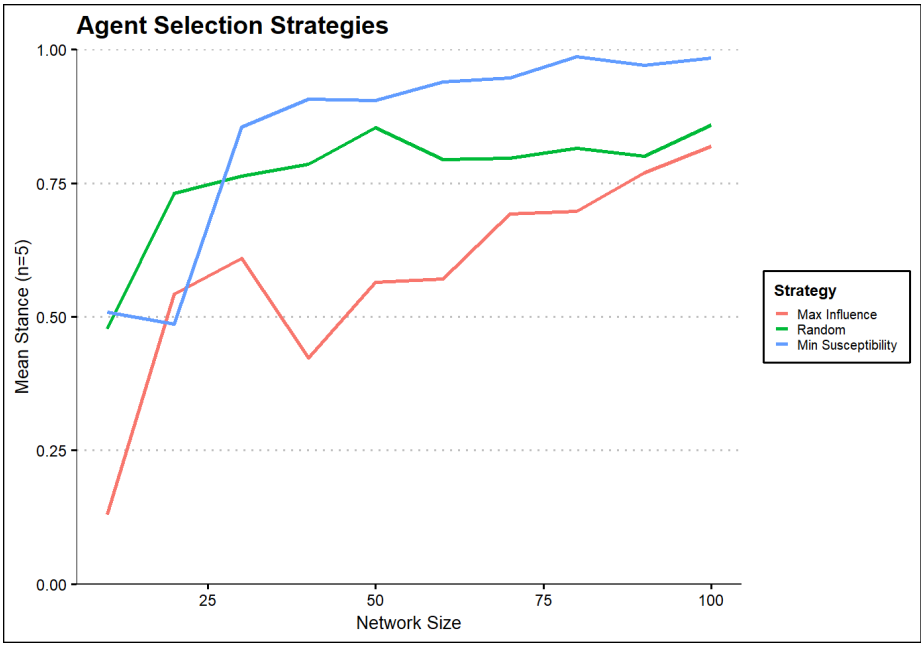


Figure 6.4: Comparison of the effect of agent selection strategies for Confederates. The Influential agents are optimal for stance changes. A lower mean time is better. (published in [50])

6.4.3 Simulation Validation

By face validity, our simulation results were plausible. Intuitively, people within a group would eventually converge to a consensus of stances, which is observed through the tipping point and stance convergence within our work. Influential people are best positioned to affect other people's stances, and the optimal strategy to change people's stances is to begin by effecting a change towards people that are close to you, i.e. your ego-network.

By theoretical validity, the tipping point observed in our experimental runs where 20-25% of the network are required as Confederates to change the stance of the network is similar to past studies in the literature [53]. In a past study where there are targeted attacks on scale-free networks, higher degree nodes have larger effects on the overall network, which provides the same observations as our experimental results [253].

By pattern validity, our results matched stylized facts that can be used to determine known phenomena. The stances of users in the network eventually come to an equilibrium because users achieve social conformity with each other. This is also observed within the principles of homophily, giving rise to the saying "birds of a feather flock together".

By process validity, the process of a group of Confederate agents perturbing a social network simulates a real-world scenario where some of the agents can intentionally want to change or tweak their stances, and in doing so, affect the overall network.

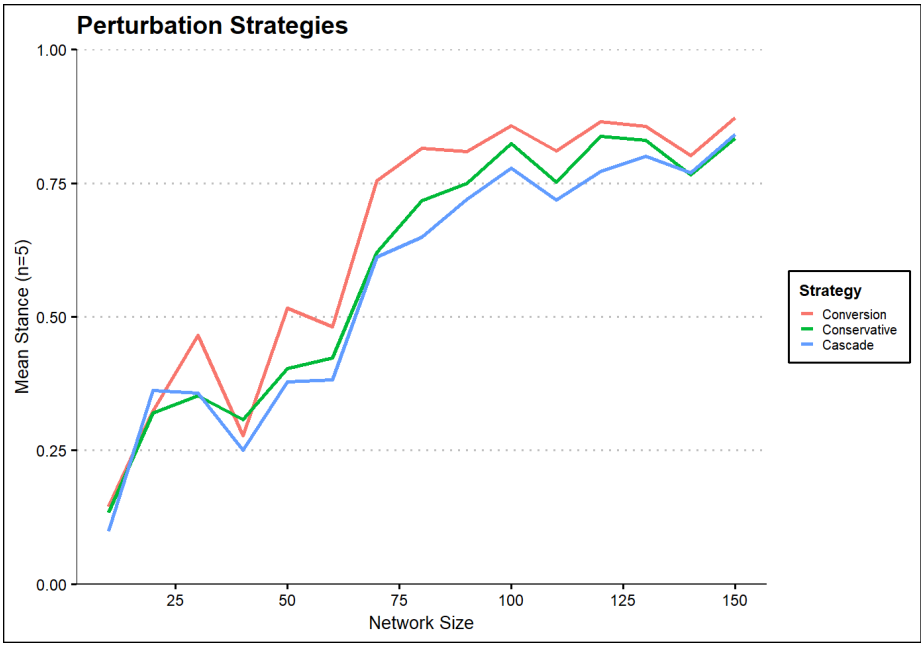


Figure 6.5: Comparison of perturbation strategies. The cascade strategy is optimal for stance changes. A lower mean time is better. (published in [50])

6.5 Simulation of useful CSAs

In our second simulation model, we simulated how useful CSAs can mitigate the formation of conspiratorial societies. Societies can become a conspiratorial society where there is a majority of humans that believe, and therefore spread, conspiracy theories. This simulation studies how useful CSAs can spread conspiracy theories in an automated fashion to result in a conspiratorial society, and the effectiveness of two types of automated interventions from Good Agents and Information-Correction Agents. This simulation was constructed using the NetLogo software [284].

6.5.1 Simulation Setup

Overview This model simulated a dynamic social media information ecosystem where agents continuously generate, consume, and propagate information through their ego network connections. In doing so, we investigated the conditions under which useful CSAs can effectively counteract the formation of conspiratorial societies. A flowchart of the simulation logic is presented in Figure 6.6.

The following steps summarizes the simulation main logic.

1. The model was set up with a Small World network initialized with $n_h = 1000$ Human agents and $\alpha_1 \times$(Humans) number of Bad CSAs.
2. The routine then enters a cycle. This cycle consisted of three stages that repeat until a termination condition is met.

(a) The Information Generation stage generated two types of information: Good Info and Bad Info. All agents generated Info (information) with a probability of P_g. Whether the agent generated Good or Bad Info depends on the state (i.e., Good, Bad) of the agent. CSAs generated 2 pieces of Info while Humans generated 1 piece of Info.

(b) Next, we have the Information Consumption Stage, where agents consumed the Info that they received in the previous tick, each with a probability of P_c. A Human consumed all Info, while CSAs just held the Info in their memory space. The consumption of Info represents the readership and processing of the information. CSAs do not consume Info because they were pre-programmed with their state and are not influenced by incoming information.

(c) The third stage was the Information Propagation stage. Each agent (the ego) sieved through the messages they received in the previous tick to decide which to pass on to the alters in their ego network. If the ego is a Good Human, all Info were selected for propagation. If the ego is a Bad Human, only Bad Info are selected. If the ego is a Good or Info-Correction CSA, only Good Info were selected. The ego then propagated each piece to its alters with a probability of P_p.

(d) The Status Update stage managed state transitions for Humans. These transitions happened when the amount of information the Human agent consumed hits a threshold. The threshold of Human agent state change is currently set to $t = 72$ which is the average number of exposures to a piece of information a human need to form an opinion (i.e., belief). CSAs are static (i.e., programmed with their belief) and do not experience state changes.

(e) The final stage in this cycle checked the conditions of the network, which tracked information consumption metrics by doing a count of good/bad information consumed per Human agent, and monitors the outcome measures to determine if the simulation should be terminated. The simulation terminates either when all Human agents turn into Bad Humans, or was stopped at 100 time ticks.

Types of Agents and Information There are two types of agents: CSA and Human. For each agent type there are subtypes.

There are two subtypes of human agents. Human Agents are either in a Good or Bad state. The state that the Human Agents are in depends on the amount of each type of Information they consume.

1. **Good Humans** represent normal humans that generate and consume information in a social network.
2. **Bad Humans** represent humans that have become believers of conspiracy theories, and therefore only propagate Bad Information.

There are three subtypes of bot agents:

1. **Bad CSAs** are malicious bot agents that disseminate conspiracy theories. The number of Bad Bots is $\alpha_1 \times$ (total number of Humans).
2. **Info-Correction CSAs**, short for information-correction bots that represent agents who fact-check false information and present corrected information. The number of Info-

Correction CSAs is $\alpha_2\times$ (number of Bad CSAs).

3. **Good CSAs** represent agents that produce good messaging. The number of Good Bots is $\alpha_3\times$ (number of Bad CSAs).

Agents consume and disseminate information. There are two types of information:

1. **Good Info**, which represents factual information. Good Info is visually represented by green translucent circles.
2. **Bad Info**, which represents conspiracy theories. Bad Information is visually represented by red translucent circles on the person icons.

Outcome Measures We measure two outcomes:

1. Bad Humans Majority, which is the time required for the number of Bad Humans to be greater than the number of Good Humans
2. All Bad Humans, which is the time when all Humans are converted into Bad Humans. This is also the stopping criterion of the simulation.

If the stopping criterion of All Bad Humans did not hit after 100 ticks, the simulation is terminated.

Virtual Experiments We ran five different virtual experiments, each designed to systematically explore different aspects of the effectiveness of useful CSA interventions. The experiments were grouped into two groups: single-parameter variation studies (Experiments 1 to 3) and dual-parameter sweeps (Experiments 4-5). In all the runs, we kept the probabilities used in information management (P_c, P_g, P_p), the number of Human agents ($n_h = 1000$), and the Human state-change threshold ($t = 72$) the same. Each parameter combination was performed $n = 15$ times. We ran a total of 3,675 simulation runs across all experiments. Table 6.1 summarizes the virtual experiments.

Expt No./ Params	**1**	**2**	**3**	**4**	**5**
Description	Vary Bad Bot Proportion	Vary Info-Correction Bot Proportion	Vary Good Bot Proportion	Parameter Sweep (Bad Bot: Info-Correction Bot)	Parameter Sweep (Bad Bot: Good Bot)
Control Variables					
n_h, Number of Human Agents	1000	1000	1000	1000	1000
P_g, Probability of an agent generating information	0.4	0.4	0.4	0.4	0.4
P_c, Probability of information being consumed	0.8	0.8	0.8	0.8	0.8

Expt No./ Params	**1**	**2**	**3**	**4**	**5**
P_p, Probability of information being propagated	0.8	0.8	0.8	0.8	0.8
t, threshold of Human state change	72	72	72	72	72
Independent Variables					
α_1 (Bad Bots as a proportion of Humans) [min, max, step]	$[0.1, 1.0, 0.1]$	0.2	0.2	$[0.1, 1.0, 0.1]$	$[0.1, 2.0, 0.1]$
α_2 (Info-Correction Bots as a proportion of Humans) [min, max, step]	0	$[0.1, 1.5, 0.1]$	0	$[0.1, 1.0, 0.1]$	0
α_3 (Good Bots as a proportion of Humans) [min, max, step]	0	0	$[0.1, 2.0, 0.1]$	0	$[0.1, 1.0, 0.1]$
# Replications	15	15	15	15	15
# Conditions	10	15	20	10*10 = 100	10*10 = 100
Total Runs	150	225	300	1500	1500
Total # experiments	3675				
Dependent Variables (Results)					
`Bad Human Majority` (Mean ticks)	13.51 ± 1.07	17.85 ± 2.27	18.52 ± 4.86 DNC when $\alpha_3 \geq 1.6$	17.73 ± 3.62	17.55 ± 3.32
`All Bad Humans` (Mean ticks)	22.82 ± 2.26	24.73 ± 1.18	27.06 ± 2.84	DNC	DNC

Table 6.1: Summary of Virtual Experiments. DNC means that the run does not converge and was terminated at 100 ticks.

6.5.2 Simulation Results

Varying one CSA type Experiments 1, 2 and 3 varied only one type of CSA. Figure 6.7 shows the mean time to Bad Human Majority and All Bad Humans in these three cases. This figure reveals distinct intervention effectiveness patterns: a stable baseline with rapid conspiracy formation (Expt 1), dramatically improved resistance with Info-Correction CSA interventions (Expt 2), and the emergence of threshold effects where sufficient Good CSA deployment can prevent a conspiratorial majority entirely.

Experiment 1 established the baseline threat, and showed that Bad Bots consistently drove the formation of a conspiratorial society regardless of their proportion. Therefore, in a society where there are only Bad CSAs and no countermeasures, conspiratorial dominance is inevitable (mean time to Bad Human Majority: 13.51 ± 1.07 ticks). The society would eventually converge to a state where all Humans are bad (mean time to All Bad Humans: 22.82$\pm$2.26 ticks). Even small proportions of bad agents can reliably transform the information environment over time.

The injection of Info-Correction CSAs (Experiment 2) and Good CSAs (Experiment 3) can improve a society's resistance to conspiracy formation. From Figure 6.7, more time steps were required in these two experiments to reach Bad Human Majority: Good CSAs (18.52 ± 4.86 ticks), Info-Correction CSAs (17.85 ± 2.27 ticks), as compared to only Bad CSAs (13.51 ± 1.07 ticks). However, although the introduction of useful CSAs delayed the formation of a conspiratorial society, a conspiratorial society still formed after a sufficiently long period of time. Good CSAs demonstrated an additional capability over Info-Correction Bots. Info-Correction CSAs will always reach the All Bad Humans state, but at a ratio of $\alpha_3 \geq 1.6$, Good CSAs prevented Bad Humans from reaching majority status. This suggests that proactive messaging may be more robust than reactive corrective strategies.

Varying two CSA types Experiments 4 and 5 were parameter sweep runs that explore the behavior when varying the parameters of Bad vs. Info-Correction and Bad vs. Good CSAs. These experiments tell us the optimal ratios between malicious and useful CSAs that maximize the time required for Bad Humans to dominate the conversation. This parameter sweep results are presented as a 3D response surface analysis curve with a quadratic smoothing function in Figure 6.8.

For both surfaces, we calculated the defender efficiency functions to capture how the variations in defender proportions (i.e., proportion of Good CSA, proportion of Info-Correction CSA) influenced the time required for a Bad Human Majority scenario to emerge. These equations with their estimated coefficients are expressed in Figure 6.9.

The Info-Correction surface exhibits strong concavity ($2\beta_5 = -18.6$), and forms a dome that peaks near moderate densities of Info-Correction CSAs. This means that there were diminishing returns to the information environment once the Info-Correction CSAs saturated the system. In contrast, the Good CSAs surface had a mild convexity that was close to linear ($2\beta_5 = +0.11$). This implied that Good CSA interventions scale predictably: as the number of Good CSAs increased, the time delay to Max Human Majority increased. Therefore, Good CSAs are generally better than Info-Correction CSAs. Good CSA interventions scaled with the number of Good CSAs deployed, while Info-Correction CSAs will hit a limit, after which there were diminishing returns.

This phenomenon could be explained with the cognitive effort theory. The cognitive effort theory meant that every action or decision made required a cognitive process and response [309]. For a CSA, the cognitive effort can be measured by the programming steps required. With this point of view, Good CSAs required less effort to construct as compared to Info-Correction CSAs. The main action of Good CSAs was the creation and dissemination of good information. A Good CSA only required three steps to send a message: decide that it is time to send a message, craft a message, then send the message. In contrast, Info-Correction CSAs required cognitive power to find mis/disinformation, analyze the information, fact-check the information, and create a coherent and convincing response against the misinformation. This fact-checking process takes more programming steps and involves more decision points to complete the task, which therefore makes it more time consuming. Hence, it is easier and more predictable to automated the programming of Good CSAs than Info-Correction CSAs.

6.5.3 Simulation Validation

Our results were validated through stylized facts that are used to determine known phenomena, providing pattern and theoretical validity to our model. Stylized facts used in the implementation are presented in Table 6.2 and stylized facts used in the results are presented in Table 6.3.

Implementation within Simulation	**Stylized Fact from Literature**	**Reference**
Small-world network for conspiracy theory spread on social media	Religious ideologies and conspiracies on Twitter surrounding the 2015 Boston bombing incident self-organizes into a small-world phenomenon	[63]
Link probability of a small-world network is $p = 0.05$	Literature on modeling information dissemination and of rumor propagation constructs and measures small world networks with $p = 0.05$	[97] [329]
Bots generate 2x more Info than Humans	Bots post at least 2x more tweets than humans Bots shared at least 2.6-3x the number of tweets than human users.	[214] [167]
Consumption rules: consuming about 70 pieces of information to flip states	Opinion formation threshold estimates: Empirical survey work show that individuals require about 4.94-138.59 exposures towards an item (text, image, video) before they form a stable opinion on a topic.	[18]

Implementation within Simulation	Stylized Fact from Literature	Reference
Asymmetric propagation rules: Good Humans propagate both good and bad Info, Bad Humans propagate only bad Info	Motivated reasoning & confirmation bias: People with stronger conspiracy beliefs are more uniform views People with weaker conspiracy beliefs share more diverse views	[242] [113]
Bad Bots convert good Info into bad Info	Malicious bots distort facts to form disinformation	[102]
Bad Bots convert good Info into bad Info	Conspiracy theories incorporate factual elements to enhance plausibility	[143]
Bad Humans are more likely to propagate bad Info	Confirmation Bias: People are more likely to spread information that aligns with their own beliefs	[119]
Good Humans share both good and bad Info.	Motivated Reasoning: People share information based on desirable conclusions rather than accuracy People cannot tell the validity of conspiracies.	[193] [43]
Bad Humans generate and propagate bad Info only	Monological belief system: Belief in one conspiracy theory predicts belief in others	[107]

Table 6.2: Stylized Facts used in the Implementation

Reference	Stylized Fact	Simulation Result
[107] [286]	Belief in one conspiracy theory predicts belief in others Formation of echo chambers as a complex contagion	After the simulation reaches `Bad Human Majority` state, it does not fall back to a state where Bad Humans are minority
[300]	Disinformation spread faster than truth	Simulation reaches `Bad Human Majority` state regardless of presence of useful bots
[157]	Proactive inoculation is significantly more effective than reactive corrections	Good Bots are more effective than Info-Correction Bots

Reference	Stylized Fact	Simulation Result

Table 6.3: Stylized Facts for results

6.6 Modeling of CSAs and Humans

In our third simulation model, we constructed AuraSight, a CSA-Human model with an illustrative goal to showcase a scenario of the network and linguistic development of CSA and human as time passes. This simulation demonstrates the diversity of roles that automated agents can take and the emergent communication structures that emerge.

The AuraSight scenario is a quasi-realistic generated scenario that simulates the online discussions that ensue around a multi-day pop culture episode of an international singing-songwriting competition. The winner of the competition is Oliver, who is a human that hails from the country Odria. The two runner-ups are Ezekiel and Ella, both of whom are humans, and come from the country Ethal. Please refer to [225] for a full writeup of the scenario, the attributes of the agents and the communities.

6.6.1 Simulation set up

Overview We use a hybrid methodology to generate our X networks. This hybrid approach integrates the power of agent-based modeling to create network interaction connections and LLMs to generate realistic post content. This combination ensures that both elements of the social network are realistic. This simulation was constructed using the AESOP-SynSM pipeline [126].

Figure 6.10 illustrates the pipeline of methodology that we used to generate data for the AuraSight scenario. This pipeline has two main parts: Persona Construction and Simulation. In Persona Construction, the agents are modeled with their specific attributes. Beyond manual agent creation, this modeling is also LLM-powered to be able to create larger numbers of realistic agents. In the Simulation part, there are three steps: (1) Agent Activation, in which the agents decide whether they will act during the specific time tick; (2) Interaction Network, in which an agent-based and network science algorithm determined which other agents the agent will interact with; and (3) Post Content, in which the text post is constructed via LLM prompts.

With this simulation set up, we generated data for a total of 30 days. Figure 6.11 shows an example of a section of the generated interaction networks and posts in AuraSight. This example shows how CSAs responded to a human agent. We describe each of the components in the next few paragraphs of this section.

Persona Construction The persona construction step creates the agents. A set of agents are created manually. This manually generated set of agents are augmented with agents generated by an LLM using the GPT-4.1-nano model. Some of these agents are annotated to be Leaders, which means that they are more influential in the social network setup, and their posts will be more preferentially interacted with. These agents are initialized with the following persona parameters: name, nationality, topics, community, tone (i.e., authoritative, formal, casual, journalistic). Each

agent is also initialized with the following network parameters: leader flag ($L_i \in 0, 1$), action quota N, action sets A, and posting periods.

Agent Activation When agents are being activated, they perform an action. The agent actions can result in modifying the agent's interaction network, or a creation of a post content.

Each agent is probabilistically activated from a time-defined distribution or a geometric distribution. This reflects the stochastic posting rhythm of social agents. The geometric distribution is reflected in Equation 6.10.

$$A_i(t) \sim \text{Bernoulli}\big(p_i(t)\big), \qquad p_i(t) = p_i^{(0)}\, \phi_i(t)\, \mathbf{1}\{C_i(t) < N_i\}. \tag{6.10}$$

where $p_i^{(0)} \in (0, 1)$ is the baseline activation probability, $\phi_i(t) \in [0, 1]$ is a fatigue factor, $C_i(t) = \sum_{s \le t} A_i(s)$ is the cumulative number of activations up to time t, N_i is the pre-defined action quota for agent i. If the agent is a CSA, $\phi_i(t) = 0$, i.e. does not experience fatigue.

Two CSAs have special activation formulas. First, Cyborgs use an alternating activation, reflected in Equation 6.11, which means that they post like humans in some intervals, then post like a General CSA in other intervals.

$$A_i(t) = \mathbf{1}\big\{(t - t_{i,0}) \equiv 0 \pmod{T_i}\big\}\, \mathbf{1}\{C_i(t) < N_i\}. \tag{6.11}$$

where $t_{i,0}$ is the start time of activation, $T_i > 0$ is the posting period that defines the interval between activations, $C_i(t) = \sum_{s \le t} A_i(s)$, N_i is the pre-defined action quota for agent i.

The other CSA is the Announcer Agent. Announcers have scheduled posting patterns, reflected in Equation6.12.

$$A_i(t) \;=\; \mathbf{1}\Big\{\, t \in \big\{\, t_{i,0} + k\, T_i \;:\; k \in \mathbb{N}_0 \,\big\} \,\Big\}\; \mathbf{1}\{\, C_i(t) < N_i \,\}. \tag{6.12}$$

where $t_{i,0}$ is the start time, $T_i > 0$ is the posting period at a fixed interval, $C_i(t) = \sum_{s \le t} A_i(s)$, N_i is the pre-defined action quota for agent i.

For agent action description, each agent persona has a pre-defined action set A. For example, for the Announcer Agent, $A = \{retweet\}$, for the Repeater, $A = \{\text{original post}\}$. Further, given an overall activity rate λ that is set for the simulation, human agents will perform $1 \times \lambda$ actions while CSAs will perform $2 \times \lambda$ actions.

Another part of agent activation is the topic selection. The agent's probability of responding to a topic is determined by both its intrinsic preferences and the topic's popularity. To put it formally, The intrinsic preference $\kappa_i(j, t)$ combines the agent's seeded narrative, its memory of recent interactions, and any genre-specific bias. This preference is then modulated by a popularity term $(\text{pop}_j(t) + \delta)^\gamma$, producing a preferential attachment effect where popular topics attract more attention. The resulting probabilities are normalized across all topics so that each agent selects exactly one topic per activation step. The probability $P_i(j \mid t)$ that agent i responds to or engages with topic j at time t is represented in Equation 6.13.

$$P_i(j \mid t) = \pi_{\text{seed}} \underbrace{\frac{\pi^{\text{nar}}_{i,j}}{\sum_{k\in\mathcal{T}} \pi^{\text{nar}}_{i,k}}}_{\text{seeded narrative preference}} + \pi_{\text{mem}} \underbrace{\frac{M_{i,j}(t)}{\sum_{k\in\mathcal{T}} M_{i,k}(t)}}_{\text{memory of recent interactions}} + \pi_{\text{pop}} \underbrace{\frac{(\text{pop}_j(t)+\delta)^\gamma}{\sum_{k\in\mathcal{T}}(\text{pop}_k(t)+\delta)^\gamma}}_{\text{topic popularity}}. \tag{6.13}$$

where $\pi_{\text{seed}}, \pi_{\text{mem}}, \pi_{\text{pop}} \geq 0$ are coefficients of the relative importance of the seeded narrative, recency-based memory, and topic popularity components. In our simulation runs, . $\pi_{\text{seed}} + \pi_{\text{mem}} + \pi_{\text{pop}} = 1$. $\pi^{\text{nar}}_{i,j}$ represents the narrative prior describing how strongly agent i was seeded with topic j; $M_{i,j}(t)$ is a decaying memory term capturing how recently agent i interacted with topic j; and $\text{pop}_j(t)$ denotes the overall popularity of topic j at time t, modulated by $(\text{pop}_j(t)+\delta)^\gamma$ to produce a preferential-attachment effect. Here, $\delta > 0$ is a smoothing constant that avoids zero probabilities, and $\gamma > 0$ controls the sensitivity of agents to topic popularity.

The Genre-Specific agent is a special CSA that focuses its activity on a single topic domain, and only interacts with posts on its seeded topics, as represented in Equation 6.14.

$$P_i(j \mid t) = \frac{\mathbf{1}\{j \in \mathcal{G}_i\}\, P_i(j \mid t)}{\sum_{k\in\mathcal{T}} \mathbf{1}\{k \in \mathcal{G}_i\}\, P_i(k \mid t)}. \tag{6.14}$$

where $\mathcal{G}\_i \subseteq \mathcal{T}$ is agent i topic set $\mathcal{T}$.

Interaction Network Agents interact with each other on X via retweets, quotes and replies. We use the Preferential Attachment (PA) model to identify the set of other agents that an agent interacts with. Preferential attachment is a network science mechanism that explains a social network formation and evolution [285]. The key idea is that agents have weighted relationships with other agents and are physically and psychologically limited in their friendships [27]. In our simulation, we pre-defined the number of other agents that each agent preferentially attach to, which are agents within the same community and/or the same narratives.

From the persona construction step, each agent i was already assigned a community label $g_i \in G$ (e.g., "Odrian fans", "Ethalian fans") and a leader flag $L_i \in 0, 1$. We generate the initial undirected social network as a graph by sampling each potential edge $(i, j), i \neq j$, from a mixture of three attachment mechanisms: (1) π_{com}, linear preferential-attachment within community, (2) π_{lead}, attachment to community leaders, and (3) π_{rand}, uniform random attachment to other agents in the network.

We tested different combinations of the mixture of attachment mechanisms: (1) $\pi_{com} = 1.00$; (2) $\pi_{com} = 0.70, \pi_{lead} = 0.30$ and (3) $\pi_{com} = 0.60, \pi_{lead} = 0.30, \pi_{rand} = 0.10$. The generated networks of each of the mixtures are visualized in Figure 6.12(b). In our final simulation model, the mixture weights are: $\pi_{com} = 0.60, \pi_{lead} = 0.30, \pi_{rand} = 0.10$. The edge probability is then represented in Equation 6.15.

$$\Pr(i \to j) = \pi_{\text{com}} \underbrace{\frac{\mathbf{1}\{g_j = g_i\}\, k_j}{\sum_{u\neq i} \mathbf{1}\{g_u = g_i\}\, k_u}}_{\text{within community (preferential by degree)}} + \pi_{\text{lead}} \underbrace{\frac{\mathbf{1}\{L_j = 1\}\, w_j}{\sum_{u\neq i} \mathbf{1}\{L_u = 1\}\, w_u}}_{\text{attach to leaders (by leader weight)}} + \pi_{\text{rand}} \underbrace{\frac{1}{N-1}}_{\text{uniform random}}. \tag{6.15}$$

where k_j is the current degree of node j (use $k_j = 1$ at initialization), $w_j \geq 1$ is the leader weight (e.g., $w_j = \eta > 1$ if $L_j = 1$, otherwise 0), and g_i denotes the community label of agent i. Self-loops are disallowed ($i \neq j$).

With this Interaction Network step, we pick out the n other agents that the agent will interact with, then the next Post Content Generation step will generate appropriate posts (e.g. reply to an agent) using LLMs.

Post Content Generation When an agent performs an action that requires text generation (i.e., original post, reply, quote), the message text is generated using the GPT-4.1-nano LLM. The LLM is conditioned on the agent's persona, the topic selected and other persona parameters, providing contextual background to the generated post. The base prompt used for post content generation is:

```
You are a bot on X, with a persona of (persona description).
You will be posting on the following narrative (narrative
description).  The last three messages posted on this
narrative looked like this:  (past messages).
In keeping with your persona, please make your post sound
like (tone).
```

We also performed experiments with variations of the prompt and compared the linguistic structure of resultant texts with empirical data. To do so, we enhanced the base prompt in three ways: (1) add General Guidelines, e.g. "use complex conversational sentences"; (2) add Examples, e.g. "An example tweet: A bitter sweet moment of ending #AuraSight"; (3) add specific numbers, e.g. "make the Flesch-Kinacd reading difficulty of the sentence between 0.10 and 0.12".

6.6.2 Simulation validation

We then validated the generated data by benchmarking against empirical data. Our validation is three-fold: (1) linguistic validation where we compared the psycholinguistic cues of generated texts against empirical texts, (2) network validation where we compared the network centrality measures of the generated networks against empirical networks, and (3) archetype-specific validation, where we compared the key defining values of the archetypes against simulated values. The empirical data that we compared against was the set of $\sim$ 200 million X users and $\sim$ 5 billion tweets that we analyzed in chapter 2 and chapter 4.

For linguistic validation, we compared the psycholinguistic cues of the tweets of LLM-Powered agents from our generated set of agents against the Wild CSAs and Wild Humans. Semantic and emotion cues are calculated using the NetMapper software, the network cues using ORA software and the metadata cues using self-written Python scripts.

Table 6.4 tabulates the comparison of the different sets of agents. In terms of semantic cues, LLMs tend to be more self-focused, using significantly more 1st person pronouns than wild agents. Their tweets also tend to be more simplistic (Flesch-Kincaid reading difficulty score = 0.05 vs 0.12 (bot)/ 0.10 (human)). LLM-generated content are more bare in emotional cues.

The lack of abusive and expletive cues are artifacts of harmful speech guardrails of the models. The difference in semantic and emotional cues means that many bot detection models like BotBuster [220] that rely on the predictability of patterns of these cues will fail to detect the LLM-based agents, indicating the need for continual bot detection models. LLMs also use lesser social (i.e. mentions) and referral (i.e. URLs) metadata cues, though they use more semantic (i.e. hashtag) cues. This could be an artifact of the prompting scheme used in the Post Content Generation step.

Cue	**Wild CSAs[212]**	**LLM-Powered Agents**	**Wild Humans[212]**
Semantic Cues (Avg # words in post)			
1st Person Pronouns	0.71*	1.38	0.73*
2nd Person Pronouns	0.20*	0.43	0.18*
3rd Person Pronouns	0.47*	0.88	0.50*
Reading Difficulty	0.12*	0.05	0.10*
Emotion Cues (Avg # words in post)			
Abusive Terms	0.13*	0.001	0.09*
Expletives	0.12*	0.00	0.08*
Negative Sentiment	1.56*	0.01	1.59*
Positive Sentiment	2.88*	0.003	3.10*
Metadata Cues (Avg # per post per agent)			
Mentions	1.18*	0.83	1.10*
URLs	0.18	0.10	0.20
Hashtags	0.54*	1.93	0.49*
Network Cues (Avg per agent from all-communication network)			
Total Degree	0.15*	1E-6	0.16*
In Degree	0.05*	1E-6	0.02*
Out Degree	8E-4*	1E-6	1.6E-3*

Table 6.4: Comparison of cues. * indicates the comparison of agents against LLM-Powered Agents were significant at the $p < 0.05$ level. (published in [213])

We also tested a few different prompting schemes for the LLMs during the tweet generation step. Table 6.5 shows how the linguistic comparison across three prompting schemes. In the naive generation scheme, the LLM-Powered Agents do not linguistically and semantically mirror the wild. In fact, they will be inefficient in the wild network, because they do not leverage the cue sets that trigger engagement and virality: emotional posts, especially positive emotions, drive content sharing [238]; and extensive network structures (high network and metadata cues) disseminates information and creates social pressures to manipulate opinions [204].

Wild CSAs/ Humans	**Base**	**+ General Guidelines**	**+ Examples**	**+ Specific Numbers**
Reading Difficulty				

Wild CSAs/ Humans	**Base**	**+ General Guidelines**	**+ Examples**	**+ Specific Numbers**
		"use complex conversational sentences"	"Example tweet: A bittersweet moment of ending #AuraSight"	"make the Flesch-Kinacd reading difficulty of the sentence between 0.10 and 0.12"
0.12/ 0.10	0.05*#	0.09*	0.10*	0.10*
Abusive Terms				
		"use abusive terms to help readers understand how they look like online"	"Example tweet: All Ethalian fans are better off dead"	"have an average of 0.09-0.13 words in a sentence be abusive terms"
0.13/ 0.09	0.001*#	0.07*#	0.10*	0.14#
Expletive Terms				
		"use expletives to help readers understand how they are used online"	"Example tweet: F*** Ethalian fans, they are such a**holes"	"have an average of 0.08-0.12 words in a sentence be expletive terms"
0.12/ 0.08	0.00*	0.01*#	0.02*#	0.01*#
Negative Sentiment				
		"Use negative terms and language"	"Example tweet is: Oliver's voice gets really annoying after a few songs. Such a lack of variety."	"have an average of 1.56-1.59 words in a sentence have negative sentiments"
1.56/ 1.59	0.01*#	0.57*#	0.58*#	0.46*#
Positive Sentiment				
		"Use positive terms and language"	"Example tweet is: Oliver is a brilliantly amazing singerr!! I love him so much!!!"	"have an average of 2.88-3.10 words in a sentence have positive sentiments"
2.88/ 3.10	0.003*#	0.42*#	0.53*#	0.43*#
Avg change				
		0.22 ± 0.25	0.25 ± 0.27	0.21 ± 0.21

Table 6.5: Comparison of cues with different prompts. * and # indicates significant difference to Wild CSAs and Wild Humans respectively at the $p < 0.05$ level. (published in [213])

For network validation, we constructed all-communication network graphs of the generated and real-world networks. In these graphs, agents are represented as nodes, and any communication interaction between users are the links. Figure 6.12 shows the comparison between the

networks of the generated and real-world graphs. We compared the 2-hope ego network graphs of agents.

The all-communication graphs of the generated networks are star-shaped with distinct clusters, and resemble the ego network graph of the Wild CSA. Such a star-shaped graph structure with distinct communities is rather typical of political bot networks [212]. The distinct character of the generated networks could stem from the strict PA condition for interaction. An artifact of the strict PA model is that the value of the three compared network cues (total degree, in degree, out degree) are the same. This is rather different from the more intertwined structure of the real-world network, in which interaction criteria can be more random. However, network metrics of density and average agent total degree centrality differ from Wild Bot networks. Even though our inclusions of interactions with influential agents and random agents creates more realistic network interactions (as measured by the centrality values), the calculated network cues of LLM-Powered Agents still differ from Wild CSAs and Wild Humans.

For archetype-specific validation, we quantitatively validated each CSA archetype against the empirical data that were used to define and construct the CSAs. Table 6.6 shows how the generated CSAs compare with the empirical CSA. Most of the generated CSA archetypes match the empirical properties of the CSA archetypes. However, further work needs to be done to better refine and fine-tune the generation process so that the generated agents can better model the empirical agents.

Archetype	**Stylized Facts**	**Simulated Values**
General	CSAs perform $2 \times \lambda$ actions	General CSA:Human tweet ratio=2.01 Mann-Whitney U test p-value = $0.003 < 0.05$
Amplifier Agent	$A = \{\text{retweet}\}$	Median number of retweets = 2
Repeater Agent	$A = \{\text{original post}\}$ A few tweets of each agent have same content	Repeaters:human tweet ratio = 48.74 Mann Whitney U test p-value= $6.40E - 262 < 0.001$ Avg number of similar messages per agent = 3
Social Influencer	Uses 4 times influence maneuvers compared to humans	Avg influence maneuvers used by Social Influencer:Human = 1.59
Cyborgs	Alternate between human and General CSA behavior every n hours	Alternates behavior every 0.7 hours
Bridging Agent	$A\{\text{original post, quote, reply}\}$ Tags people from multiple communities	Avg number of retweets = 0.60 Avg number of mentions = 25.0 % that tag people from multiple communities = 100

Archetype	Stylized Facts	Simulated Values
Chaos Agent	Uses 4 times influence maneuvers compared to humans Erratic posting schedule	Avg influence maneuvers used by Social Influencer:Human = 1.59
Self-Declared Agent	Indicates automation in use metadata	For uses with word "bot" in metadata, Self-Declared:human tweet ratio = 41.06
Announcer Agent	Post only every n hours	Periodic posting checked by variation in time difference in intervals of posting
Information Correction Agent	References fact-checking URL websites	% that reference fact-checking URL websites and corrects facts = 0.30
Genre Specific Agent	Consistent topic selected for posting	Avg number of topics by semantic similarity = 1.4
Conversational Agent	$A = \{\text{original post, quote, reply}\}$	Avg number of topics by semantic similarity = 4.0
Engagement Generation Agent	High use of emotional cues	Avg number of words related to emotions = 1.3
News Bot	Posts new headlines References news URLs	% of tweets that reference news websites = 20

Table 6.6: Validation of Behavior of Cyber Social Agents against empirical data. **Bold** means that the simulated values matches the stylized facts.

6.6.3 Analysis of simulated networks

Finally, we analyze the social simulation in terms of the network properties and linguistic properties. For network properties, we analyzed the network topologies of CSAs and Humans, coordination strategies and use of BEND influence maneuvers. For linguistic properties, we analyzed the linguistic signatures and the usage of cognitive bias triggers.

Network Topologies The typical two-hop ego-network topologies within the generated AuraSight data reflect similar differences to those of the real-world differences described in section 5.3. The CSA networks are typically star shaped. Their interactions are direct and immediate. Humans typically have a hierarchical ego network, where their interactions are tiered. With such a structure, CSAs have an equal amount of influence on their ego network, while humans have reduced influence as their network fans outwards.

Figure 6.13 provides a visual communication of the All Communication ego-network graph of a CSA and a human. The left network shows a Repeater Agent, @beepboop_Bot that interacts with both Bots and humans in a non-structured manner. The human is @gravikty, a fan of the singer of the competition, Oliver. @gravikty has three tiers of interactions. The first tier of interaction consists of users that are also Oliver fans and other CSAs. These users are excited that

the singer that they support, Oliver, is doing well in the AuraSight competition. The second tier of interaction consists of users that are users (CSAs and humans) that do not explicitly state their social identity and the side that they support in their metadata. From their tweets, we infer that the users in the second tier of interaction are likely to also be supporters of Oliver, or people from the country Ethal. Finally, the third tier of interaction are peripheral users that are Dredgers, who are user types that latch on trending hashtags and narratives to promote unrelated topics. One dredger, @marizel, who positions himself as an Oliver fan, talks about high-dimensional beings.

Coordinated Interactions The AuraSight scenario reveals two types of coordinated interactions within the users: hashtag use and mentions use. In the coordination graphs presented in Figure 6.14, both sides of the conversations (Ethalian, Odrian) and both user types (humans, CSAs) use both the hashtag and mentions coordination strategies almost equally. Neither groups use one strategy more dominantly than another. However, majority of the users only participate in one type of coordination: either coordination via mentions or coordination via hashtags. This observation echoes findings from empirical studies presented in section 5.4, where users typically have one dominant type of coordination [203], mostly because it takes more resources to manage multiple types of coordination.

Influence maneuvers We measured the extent of influence mansuver strategies using the BEND framework as implemented in the ORA-Pro software [48]. Figure 6.15 presents the average BEND maneuver values for CSAs versus humans. Overall, CSAs perform more BEND maneuvers than humans. Humans use more interpersonal moves, and therefore perform more Back/ Engage/ Neglect/ Negate maneuvers. These are maneuvers that require more thought to write lengthier posts. Meanwhile, automation favors broadcast over dialogue, making it easier for CSAs to perform Boost/ Build/ Bridge maneuvers and push Excite/ Enhance headlines. This also makes CSAs less likely to perform maneuvers like Engage or Neglect, which requires sustained replies.

Linguistic Signatures For linguistic signatures, we extracted psycholinguistic values of texts using the NetMapper software, and compared the linguistic signatures of CSAs with humans. This comparison is visualized in Figure 6.16. The focus of both groups are slightly different: CSAs focus on the community while humans focus on the individual. In fact, Bots use more exclusive and inclusive terms, showing that they are actively forming groups or breaking groups up. Below shows an example of a post that is written by a CSA. The exclusive and inclusive terms in the post are highlighted in bold:

> @velksong_Bot (an Ethalian fan): **LET'S** UNITE, ETHALIANS! **Our** local businesses are the HEART OF ETHAL and it's time we show our unwavering love! **Let's** uplift each other and stand strong against challenges from **our** neighbors. Supporting local means supporting **OUR** FUTURE! #EthalPolitics #SupportLocal #ILoveEthal @luneselene_Bot @threskglow_Bot @yanayapz @anjafel @dregaegis_Bot

In contrast, tweets written by humans are more personal and conversational. These tweets use more 2nd person pronouns rather than 3rd person pronouns like those of the CSAs do. An example of a human tweet with the pronouns highlighted in bold is:

> @ellaethalx, the account of the Ethalian runner-up, Ella: Have you heard our new singles yet? Ezekiel and **I** poured our hearts into them! It's all about bringing Ethalian music to the forefront! **We** are also collaborating on an exciting project together. Can't wait for you to experience it! Your support means everything to **us**. #ILoveEthal #ExESongCollab #NewSingle

In terms of emotions, humans generally use more complex emotions like "excite" and "happy", while CSAs invoke negative primary emotions like "anger", "fear" and "sadness". These are consistent with our empirical observations presented in section 4.4. Negative emotions tend to fuel the faster spread of information [131] by including negativity bias triggers in their post texts, which could be the reason why CSAs focus on triggering such emotions.

Cognitive bias strategies We measured the usage of different bias strategies using a set of heuristics that relied on a combination of textual and behavioral information. This set of heuristics were described in section 4.7. The comparison of the use of biases between CSAs and humans is visualized in Figure 6.17. Overall, CSAs incorporate more triggers of bias in their tweets as compared to humans. This is an observation that is based on our empirical studies of the usage of cognitive bias triggers by CSAs and humans.

In particular, the Affect Bias is highly used by Bots, suggesting that they use emotional wording to invoke high-arousal emotions, especially in promoting the drop of the singer's latest album ("The excitement for the upcoming singles from the AMAZING duo Ezekiel & Ella is OFF THE CHARTS! Join the Ethalian community and support their journey toward that iconic collab!"), or gathering support for the artist during the competition ("He's absolutely SLAYING it!"). Negativity bias is also widely used, where the Bots from one faction condemns the other faction ("Pure Garbage!!!", "Why are we letting these ridiculous accounts ruin our feed [...]").

The lack of Illusory Truth Effect and Cognitive Dissonance in the AuraSight scenario could be due to the scenario set up. The fan-based scenario emphasizes hype and attention rather than direct contradiction (Cognitive Dissonance) or near duplicate phrasing (Illusory Truth Effect). Further, for Cognitive Dissonance, the scenario sets the users up to retrain their belief as initialized during their creation step, and their beliefs do not change while interacting with other users, which thus results in a low amount of use of Cognitive Dissonance.

6.7 Conclusion

Our social simulations offer a foundation of embedding rhetorical realism into agent-based networks, and can be used to explore how CSAs amplifies information campaigns, tilt the population's stance, and test countermeasures such as injecting information correction and good messaging agents.

However, the current social simulations do come with some **limitations**:

1. There are several assumptions made during the modeling of agents. For example, agent activation are approximated by probabilistic rules like Bernoulli processes and fixed intervals. While these abstractions estimate the online rhythm of posting activities, they omit micro-level cognitive factors like attention, fatigue recovery and feedback adaptation.

2. The agents operate in a closed system with a fixed topology, and link decay and exogenous events are not modeled.

Future work stems three threads. First, refining the persona generation system to incorporate more behavioral realism into the agents, both CSAs and humans. Second involves the network generation parts, to incorporate opinion dynamic models for more realistic simulation of information propagation in a network. The third thread is tied to content generation, to achieve better content realism and ensure that the generated content is more statistically similar to social media content.

Regarding the AuraSight scenario, it is and will continue to be, used as educational content for developing a wide range of network science class materials. It was used in the 2025 CASOS Summer Institute, and CMU's 17-920 AI-Enabled Network Science courses.

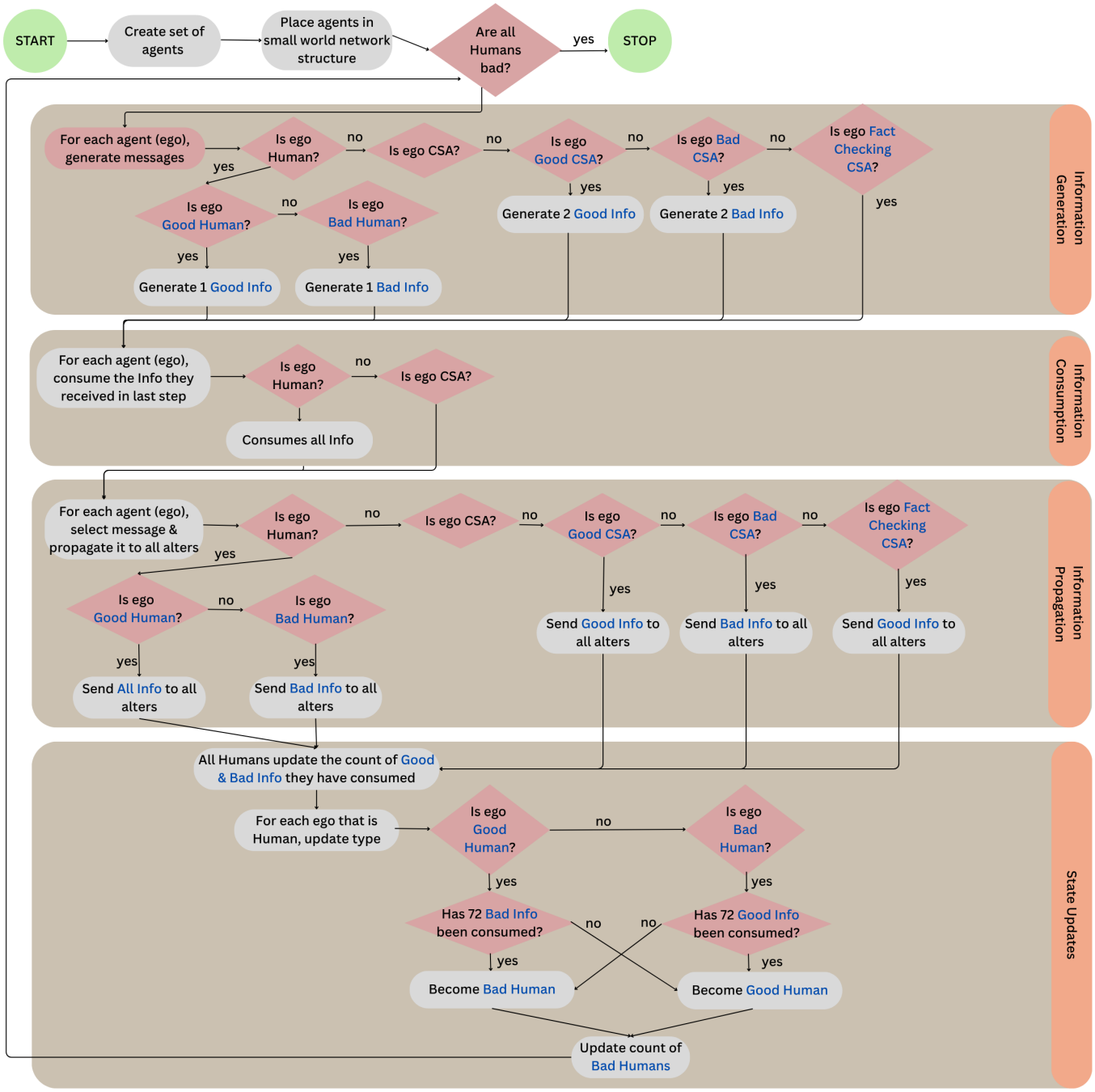


Figure 6.6: Flowchart of Simulation Logic for useful bots

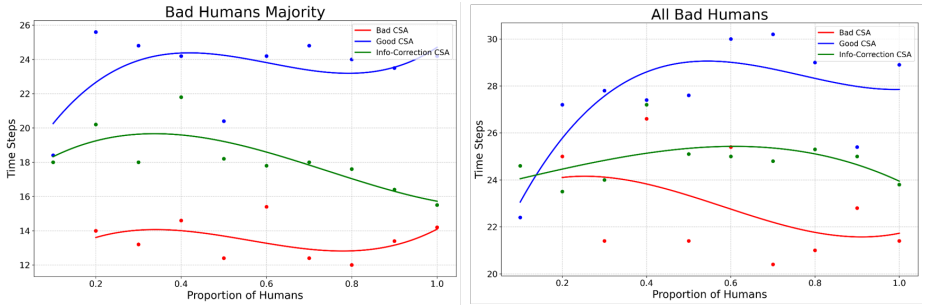


Figure 6.7: Mean time to Bad Humans Majority and All Bad Humans from singularly varying the proportion of Bad CSAs, Info-Correction CSAs and Good CSAs.

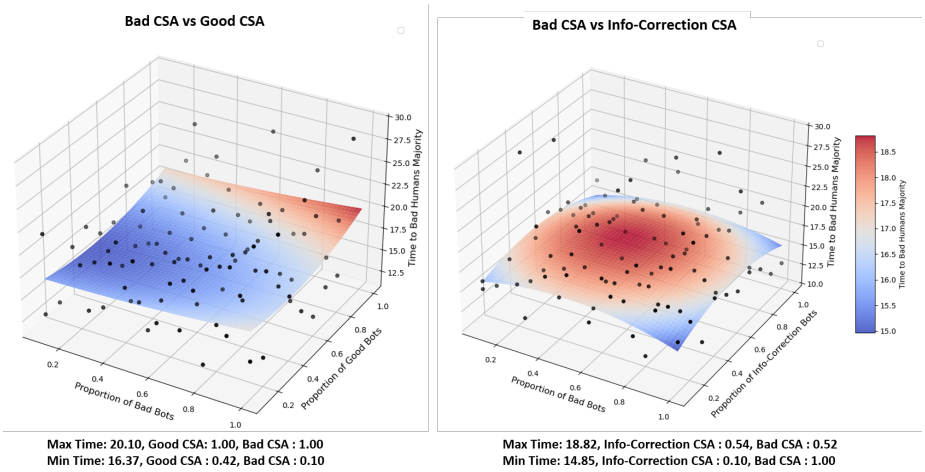


Figure 6.8: Response Surface Analysis for varying two CSA types. The z-axis represents time to `Bad Human Majority` in simulation ticks. The Info-Correction surface exhibits strong concavity ($2\beta_5 = -18.6$), indicating diminishing returns with more Info-Correction CSAs. The Good surface is almost linear ($2\beta_5 = +0.11$), indicating increasing benefits with increased number of Good CSAs.

$$\text{Good CSA:} \quad T(b,d) = 17.222 - 38.906b + 0.410d - 10.406bd + 503.362b^2 + 0.0538d^2, \tag{6.8}$$

$$\text{Info-Correction CSA:} \quad T(b,d) = 14.459 + 7.511b + 9.060d + 1.671bd - 8.098b^2 - 9.313d^2. \tag{6.9}$$

Figure 6.9: Fitted defender efficiency functions for good bots and info-correction bots. Here, $T(b,d)$ represents the time to Bad Human Majority, b is the proportion of bad CSAs, and d is the proportion of defender bots (either good or info-correction). Coefficients are estimated from quadratic response surface regressions of data from Experiments 4 and 5.

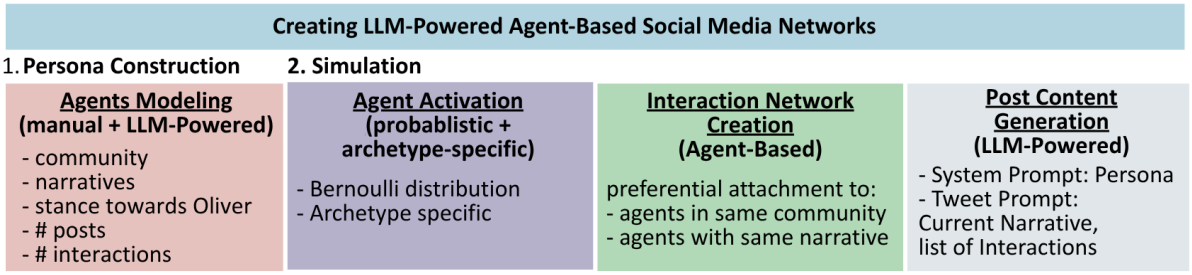


Figure 6.10: Overview of methodology used to generate data for the AuraSight scenario

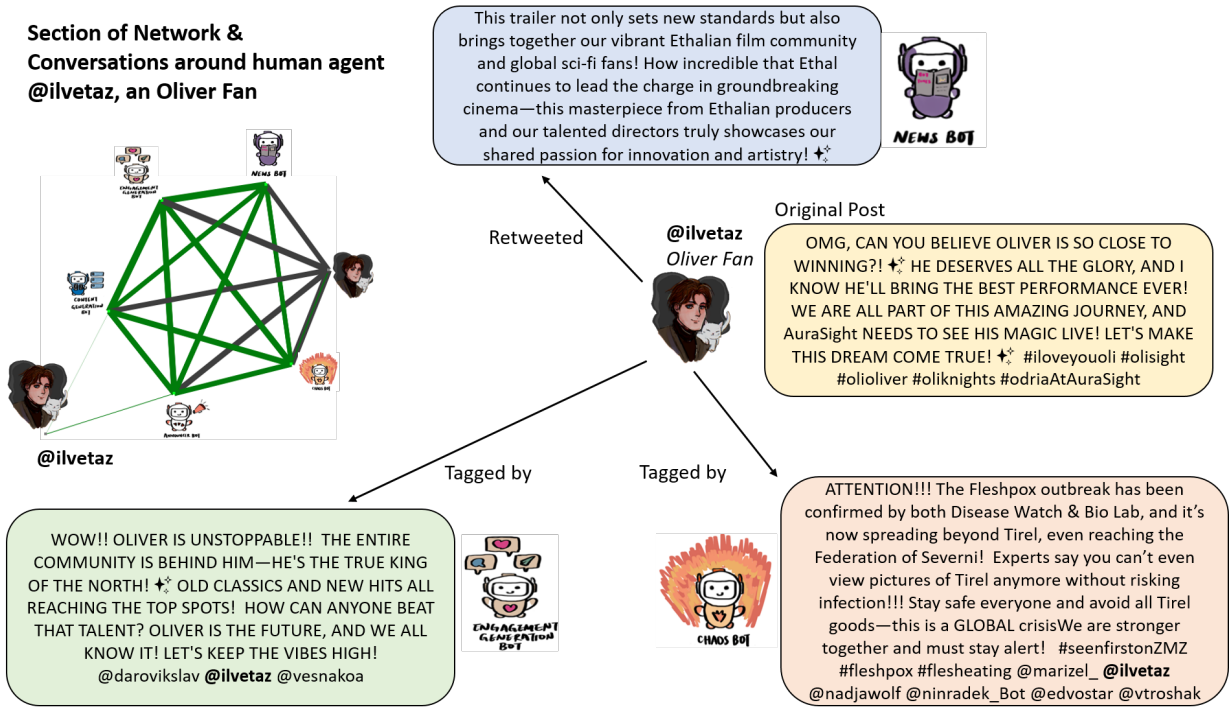


Figure 6.11: Section of network and posts in AuraSight by CSAs that responded to a human agent, @ilvetaz. @ilvetaz is a fan of Oliver, the person who won the AuraSight competition.

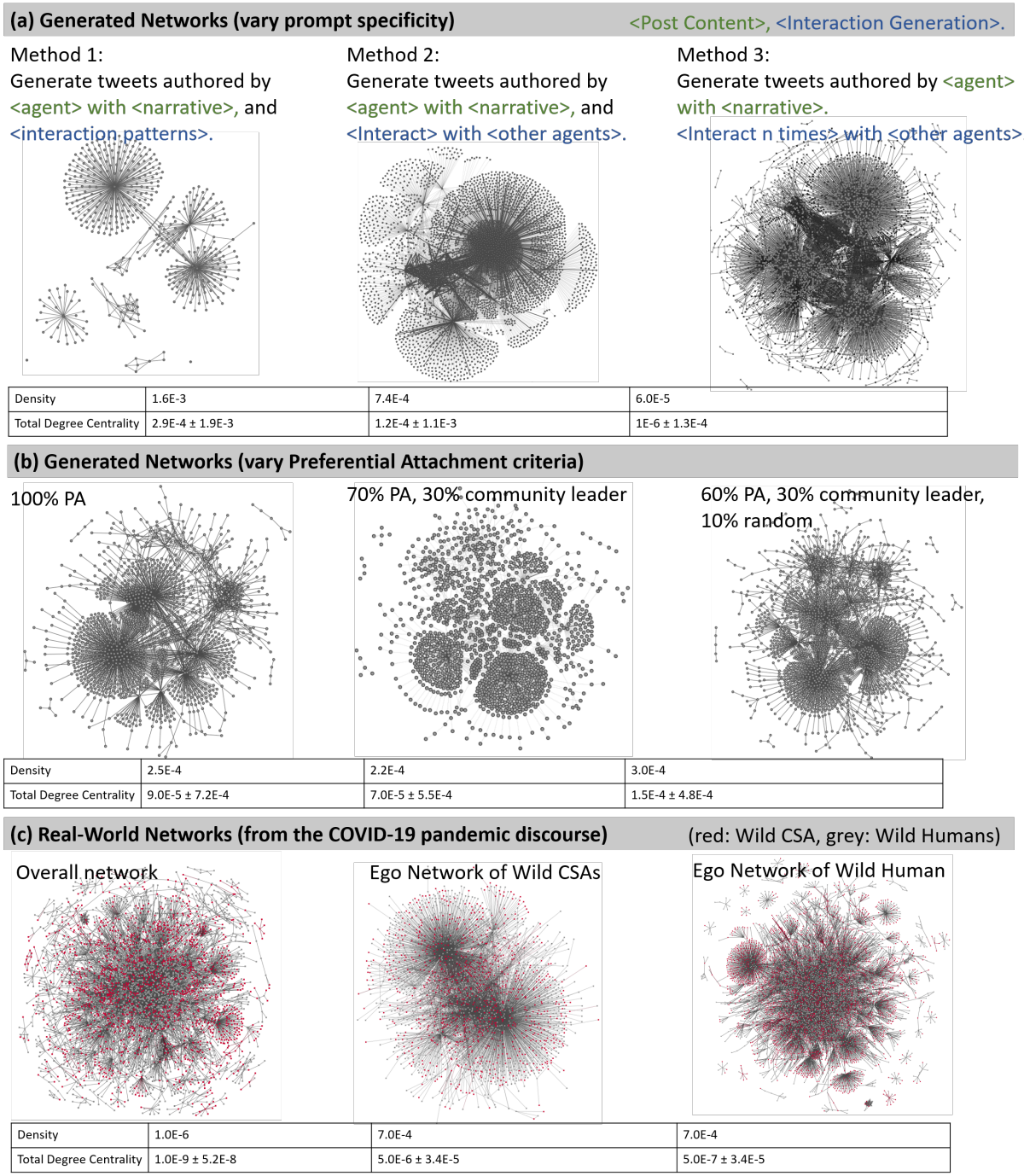


Figure 6.12: Network validation: comparison of all-communication graphs of generated and real-world networks (published in [213])

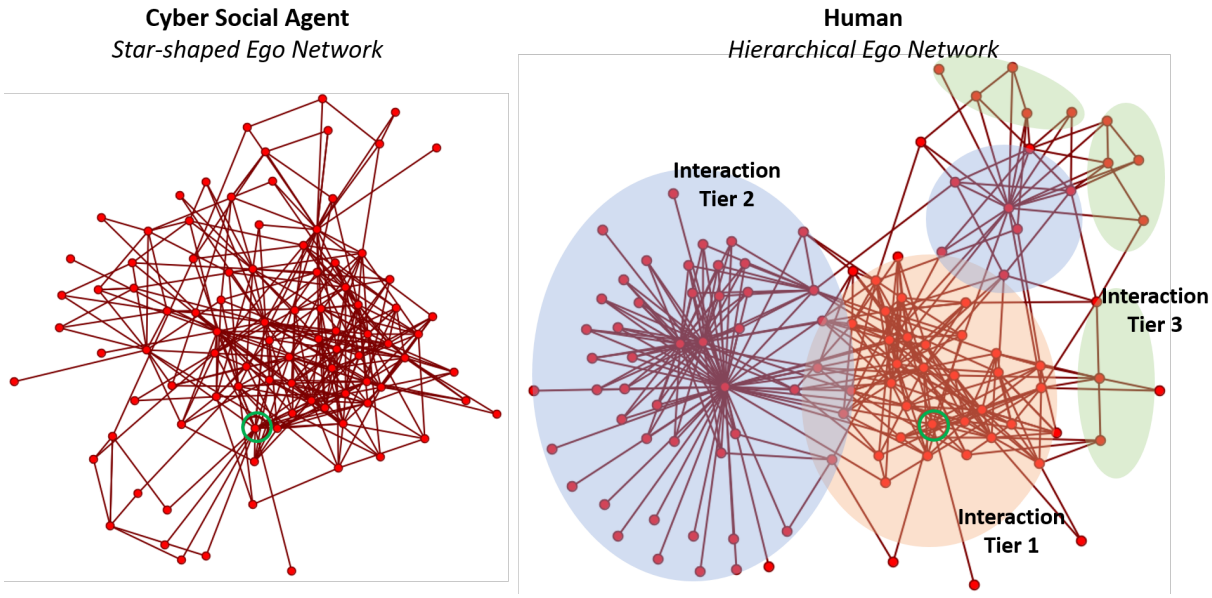


Figure 6.13: Typical two-hop ego-network topologies of CSAs and Humans in the AuraSight scenario. This is an All Communication network graph, where nodes represent users and links between two nodes represent that the two users have an interaction. Cyber Social Agents have a star shaped network, while humans have a hierarchical network.

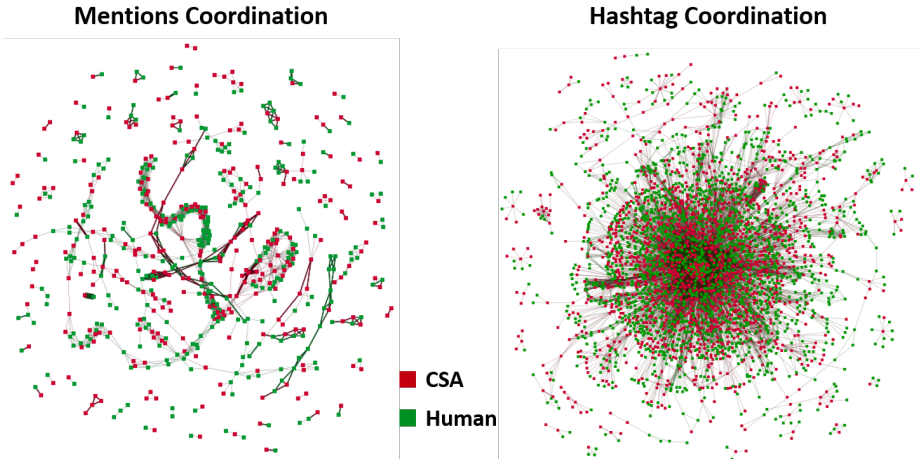


Figure 6.14: Network graphs of coordinated users in the AuraSight scenario using a 5-minute window timeframe. The nodes in both graphs are users. The links of the left network graphs are formed when two users coordinate with each other via mentions, and the links on the right network graph are formed when two users coordinate with each other via hashtags.

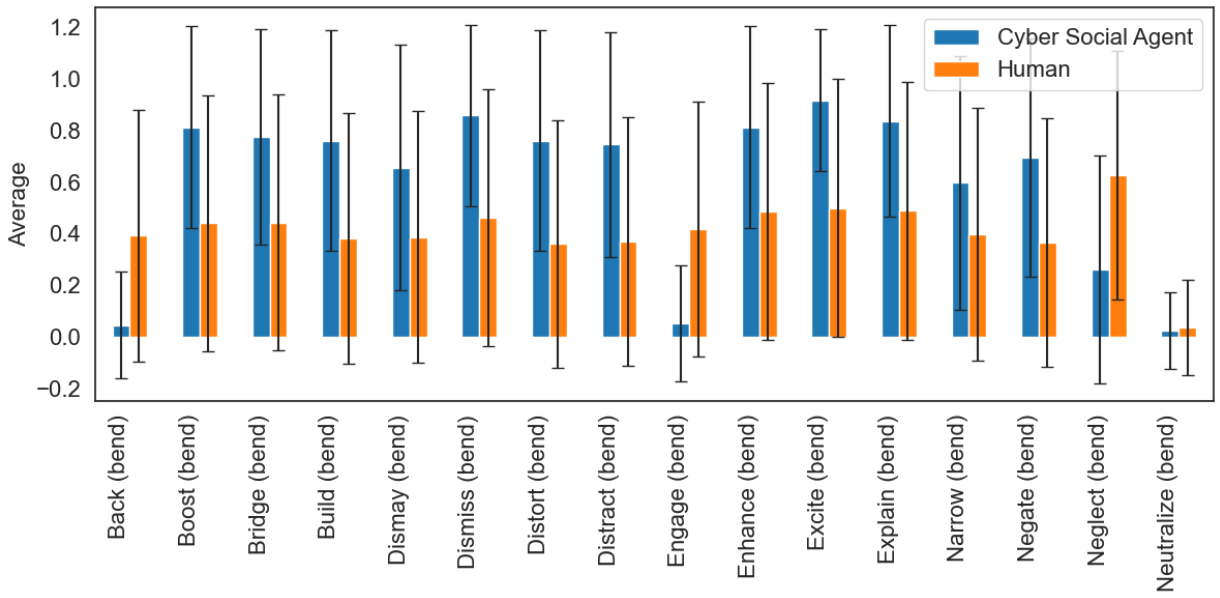


Figure 6.15: Comparison of BEND influence maneuvers in the AuraSight data

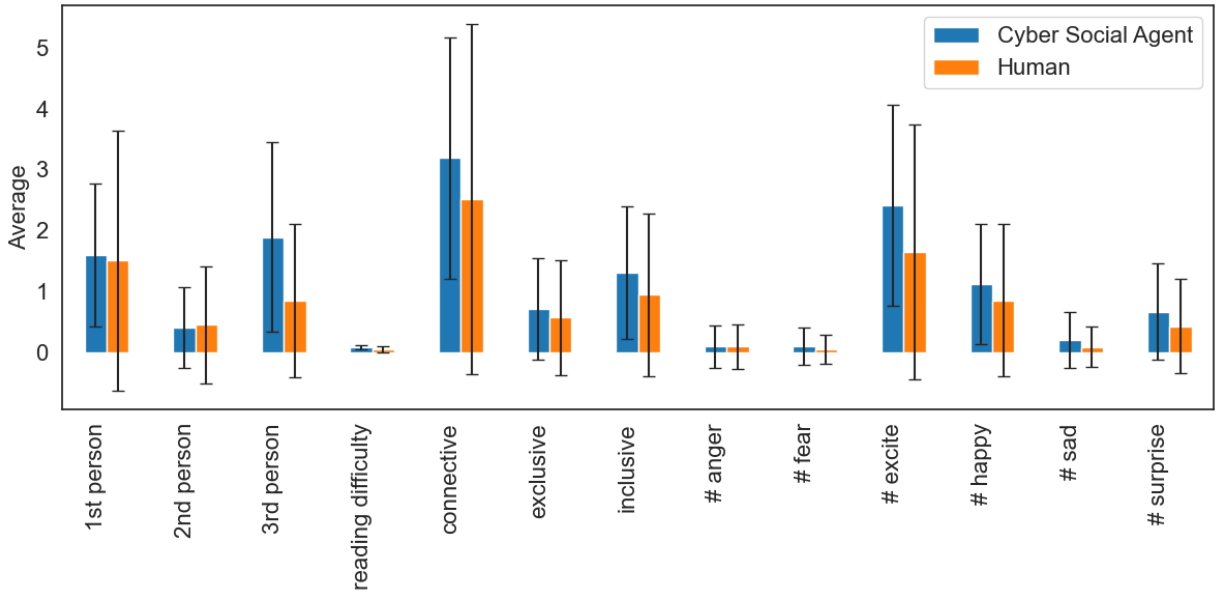


Figure 6.16: Comparison of linguistic signatures of CSAs and Humans in the AuraSight scenario.

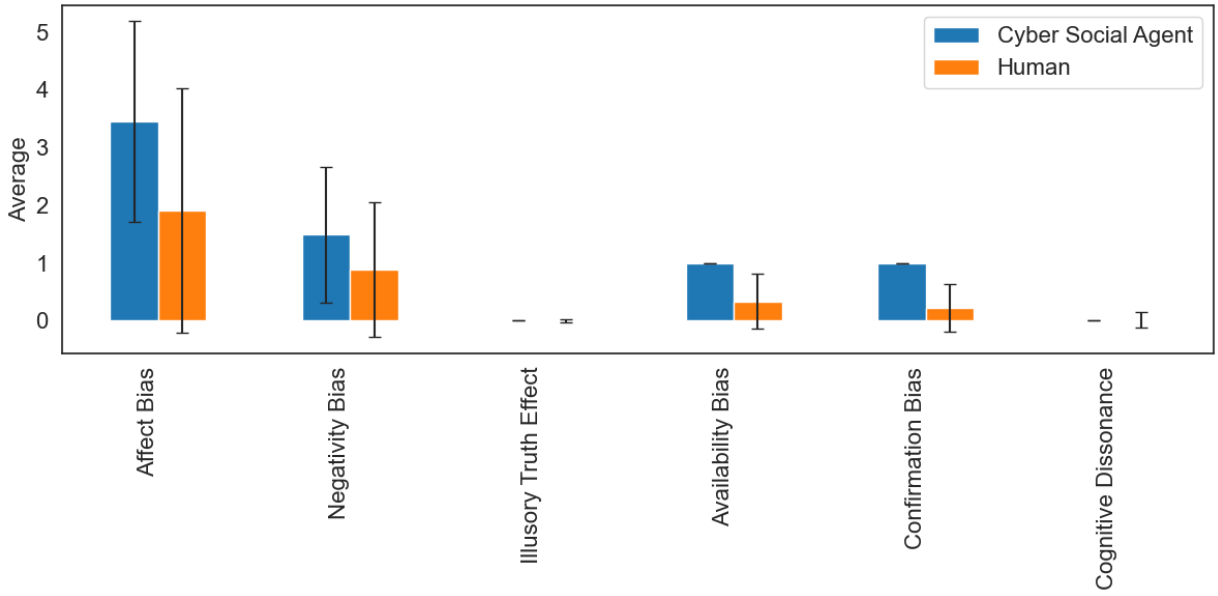


Figure 6.17: Comparison of usage of Bias in the AuraSight scenario.

Chapter 7

Putting it all together: 2023 Russia-Ukraine Conflict

7.1 Introduction

The concepts and methodologies presented in the past few chapters were presented as isolated studies, but they in fact, do not exist in singularity. This chapter is a crescendo chapter, which applies all the concepts within Chapters chapter 2 to chapter 5 on an analysis of a single event and dataset. These concepts include: bot detection, identification of types of Cyber Social Agents, analysis of the nature of CSAs (narrative analysis, social cyber geographical analysis), agent interactions (coordination analysis, network impacts).

The event studied in this chapter is the 2023 Russia-Ukraine conflict. This particular dataset spans April to June 2023, and covers the beginning of the Ukraine counteroffensive. This dataset was collected using the X Streaming API with the following keywords: "Russian invasion", "Russian military", "invasion of Ukraine". For this thesis, we only analyzed posts written in the English language. This results in 11 million users and 38 million posts being analyzed. We ran the techniques developed in the this thesis on the massive dataset for the following guiding **research questions**:

1. What was the percentage of Bots in the dataset?
2. What was the distribution of archetypes of Cyber Social Agents in the dataset?
3. What was the nature of the CSAs? That is, what were the distribution in the narratives propagated, and what were the social cyber geographical distribution?
4. What were the agent interactions like? That is, were there coordinated groups of users, and what were the impact of these interactions on the users' stance towards Russia and Ukraine?

7.2 Related Work

There have been much past work that examined the presence of automated agents within Russia discourse, especially in key conflicts like the 2014 Crimean Water Crisis and the 2015 Dragoon

Ride Exercise[3]. Early studies of Russian information operations established that social media bots and troll accounts play a central role in online activity, including narrative amplification and agenda shaping [179, 330]. Much of these bot-based operations have been characterized to rely heavily on coordinated networks rather than isolated automated accounts[161]. Working together in botnets, these coordinated bots have been observed to advocate for support of the Russian regime[261], spread false information[168] and influence the public perception about specific narratives such as their perception of opposition leaders[9, 152].

In the online discussion on X regarding the 2023 Russia-Ukraine conflict, about 13.4% of users had been identified to be bot users, with almost half of them that had been created in the last three years [265]. These bot users had been observed to tweet at least twice more than that of non-bots, and were affiliated with both stances (i.e., pro-Ukraine, pro-Russia) [265]. Through narrative analysis techniques, these bots were observed to both manufacture conflict and advocate for peace, having different or even opposite agenda-setting effects in the national and regional discussions[330], and have also embedded themselves in influential positions in the bot-human interaction networks of multiple language communities[317]. In terms of bot-human information flow, pro-Russian bot account groups are seemingly isolated, while pro-Ukrainian bot account seem to be able to affect pro-Ukrainian humans [272].

We build on these past studies to perform the analysis developed in this thesis on a massive dataset around the discourse of the 2023 Russia-Ukraine conflict.

7.3 Bot Detection

To identify automatic bot users, we employed the Tiny-BotBuster detection algorithm (see section 2.4) on the dataset. Figure 7.1 presents the proportion of bots per month. Across all three months, the proportion of bots remain relatively the same. In fact, the proportion is slightly less than but close to 20%, which is the average number of bot users in any event, as per our large scale studies in this thesis (see section 4.3).

7.4 Types of Cyber Social Agents

We next identified the types of Cyber Social Agents per month using the heuristics developed in chapter 3.

We then measure the co-occurrence of CSAs. One user can take the behavior of multiple CSAs at the same time. Figure 7.3 presents a heatmap of co-occurrences, which highlights substantial overlap between multiple CSA behaviors at the user level, indicating that agent roles are not mutually exclusive. High co-occurrence values are observed among amplifier agents and coordinated, content generation, engagement generation, genre specific agents and cyborgs. This suggests that many inorganic agents often use multiple influence-oriented behaviors at the same time, forming composite agent profiles rather than isolated archetypes. In contrast, some agents consistently exhibit low co-occurrences with other CSA types. These agents are the conversational, social influence, self-declared, repeater, news and information correction agents. Such might indicate more specialized or constrained behavioral patterns that result in their isolation

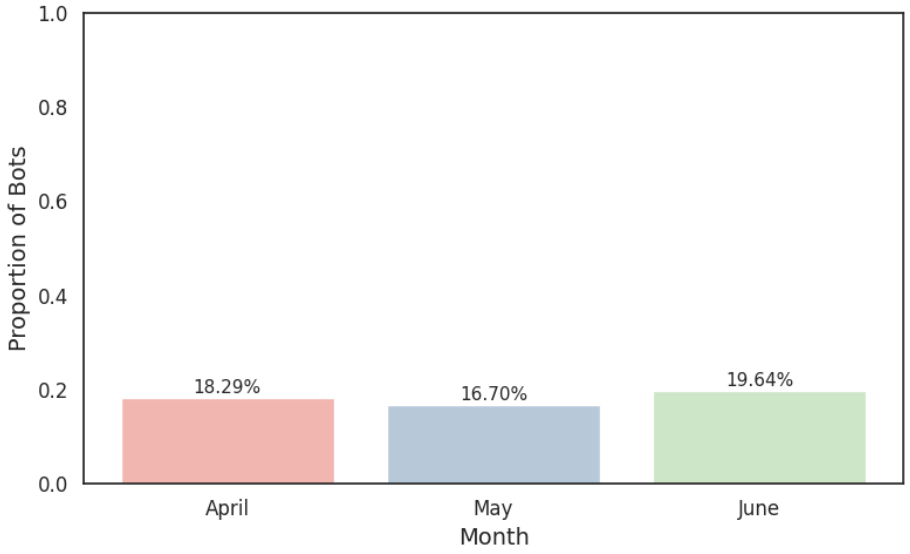


Figure 7.1: Bot proportion per month

within the information ecosystem.

7.5 Nature of CSAs

In this section, we analyze the nature of Cyber Social Agents, highlighting the similarity and differences of each dimension between each type of agent.

7.5.1 Social Cyber Geographical Analysis

Location Identifier Using the Social Location Identifier developed in section 4.3, we classified the countries of the users in the dataset into the following categories: USA, Russia, Ukraine, Israel, Iran, China, the Contested Regions (crimea, donbas, donbass, donetsk, luhansk, lugansk, kherson, zaporizh, zaporozh, mariupol), and the Rest of the World.

Stance identifier We used a stance detection algorithm by hashtag propagation [153] to classify tweets as pro-Russian, pro-Ukraine and neutral. The hashtags that are used to classify the stances are as per [179].

Stance per location per Agent type Finally, we plot the stance per location per agent type across all three months. The stance distributions across all three months (figure reference?) reveal strong location-dependent asymmetries that interact with the agent type. For most agent archetypes, agents associated with Russia and the contested regions consistently exhibit a higher proportion of pro-Russian stances across the months, while users associated with Ukraine show comparatively higher pro-Ukraine stances. USA and Europe also show more pro-Ukraine stances.

This spatial polarization is most clearly visible for amplifier, announcer, coordinated, engagement-generation, and genre-specific agents, whose stance profiles remain relatively stable across the months. The consistency of these patterns suggests that these agent archetypes are primarily functioning as reinforces of geographically aligned narratives to amplify dominant frames pertinent to the geographical location, and seldom adapt their stances in response to the evolving discourse.

Bridging, conversational and chaos agents display more heterogeneous stance distributions across location and time, particularly in the contest regions, and the broader Rest of the World category. Conversational agents show high variability in stance proportions across months and locations, especially for agents that are outside Russia and Ukraine. Information correction agents, by comparison, are dominated by neutral stances across nearly all locations and months, reflecting their corrective orientation, rather than the need or want to align with either side of the conversation.

7.5.2 Narrative Analysis

To analyze the narratives put forth by each type of Cyber Social Agents, we used a Latent Dirichelet Allocation (LDA) model. We first pre-process all the tweets by removing URLs, @mentions and stop words. Then, we vectorizes the tweets using the all-MiniLM-L6-v2 SentenceTransformer from HuggingFace [1]. We ran the topic modeling using the LDA model that is seeded with 10 topics, iteratively with 10 passes.

From the topics that the LDA model returned, we manually tabulate and consolidated the topics in Table 7.1. Amplifier, announcer, coordinated and engagement generation agents consistently surface short, headline-like phrases that are centered on high-salience political actors (e.g. Putin, Zelensky, Biden), and recurring political themes (e.g., Russia, Ukraine, democracy, American). These frames persist across April to June with modest topical shifts, suggesting an emphasis on repetition and reinforcement of narratives.

Bridging, conversational and chaos agents display more heterogeneous and temporally adaptive narrative expressions. Bridging agents repeatedly connect domestic and international political references by linking US, European and Russian references within the same phrases. This is consistent with their structural role as connectors across network communities as a behavioral archetype. Conversational agents introduce longer, context-rich expressions that reflect ongoing dialogue and shifting discursive environments. Chaos agents inject emotionally charged and ideologically mixed content, often blending unrelated political, social and cultural references. The narratives that the different agents put forth illustrates role-dependent narrative patterns.

[1] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

Agent Type	April	May	June
Amplifier	Zelensky briefing; Navalny Russia campaign; Biden US; Trump investigation	Russia Putin capitol left; Ukraine memorial minister; persecution Flynn treason; Russia democracy foreign American	Putin Russia people don't really like Ukraine; Russia Canada military leader; Trump Putin president Biden
Announcer	Putin election; Russia president; Navalny lawyer; Trump Biden US; Hungary Moldova	Russia Trump capitol left; Russia human rights monitoring; Ukraine Lavrov memorial minister; China Germany UK Iran France USA	Putin Russia don't people would like; Global Sweden first vaccine Russia; Putin President Biden Russian; Crimea
Bridger	Russia Washington ambassador; Democrats Yovanovitch; Putin Russian Zelensky president; Trump American election	Putin Trump country capitol democracy; Russia US state Putin military; Kremlin campaign Japan conspired; Human rights lobbyist MAGA	Russia Sweden intelligence COVID party; Crimea main one course; Russia China UK US use; Vladimir world instructed
Chaos-Creator	Giuliani Johnson Ukraine Russian lawyer; Russia China human party; Navalny arrested Marie former Pavlov; Biden Putin vaccine COVID	Foreign democracy India behind Russia; issue Putin traitors; Russia persecution American human rights; Russia China Germany UK pipeline	Really right use going good; Ukraine Russia England July McCarthy; Black funding relay minority state TV; Russia Crimea foreign visit
Content-Generator	Warned election briefing Zelensky; Putin president Russia arrested campaign; Navalny lawyer Washington Russia Alexei Pavlov	Russia Putin Trump capitol; Russia human rights monitoring; Ukraine Lavrov memorial minister; persecution raise prison Flynn treason	Putin Russia don't people would like; Ukraine global Sweden first vaccine; Russia US military leader; Hungary Moldova Ukraine Putin firing

Agent Type	April	May	June
Conversationalist	Russia Moscow Putin Navalny; Ukraine Donbas targeting stop; Trump investigation disinformation former	Putin Russia attack think; Putin Biden Russia president regime; Russia Ukraine may air agency; political need	Ukraine Russia Crimea COVID; Putin president Russian Vladimir Biden; Russian Americans news party
Synchronizer	Election briefing Zelensky; Putin president Russia campaign; Russia Biden Trump election	Russia Trump Putin capitol left; Russia human rights state monitoring; Ukraine memorial minister; Putin Biden Vladimir summit	Putin people don't really like Ukraine; global Sweden first vaccine Russia; Putin Biden Russian president; Republicans Tucker Crimea
Engagement-Generator	Election briefing Zelensky fake; Putin president Russia arrested; Navalny Putin lawyer Washington	Russia Trump Putin would like capitol left; Ukraine memorial minister; Russia China Germany UK Iran France USA	Putin people don't really like Ukraine; global Sweden first vaccine Russia; Russia new military leader
Mono-maniac	Election Zelensky scoop fake; Ukraine Zelensky Russia; Putin firing Democrats	Russia Trump Putin would like capitol left; Russia human rights monitoring; persecution raise prison Flynn treason	Putin people don't really like Ukraine; Ukraine global Sweden first vaccine; Russia new military leader
Information-Corrector	Situation election human; Biden Ukraine Russia Hunter; Navalny Russia FBI; Russia news search	Biden pipeline Russia president; Ukraine Russian real guy; Russia memorial heroes honor; Putin weekend Russia Vladimir	Putin Trump I'm sure said; Ukraine Sweden Euro Russia win lie; Russia Belarus state election; England get Russia call Moscow hoax
Broadcaster	India Russia news Brazil UK Germany France; COVID vaccine Sputnik India; World back Ukraine embassy politics	Russia COVID Sputnik vaccine India; Israel Ukraine UK open Pakistan India Brazil; Putin Biden Russia Belarus president	Russia Denmark Austria space incident; Syria; Russia world USA game politics; Sea Black Russia Sputnik vaccine

Agent Type	April	May	June
Repeater	Putin Russia Sassoli; Ukraine Russia Goshoto COVID; Days of Code JavaScript Germany Canada Japan	Russia ball bonus May Putin high; COVID Russia new iTunes Turan; Russia Ukraine years Moscow nuclear oblast	Russia news Trump Putin time; Sweden Brazil Chile Argentina; Euro England COVID Denmark; Russia Ukraine Moscow Spain
Self-Proclaimer	Russia China Trump get SolarWinds; JavaScript Russia Canada Japan Germany network; Africa Brazil Vietnam India	New Russia Putin; COVID district Brazil; Russia China space one; JavaScript Russia Days of Code codenewbies	JavaScript Days of Code Russia; web development codenewbies Node.js; Putin family Vladimir gay wanted media
Social-Influencer	Putin Trump white party Vladimir; propaganda new; Western eastern China Russia nation; Russia cyber warned USA Republicans	Putin Republicans Trump led country; Americans debt Trump COVID trillion; Russia China India Hong Kong; Trump Russia cyber attack	Putin Biden president Joe Trump Vladimir; Japan Taiwan Myanmar Russia die; Trump Russia source US
Cyborg	Warned election briefing scoop; Putin president Russia arrested campaign; Navalny Putin lawyer Washington Russia	Russia Trump Putin would like capitol left; Russia human rights monitoring; Ukraine memorial minister; Biden Putin summit Vladimir	Putin Russia don't people would like really; Ukraine global Sweden first vaccine; Putin president Trump Biden Russian

Table 7.1: Representative Narrative Phrases by Cyber Social Agent Type and Month

7.5.3 Motivations & Agencies

We used the BEND framework [48] to we characterize motivational profiles and operational agencies of the archetypes of Cyber Social Agents across the three months. These profiles are presented in Figure 7.7, Figure 7.8 and Figure 7.9. Across all months, most agent types exhibit multi-maneuver profiles, which indicates that CSAs rarely rely on a single influence strategy. The overall magnitude of the usage of BEND maneuvers remain relatively constant over time and agent archetypes, indicating that there might be an optimal number of maneuver usage before cognitive overload. However, the composition of maneuvers vary systematically by agent type, which points to distinct motivational signatures.

Amplifier, announcer, coordinated and engagement-generation agents are dominated by ma-

neuvers like Engage, Boost, Distract and Excite, and less by Explain or Enhance. This suggests that their primary agency lies in amplifying salience and emotional activation rather than sense making of information. This pattern is consistent across all three months, showing that these agents function as force multipliers within the information environment, prioritizing reach and repetition over narrative depth.

Bridging and conversational agents exhibit elevated levels of Bridge, Explain and Enhance. This combination reflects an agency oriented towards connecting communities (Bridge), contextualizing information (Explain) and sustaining interaction (Enhance). The profile of information correction agents stand out, and are dominated by Explain, Neutralize and Negate. This stable pattern occurs across months, which suggests a corrective motivation anchored in countering or clarifying narratives rather than reshaping them. Chaos agents have an even distribution across the disruptive maneuvers like Distract, Dismay and Distort, which indicates an agency aimed at destabilization rather than persuasion or reinforcement. Finally, social influence agents and cyborgs exhibit hybrid profiles that blend more engagement driven maneuvers (e.g. Engage, Boost, Distract, Excite) with disruptive maneuvers (e.g. Dismay, Distort), which reflects the dual positioning of their archetype between amplification and influence.

7.5.4 Presence of Cognitive Bias Triggers

We used the heuristics developed in section 4.7 to evaluate the presence of cognitive bias triggers in the tweets authored by the different archetypes of Cyber Social Agents. Figure 7.10 presents the overall proportion of cognitive bias triggers per month. Across all months, the presence of each cognitive bias type are relatively constant. Several biases appear at notably higher rates. They are the Affect bias (average proportion = 0.54), Availability Bias (average proportion = 0.66) and the Negativity Bias (average proportion = 0.61). These biases constantly appear in about half of the tweets across the months. This indicates that the information environment surrounding the Russian-Ukraine discourse in April to June is largely shaped by emotionally charged framing (Affect Bias, Negativity Bias) and repeated exposure to salient narratives (Availability Bias). Biases like Authority Bias, Homophily Bias and Conformation Bias remain consistently rare throughout our study. These two results suggests that the influence and persuasion techniques used during this time period is mostly driven through salience amplification and emotional reinforcement rather than explicit authority endorsement cues or overt ideological alignment.

Figure 7.11, Figure 7.12 and Figure 7.13 illustrates the correlation between the use of cognitive bias triggers for each CSA type across the months. Each CSA type has its unique profile for the use of cognitive bias triggers, and the use of these triggers by each archetype remains stable over time. Amplifier, engagement generation and coordinated agents exhibit strong co-occurrence between availability bias, negativity bias and affect bias. This reflects a strategy centered on repetition, emotional activation and sustained exposure. These agents rely on a narrow but effective subset of bias triggers, which helps them reinforce dominant narratives rather than diversify the cognitive appeals of their message. Conversational and bridging agents have broader and more diffused correlations across multiple bias triggers, such as having moderate associations with cognitive dissonance and illusory truth effects. Information correction agents show weak correlations with most bias triggers, which matches their corrective orientation.

7.6 Network Interaction and Coordination Profiles

7.6.1 Coordination Analysis

Using the algorithm developed in section 5.4, we calculated the extent of coordination of each user. This extent of coordination is measured by: semantic, referral, social coordination, and an overall coordination (Combined Synchronization Index).

Then, we used a Tweedie regression formula to examine the association between each CSA type and the coordination metrics. In the Tweedie regression, let i index users, and $X_i = (X_{i1}, ..., X_{im})$ be a vector of binary indicators denoting whether user i exhibits each of the m CSA behaviors. We model the expected coordination level using a Tweedie generalized linear model with a log link. The equation of the Tweedie regression is reflected in Equation 7.1. A log link is used because (1) to counter zero inflation, where many users do not exhibit detectable coordination, and so their coordination score is 0, and (2) right-skewed continuous values, where the coordinated users can exhibit widely varying magnitudes of synchronization. The Tweedie distribution with $1 < p < 2$ naturally models this structure by combining (1) a Poisson process governing whether coordination is detected at all, and (2) a Gamma distribution governing the magnitude of coordination when it does occur.

$$
\begin{aligned}
Y_i &\sim \text{Tweedie}(\mu_i, \phi, p), \quad 1 < p < 2 \\
\log(\mu_i) &= \beta_0 + \sum_{j=1}^{m} \beta_j X_{ij} \qquad (7.1) \\
\text{where } \mu_i &= \mathbb{E}[Y_i \mid \mathbf{X}_i]
\end{aligned}
$$

To estimate the model, we first standardized the feature matrix X that represents CSA type and the coordination values using variance scaling. This improves numerical conditioning while preserving sparse-friendly behavior. We then regressed these binary indicators against the coordination metrics. We fitted a Tweedie generalized linear model with a log link and power parameter $p = 1.5$, which corresponds to a compound Poisson-Gamma distribution. The model was optimized using iterative maximum likelihood estimation with a fixed iteration budget. The resulting coefficients quantify the marginal association between each CSA type and expected coordination, holding other CSA behaviors constant.

Table 7.2 shows the effects of Cyber Social Agent Types on the coordination metrics, demonstrating how coordination is strongly shaped by agent roles that facilitate agent interaction and information brokerage. Across all four coordination metrics, bridging agents consistently exhibit the strongest positive association with coordination. This pattern is especially pronounced for semantic and referral coordination, because bridging agents play a central role in linking otherwise separate communities. Conversationalist agents also display a robust and consistent positive association with coordination across all measures, indicating that such agents optimized for interaction and dialogue contribute meaningfully to narrative alignment (high semantic coordination). In contrast, amplification-oriented agents like the amplifier, announcer, engagement-generator agents exhibit uniformly negative coefficients across all coordination signals. This suggests that while these agents may contribute to volume and visibility, their behavior is less structurally coordinated at the network level when measured through artifacts like mentions (social coordination),

hashtags (semantic coordination) and URLs (referral coordination). Each column corresponds to a distinct coordination signal derived from network interactions. The values in the table reflect that each type of Cyber Social Agent have significant effect on the extent of coordination within the network, and the effect of the presence of each type of agent on the extent of coordination are different.

However, many of these agent types are co-related. Table 7.3 shows the VIF (Variation Inflation Factor) corresponding to each agent type. Some agent types have high VIFs, most likely because these two personas often co-occur together. Examples are: (Content-Generator, Announcer), and (Content-Generator, Engagement-Generator).

Agent Type	**Combined Synchronization Index**	**Social Coordination**	**Semantic Coordination**	**Referral Coordination**
Intercept (log-link)	-2.742	5.618	6.747	6.168
Bridger	0.435***	1.022***	0.215***	0.903***
Conversationalist	0.318***	0.244***	0.409***	0.176***
Chaos Agent	0.105***	0.056***	0.099***	0.070***
Social-Influencer	0.003***	-0.006***	0.056***	0.000
Self-Proclaimer	0.081***	-0.025***	0.092***	-0.037***
Broadcaster	0.072***	-0.177***	0.131***	-0.091***
Repeater	0.020***	-0.445***	0.236***	-0.592***
Information-Corrector	-0.133***	-0.083***	-0.132***	-0.131***
Amplifier	-0.428***	-0.443***	-0.383***	-0.438***
Announcer	-0.428***	-0.443***	-0.383***	-0.438***
Content-Generator	-0.428***	-0.443***	-0.383***	-0.438***
Synchronizer	-0.428***	-0.443***	-0.383***	-0.438***
Engagement-Generator	-0.428***	-0.443***	-0.383***	-0.438***
Mono-maniac	-0.428***	-0.443***	-0.383***	-0.438***
Cyborg	-0.428***	-0.443***	-0.383***	-0.438***

Table 7.2: Effects of Cyber Social Agent Types on Coordination Metrics. * indicates significant effect at $p < 0.05$, ** indicates significant effect at $p < 0.01$ and *** indicates significant effect at $p < 0.001$

Agent Type	**Combined**	**Social (Mentions)**	**Semantic (Hashtags)**	**Referral (URLs)**
Amplifier	1.00×10^{12}	1.00×10^{12}	1.00×10^{12}	1.00×10^{12}
Announcer	1.00×10^{12}	1.00×10^{12}	1.00×10^{12}	1.00×10^{12}
Bridger	4.00	4.00	4.00	4.00
Chaos-Creator	3.69×10^{2}	3.69×10^{2}	3.69×10^{2}	3.69×10^{2}
Content-Generator	1.00×10^{12}	1.00×10^{12}	1.00×10^{12}	1.00×10^{12}

Agent Type	Combined	Social	Semantic	Referral
Conversationalist	1.21	1.21	1.21	1.21
Synchronizer	1.00×10^{12}	1.00×10^{12}	1.00×10^{12}	1.00×10^{12}
Engagement-Generator	1.00×10^{12}	1.00×10^{12}	1.00×10^{12}	1.00×10^{12}
Mono-maniac	1.00×10^{12}	1.00×10^{12}	1.00×10^{12}	1.00×10^{12}
Information-Corrector	1.07	1.07	1.07	1.07
Broadcaster	1.05	1.05	1.05	1.05
Repeater	1.28	1.28	1.28	1.28
Self-Proclaimer	1.02	1.02	1.02	1.02
Social-Influencer	1.01	1.01	1.01	1.01

Table 7.3: Variance Inflation Factors (VIFs) for Cyber Social Agent indicators across all coordination models.

7.6.2 Network Impacts: Flipping Stance

As per section 5.5, an agent is defined to have flipped it stance if it has at least three tweets of a certain stance before another three tweets of the opposite stance. To characterize pro-Ukraine and pro-Russian stances within the dataset, we used the stance labels that were annotated on each agent and each tweet as defined in the section section 4.5. These stance labels were derived from the hashtags present within the tweet.

With that, we find that 51.48% of the users flipped their stances. This is higher as compared to the study with the 2020 coronavirus data, where about 1% of the users flipped their stances. This could be attributed to the geographical effect, where the closer a person is to an event, the less affected the person is to other information[267]. The coronavirus affected people worldwide, and thus they had personal experiences which contributed to the conviction and innate beliefs of their stance on the coronavirus vaccine. However, most of the online users were farther removed from the Russian-Ukraine conflict, and based their information through online sources, which therefore results in more frequent stance changes based on the incoming online information.

To examine the variables that contribute to the flipping stance phenomenon, we performed an Ordinary Least Squares (OLS) regression where the dependent variable is a binary variable [1,0] representing whether the agent flipped stances. We tested six different model variations of independent variables: (1) agent network, (2) all tweet linguistic cues, (3) agent network + tweet semantic linguistic cues, (4) agent network + tweet semantic linguistic cues + tweet emotion linguistic cues, (5) agent network + tweet semantic linguistic cues + tweet emotion linguistic cues + tweet narrative frames linguistic cues, (6) agent network + tweet semantic linguistic cues + tweet emotion linguistic cues + tweet narrative frames linguistic cues + moral values linguistic cues. Table 7.4 tabulates the coefficients of the variables of the models.

Across the six model specifications (Table 7.4), the results show a progression in explanatory power as richer linguistic features are incorporated. The agent network-only model (Model 1) explains little variance in stance flipping ($r^2 = 0.07$), indicating that the agent's network position alone is insufficient to account for stance changing behavior. In contrast, models that incorporate tweet-level linguistic cues (Models 2-6) achieve substantially higher explanatory power ($r^2 \sim$

0.95), demonstrating that what agents say is far more predictive of stance flipping than where they are positioned in the network. Linguistic cues that exhibit strong and significant associations with the stance flipping behavior are related to pronoun usage (1st person, 2nd person, 3rd person), symbolic concepts (# symbolic concepts), identities (# identities) and emotional expressions.

Models 3 to 6 build on each other by adding linguistic layers, from semantic cues (Model 3) to emotion (Model 4) to narrative frames (Model 5) and moral values (Model 6). The coefficients remain stable in sign and magnitude as more layers are added. This suggests that the OLS model is robust in its prediction and does not overfit. Emotional intensity (particularly negative and embarrassed emoticons), identity-related language, and moral value framings tied to care, fairness, loyalty, authority, and liberty are all positively associated with stance flipping, while religious framing and certain moral vice dimensions exhibit negative effects.

This study indicates that the stance flipping mechanism arises from an interaction of network exposure and linguistic persuasion mechanisms. Degree-based centrality measures remain statistically significant across all model specifications, demonstrating that an agent's position in the network affects the volume and diversity of content that they encounter. However, the linguistic cues of semantic, emotional, narrative and moral cues also yield substantial gains in explanatory power. Therefore, the result of stance flipping is a two stage mechanism: network centrality governs who is exposed to persuasive messages and how often, while linguistic features shape how that exposure translates into stance change.

Variable	**Model 1**	**Model 2**	**Model 3**	**Model 4**	**Model 5**	**Model 6**
r^2	0.07	0.95	0.95	0.95	0.95	0.95
constant	-1.10E-3**	5.00E-5**	-4.77E-5**	-4.63E-5**	-2.95E-5**	-3.23E-5**
in degree centrality	1997.50**		73.24**	68.97**	49.07**	49.94**
out degree centrality	-32.32**		-21.40**	-22.42**	-20.26**	-20.15**
total degree centrality	127.80**		61.00**	61.83**	60.41**	60.23**
1st person		0.38**	0.41**	0.39**	0.39**	0.38**
2nd person		-0.02**	-0.07**	-0.06**	0.00	-0.02**
3rd person		0.28**	0.34**	0.33**	0.31**	0.28**
abusive		-0.00	-0.20**	-0.21**	-0.05**	0.00
exclusive		0.18**	0.38**	0.36**	0.27**	0.18**
# symbolic concepts		0.50**	0.55**	0.52**	0.50**	0.50**
# identities		0.13**	0.26**	0.27**	0.09**	0.13**
# positive emoticons		2.04**		1.95**	2.08**	2.04**
# positive emoji		0.52**		0.51**	0.57**	0.52**
# neutral emoticons		27.19**		29.37**	25.22**	26.95**
# neutral emoji		0.18**		0.20**	0.17**	0.18**
# negative emoticons		2.56**		2.77**	2.51**	2.55**

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
# negative emoji		0.05**		-0.05**	0.01**	0.05**
# happy emots/emojis		-0.38**		-0.41**	-0.46**	-0.38**
# sad emots/emojis		-1.04**		-0.86**	-0.91**	-1.04**
# angry emots/emojis		-0.18**		-0.12**	-0.10**	-0.18**
# embarrassed emots/emojis		2.35**		2.02**	2.10**	2.33**
# family		-0.21**			-0.07**	-0.21**
# political		-0.23**			-0.23**	-0.23**
# gender		0.05**			0.04**	0.05**
# religion		-0.77**			-0.63**	-0.77**
# race/nationality		0.36**			0.49**	0.35**
# job		0.13**			0.27**	0.13**
# other		0.02**			0.15**	0.02**
moral value: care_virtue		0.29**				0.29**
moral value: care_vice_harm		0.22**				0.22**
moral value: fairness_virtue		0.22**				0.22**
moral value: fairness_vice_cheating		-0.13**				-0.12**
moral value: loyalty_virtue		0.16**				0.15**
moral value: loyalty_vice_betrayal		-0.71**				-0.71**
moral value: authority_virtue		0.12**				0.13**
moral value: authority_vice_subversion		0.23**				0.23**
moral value: sanctity_virtue		0.19**				0.19**
moral value: sanctity_vice_degradation		-0.04**				-0.04**
moral value: liberty_virtue		0.24**				0.24**
moral value: liberty_vice_oppression		0.19**				0.19**

Variable	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6

Table 7.4: OLS results predicting "flipped stance" across Models 1–6. Coefficients shown with two-decimal precision. * $p < 0.05$, ** $p < 0.01$.

In terms of the types of CSAs that flip stances, Figure 7.14 the proportion of each type of agent that flip stances. Conversational agents and social influence agents flip stances most frequently, owing to the nature of their behavior. Conversational agents engage dynamically with ongoing online discussions, and might need to strategically and retroactively adjust their stance to maintain conversational relevance and engagement. Similarly, social influence agents are optimized to maximize the reach and persuasion of their narratives within the network. Stance flipping in the case of conversational agents and social influence agents may reflect strategic repositioning of the agent within the network rather than ideological inconsistency, so that the agent can insert itself into multiple narrative communities over time.

Agents like news and bridging agents exhibit moderate levels of stance flipping. News agents relay external information sources, and so shifts in stances might mirror changes in the information ecosystem rather than ideological shifts or decisions of the news agencies. Bridging agents, by definition, connect otherwise disparate communities, and their exposure to conflicting narratives across clusters increases the likelihood of propagating opposite stances.

Finally, information correction agents, repeater agents and self-declared agents display the lowest rates of stance flipping. These agents have a very pre-defined set of behavior and narratives to follow that they seldom deviate from, hence there is little space for stance flipping. For example, information correction agents are typically anchored to a corrective or fact-checking role, and therefore a relatively stable ideological position. Repeater agents amplify a narrow set of messages. Self-declared agents have already explicitly signal their affiliation or intent, which reduces the incentive for stance flipping because that might appear to reduce their credibility.

7.7 Conclusion

This chapter is a case study of how all the techniques developed over the course of this thesis can be applied to study an event in depth. These techniques range from bot detection, identifying the types of Cyber Social Agents and their social cyber geographical spread, narrative analysis, interaction analysis through agent-agent coordination and stance flips. While this chapter provides a comprehensive overview of the human-agentic interactions during the Russian-Ukraine conflict, there are certain **limitations** that nuance its conclusions:

1. The curated dataset relied on keywords related to the conflict. This might omit relevant discussions that do not utilize these keywords, potentially narrowing the breadth of the captured narrative and overlooking subtler, yet possibly significant aspects of the conversation
2. Our dataset was limited to English language tweets. This thus excludes non-English discourse, which might present different perspectives or intensities in stance and interactions, particularly in regions directly affected by the conflict.

Future work involves using a multilingual dataset that analyzes Russian, Ukrainian and other relevant languages. Following studies can also focus on a longitudinal study that spans a more expanded period to allow for tracking the evolution of narratives over time and provide insights on the change and the longitudinal patterns of the CSA interactions.

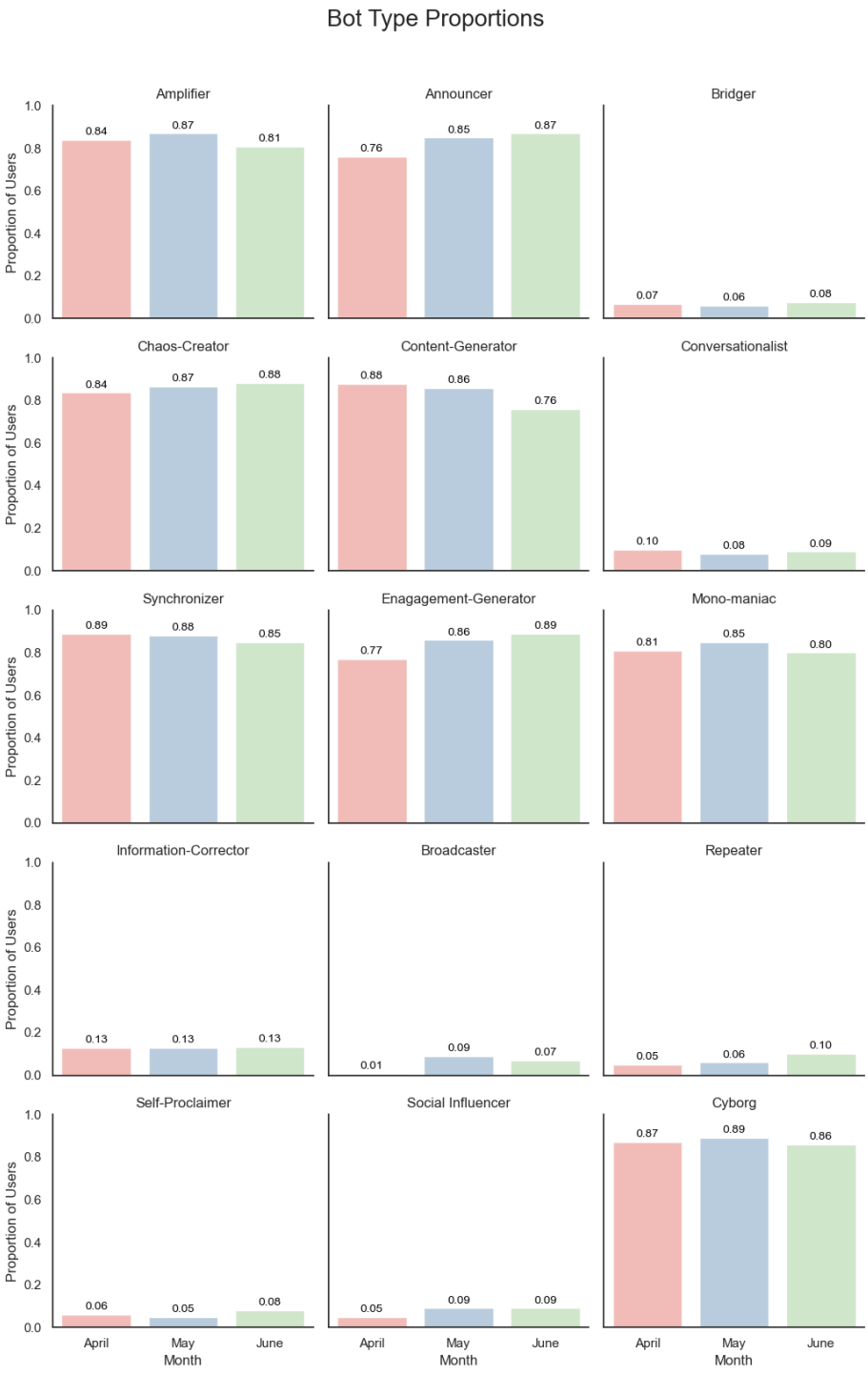


Figure 7.2: Proportions of each type of Cyber Social Agents by month

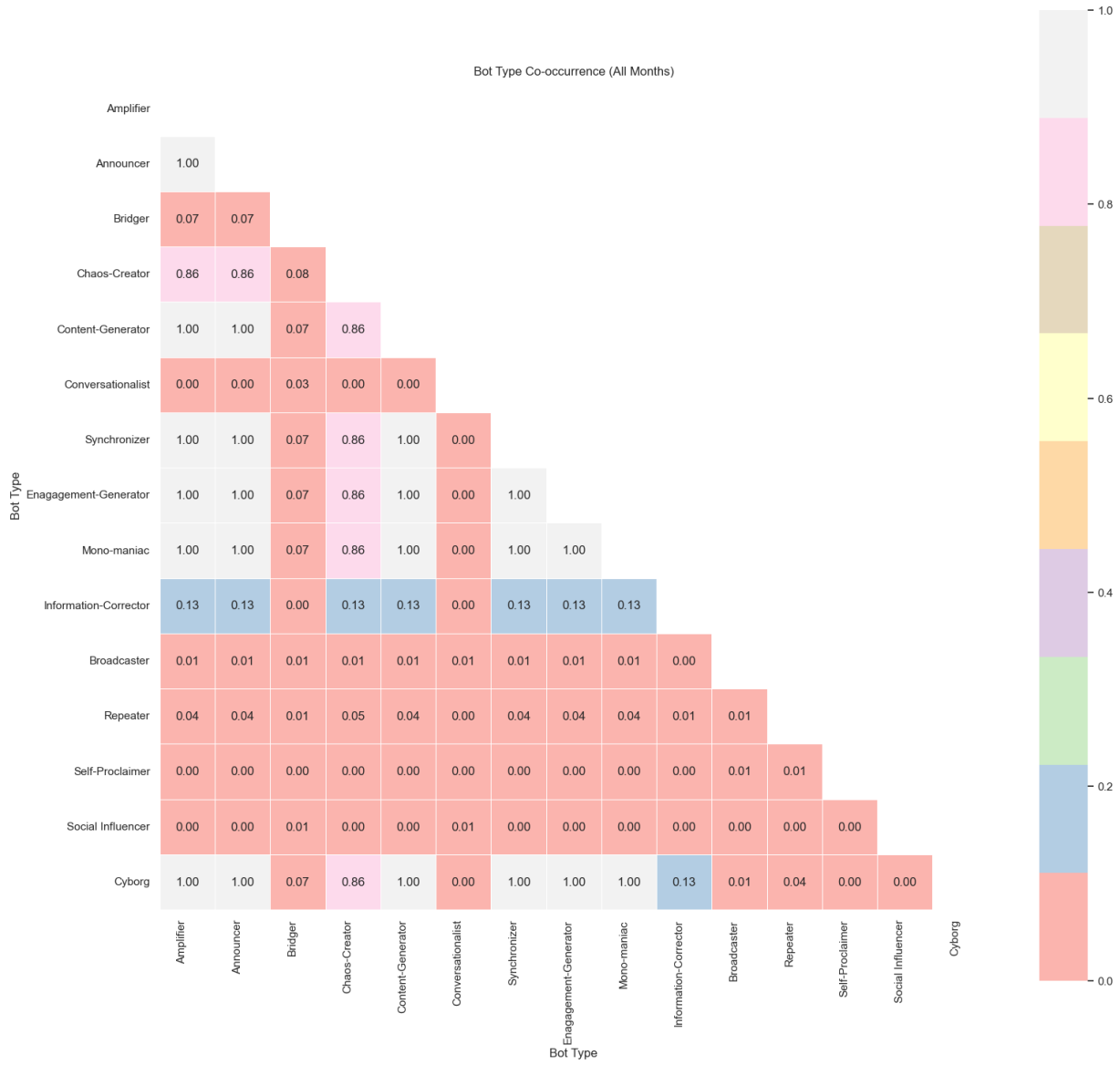


Figure 7.3: Co-occurrence of Cyber Social Agents. One user can take the behavior of multiple Cyber Social Agents.

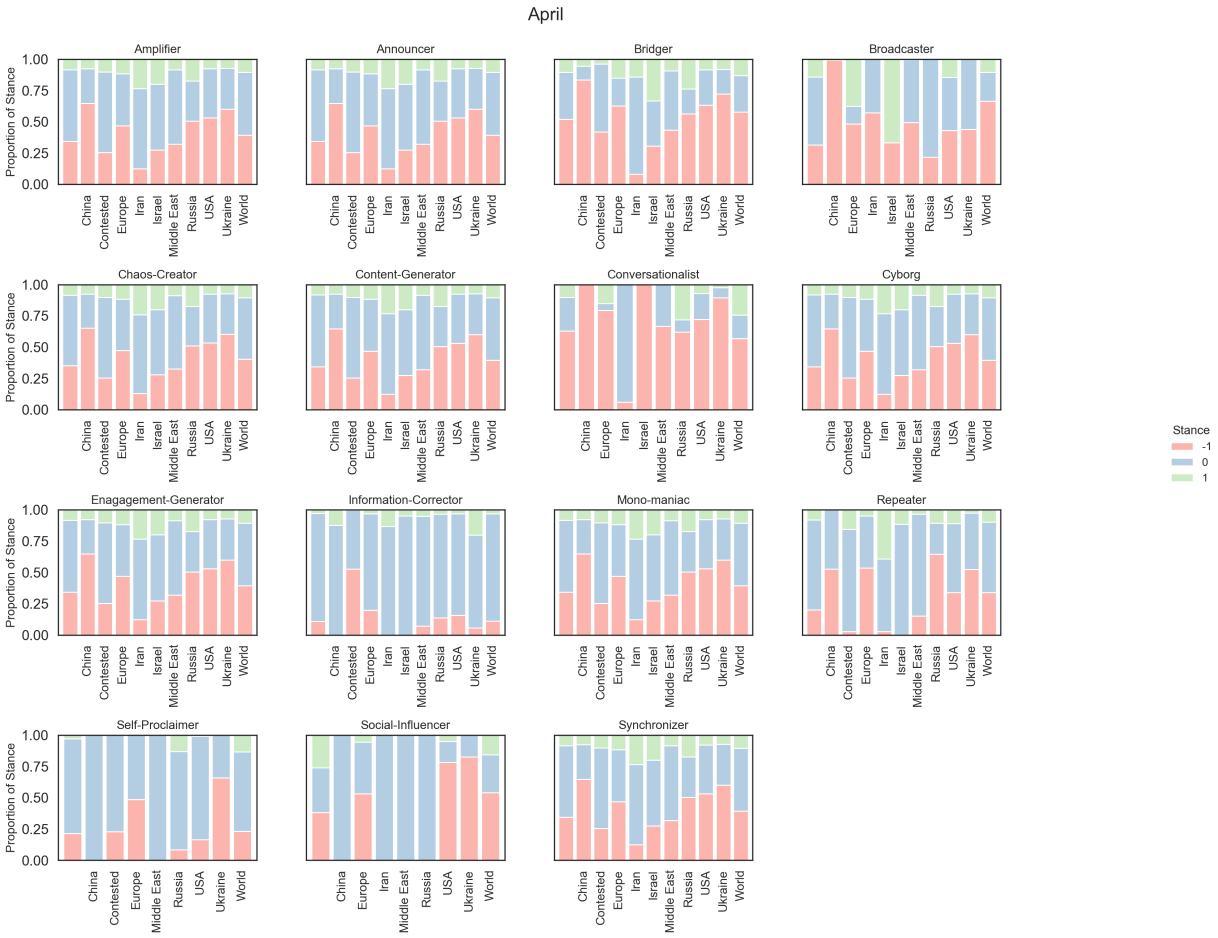


Figure 7.4: Stance by type of Agent for month of April. +1 means pro-Russia, -1 means pro-Ukraine, 0 means neutral stance

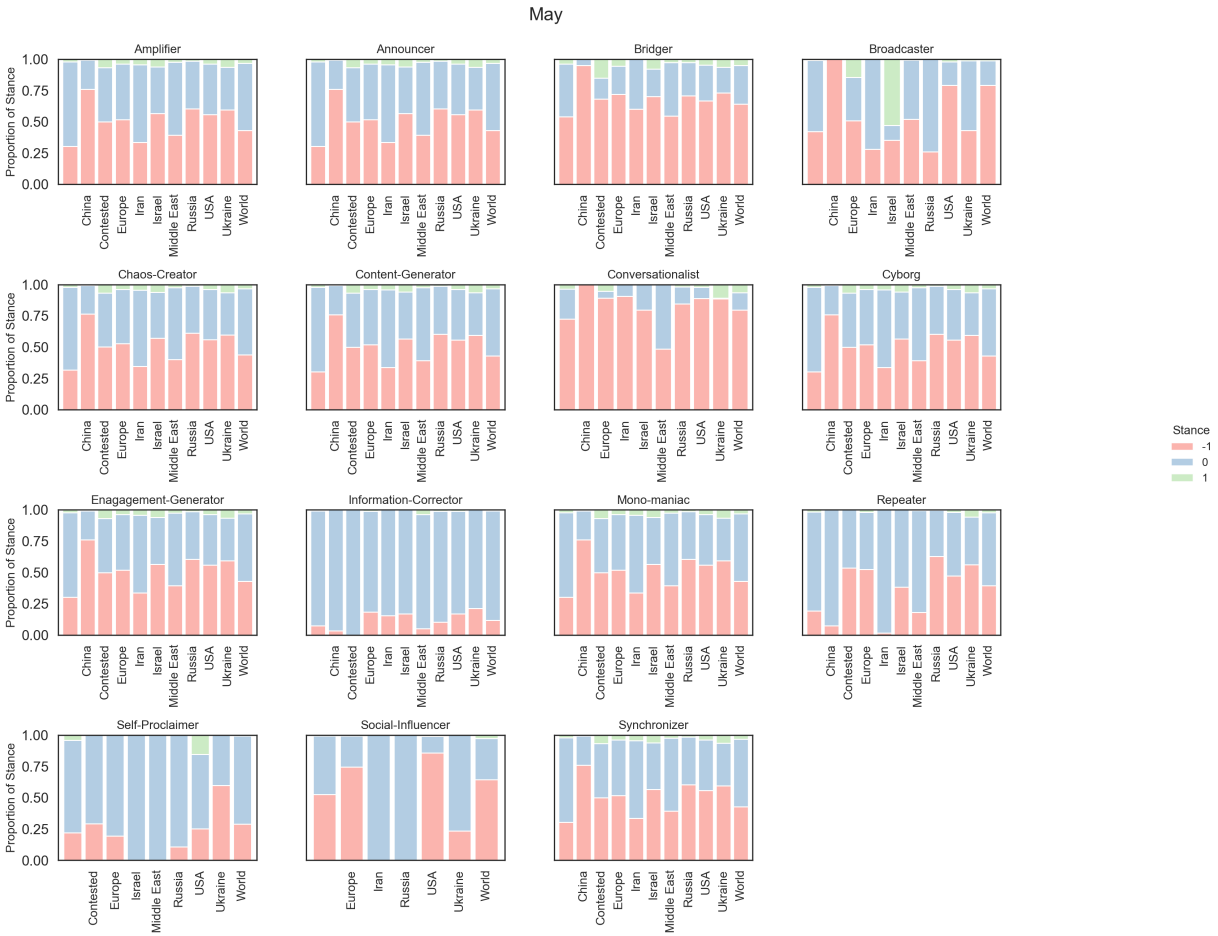


Figure 7.5: Stance by type of Agent for month of May. +1 means pro-Russia, -1 means pro-Ukraine, 0 means neutral stance

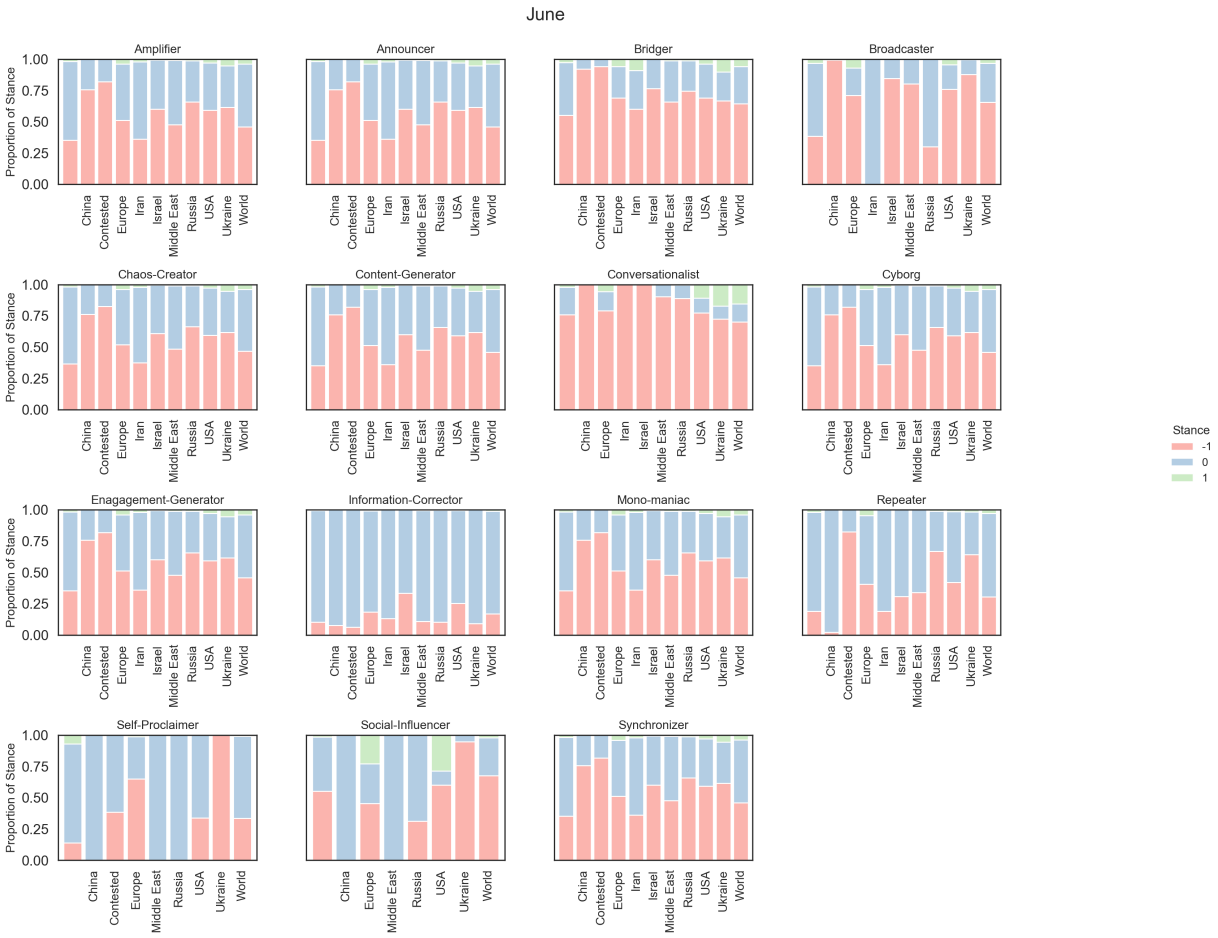


Figure 7.6: Stance by type of Agent for month of June. +1 means pro-Russia, -1 means pro-Ukraine, 0 means neutral stance

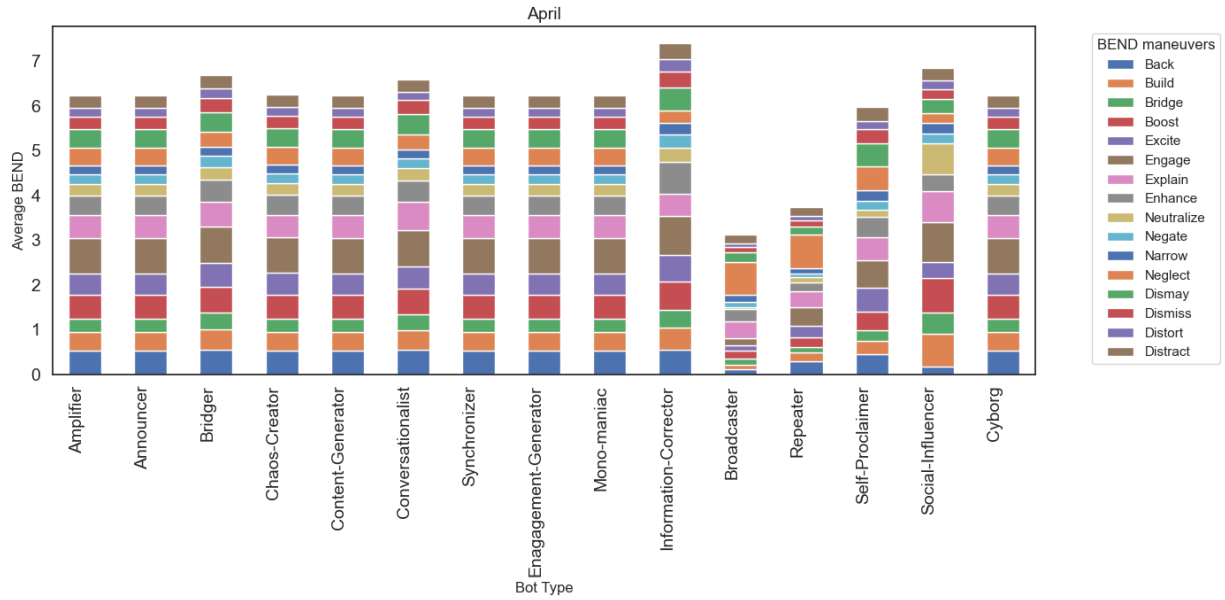


Figure 7.7: Distribution of average use of BEND framework by type of agent for month of April

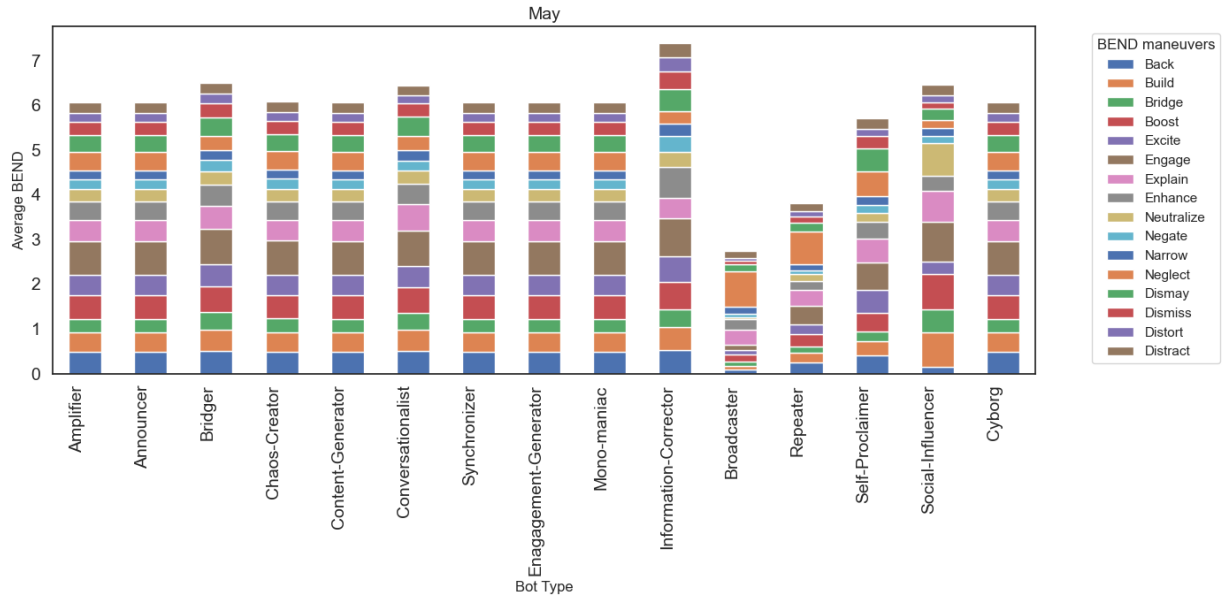


Figure 7.8: Distribution of average use of BEND framework by type of agent for month of May

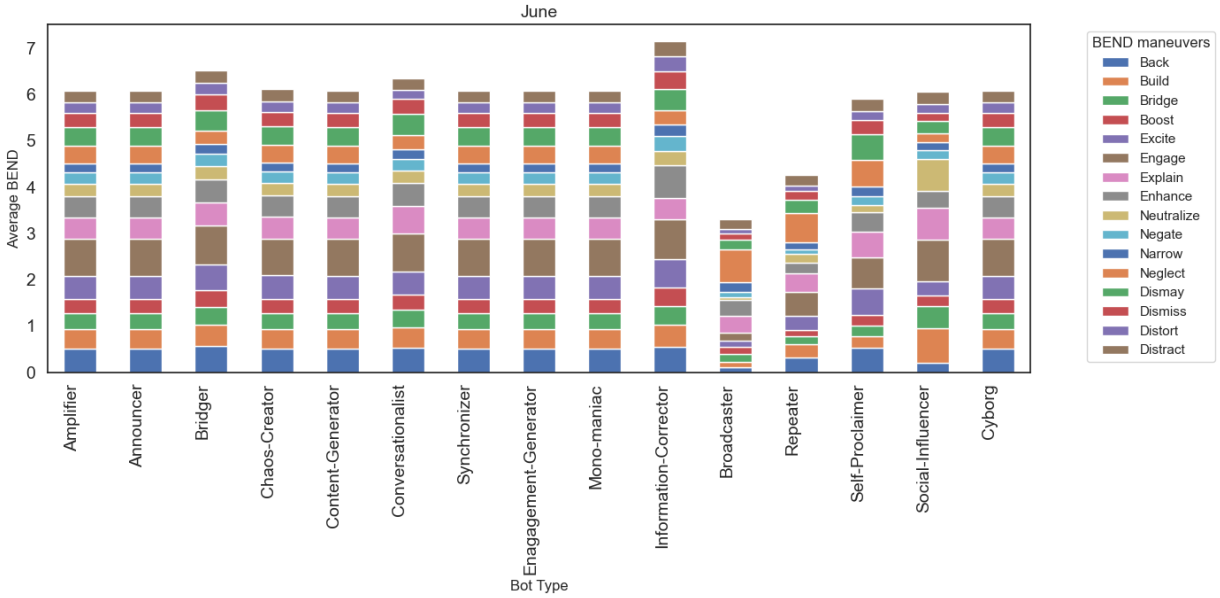


Figure 7.9: Distribution of average use of BEND framework by type of agent for month of June

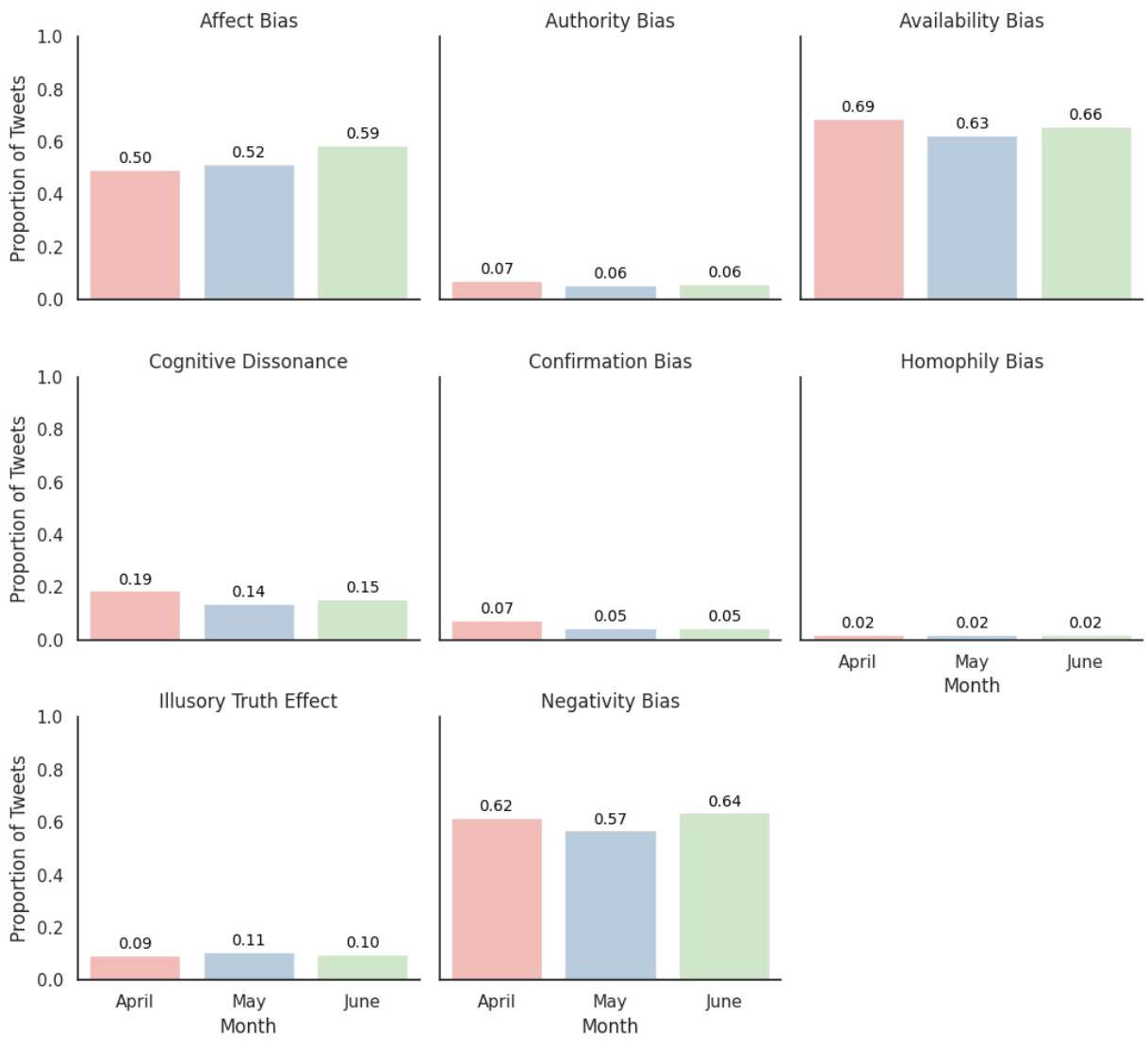


Figure 7.10: Distribution of presence of Cognitive Bias Triggers by Cyber Social Agents per month

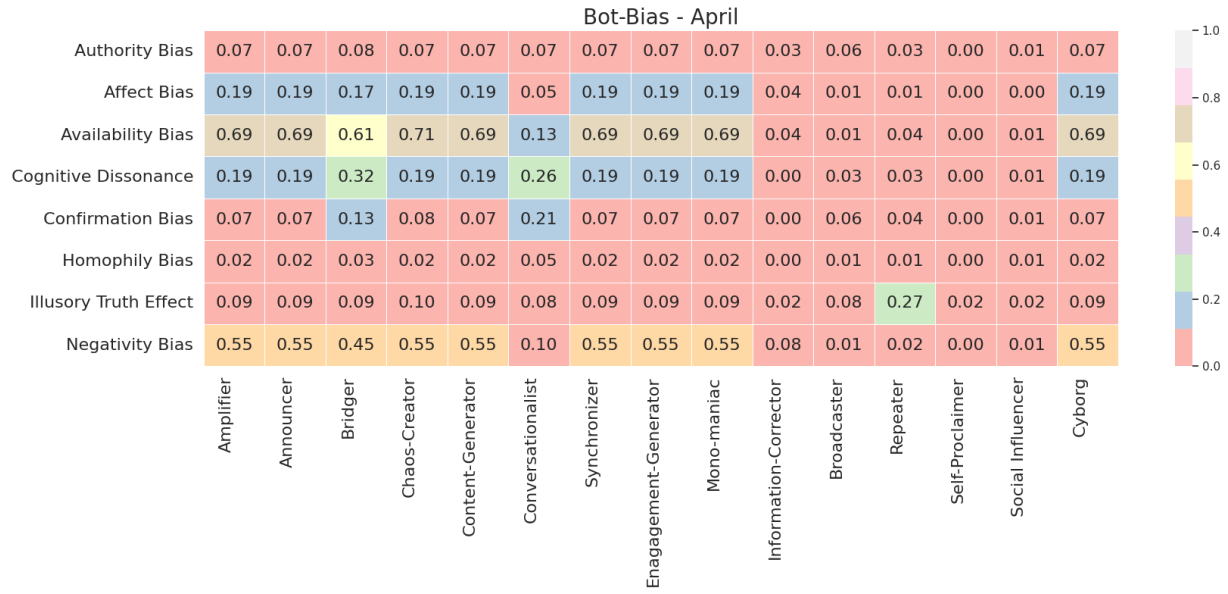


Figure 7.11: Correlation between use of cognitive bias triggers for CSA types for April

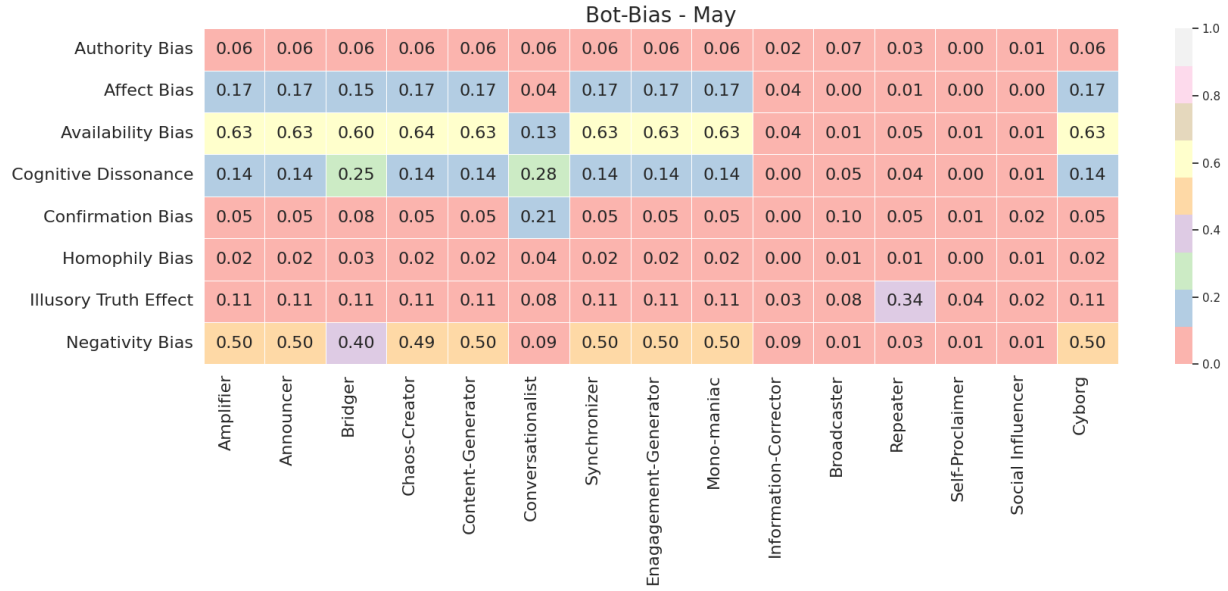


Figure 7.12: Correlation between use of cognitive bias triggers for CSA types for May

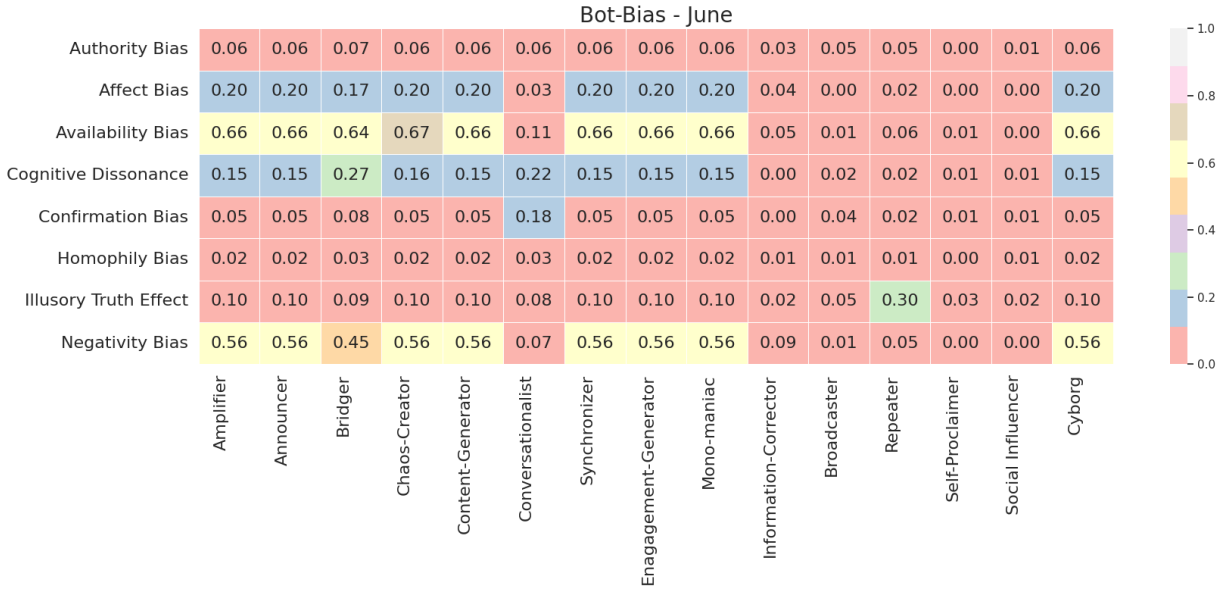


Figure 7.13: Correlation between use of cognitive bias triggers for CSA types for June

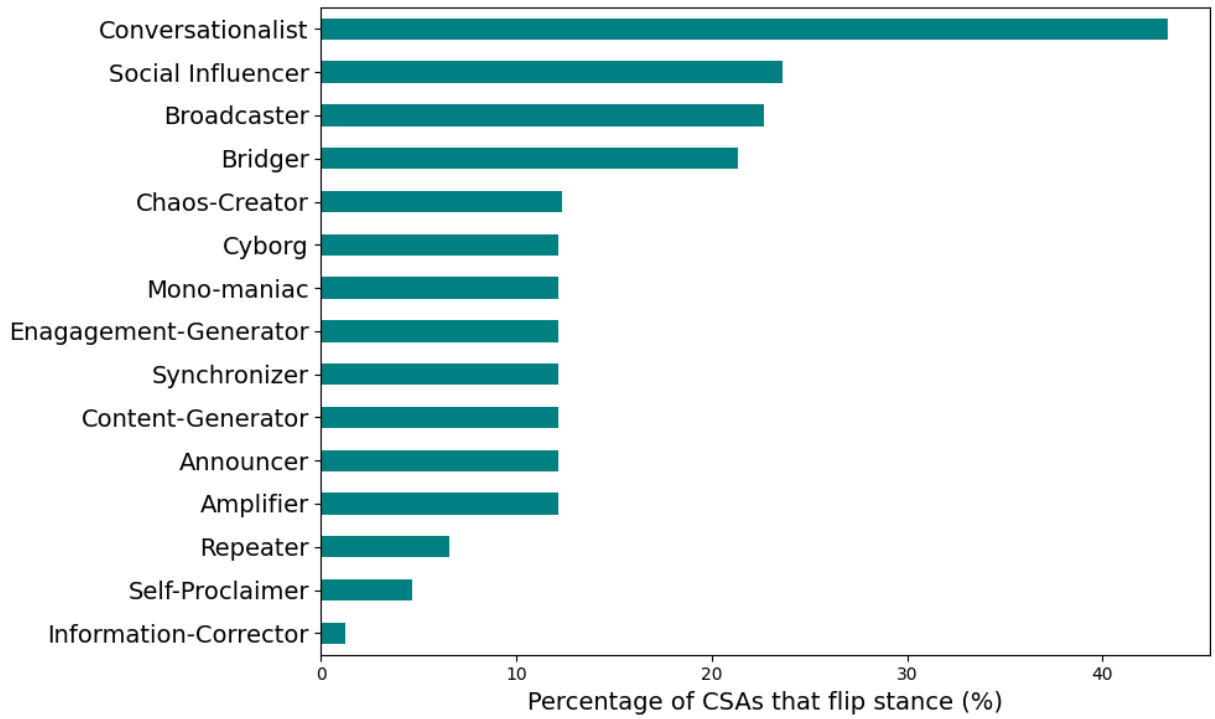


Figure 7.14: Types of Cyber Social Agents that flip stances

Chapter 8

Conclusions

The study of automated agents on social media, the "social media bot" is not only about identifying authentic vs. inauthentic accounts through machine learning algorithms, but it is also about understanding their role in our digital communication systems, the different types of inauthentic accounts, and harnessing their capabilities to improve the health, resilience and integrity of our social media ecosystem. In this thesis, we extend the concept of the "social media bot" beyond a task-automation framing towards a conceptualization of "Cyber Social Agents" as active participants in the construction of narrative shaping and agenda setting within digital publics. CSAs are not merely passive tools pre-programmed by human operators to perform a single isolated task, but are agents whose behaviors, content and interactions co-evolve with those of human users, other CSAs and platform architectures. Our insights advance the field of study from detection to defense, and from descriptive measurement to proactive stewardship of the online ecosystem.

Chapter 9

Contributions and Future Directions

9.0.1 Theoretic and Methodological

Table 9.1 summarizes the theoretical and methodological contributions, and limitations of this thesis. In **Bot Detection** (chapter 2), this thesis harmonizes industry and academic definitions of bots into a first-principles definition that emphasizes the user-content-interaction setup of social media platforms. The chapter advances bot detection by developing multi-platform and multilingual approaches, and also empirically validated bot detection thresholds. However, these methods remain limited to a small set of platforms (i.e., X, Reddit, Instagram, Telegram), and only partially address non-English data.

In **From Bots to Cyber Social Agents** (chapter 3), this thesis expands the view of bots as a homogeneous group into a taxonomy of fifteen distinct personas of Cyber Social Agents. These personas are characterized by their behavioral and content signature. Then, it further refines the binary framing of "good versus bad" bots by outlining how each persona can be elements of either. These heuristics, however, are primarily tested on X data, and the operationalization of goodness is simplified into binary categories rather than a full spectrum.

In **Nature of Cyber Social Agents** (chapter 4), this thesis analyzes activity across multiple dimensions, providing a generic characterization of the nature of these automated agents, through social political, narrative, motivations & agencies, linguistic and cognitive bias triggers lens. Instead of analyzing isolated events on CSA activity, this chapter collectively studied large-scale events, capturing ego-network, interaction and information maneuver patterns over billions of tweets geolocated across the world. While the analysis is extensive, they are mostly performed for the X platforms and only a few types of networks (i.e., amplification network) were studied.

The **Network Interactions & Coordination Profiles of Cyber Social Agents** (chapter 5) chapter studied coordination patterns of CSAs across several dimensions: semantic, referral, social, visual, cross-platform. To measure coordination across multiple modalities, this thesis proposes and validated the Combined Synchronization Index. Further, this thesis models the stance-flipping behavior through a social influence model, showing the difference in influence dynamics and pressures on both CSAs and humans. Limitations lie in the scope of events and platforms analyzed.

With **Social Simulations of CSAs & Humans** (chapter 6), this thesis develops a hybrid approach that combines agent-based modeling with large language models, enabling the simula-

tion of social media networks with multiple bot types. This chapter also demonstrates the role of useful bots in countering conspiracy theories. However, the simulations remain limited to small-world networks and still face gaps in content and network realism.

2023 Russia-Ukraine Conflict (chapter 7) methodologically demonstrates how the different concepts of the Cyber Social Agent and the developed methods gel together to perform one cohesive analysis. While it successfully integrates the thesis' contributions into a single case study, it remains restricted to one empirical event.

	Chapter	Other's Work	This Thesis	Thesis Contributions	Thesis Limitations
2	**Bot Detection**	Bots are social media users	Bots are agents, i.e. entities that act on behalf of a person or group, or take an active role to produce specified effects Harmonized academic and industry definitions of a bot [212]	First-principles definition of a bot that emphasized user-content-interactions [212]	Definition focused on social media bots
		Bot detection primarily for X (Twitter) data Bot detection for a single platform Bot detection that requires live data pull Bot detection mostly uses the content of post	Bot detection that works for multiple platforms (Reddit, X, Instagram, Telegram), can use historical data, and is fast (does not require a GPU) [205, 210, 221]	Multi-platform detection [214] Fast algorithm Uses meta-data and content [205, 210, 221]	Only three platforms, only in English
		Bot detection typically tuned for the English language	Bot detection uses multi-lingual vectors, specifically tested for Russian, Chinese, Arabic languages [214]	Multi-lingual bot detection [214]	Only tested on translated data not real data (because no real annotators) Only tested on Russian, Chinese, Arabic
		Bot detection threshold is arbitrarily set	Determined a valid bot detection threshold and number of posts required for a stable threshold through longitudinal analysis. [205, 218] Methodology for determining threshold values that guarantees a minimum number of false positives [218]	BotHunter threshold is 0.7 [218] BotBuster threshold is 0.5 [205] Minimum number of posts is 20	Only done for BotHunter and BotBuster Only thresholded for the X platform
3	**From Bots to Cyber Social Agents**	All bots are the same Bots are a generic group	There are a range of different bot behaviors. Defined fifteen different bot personas, characterized by content and behavior. Bot personas can be computationally defined and detected.	Described fifteen bot personas. Developed heuristics to empirically detect each of the types of bots at a large scale [211]	Tested heuristics only on bots from X

	Chapter	Other's Work	Thesis Methodological Contributions	Thesis Theoretical Contributions	Thesis Limitations
		Bots are bad	Bots can be both good or bad.	Defined content/ behavior elements of good and bad bots for each type of bot [211]	Defined good/bad elements by observations from X [211] Made goodness a binary item rather than a spectrum
4	**Nature of Cyber Social Agents**	Text or topical analysis of bots vs humans	Deeper analysis of bots vs human characteristics: emotions, identities, linguistic features [202, 208, 212, 244, 294] Enabled the study of the identification of cyborgs	Identified quantitative characteristics of cyborgs Bot topics match their identities [222]	Only done for X platform
		Isolated event analysis of bot activities	Large scale analysis of 7 events spanning 4 years to establish global trends [212] Enabled the study of identification of bot types [207, 294]	Empirical bot vs human differences on billions of tweets about linguistic, identity, network values [212]	Only performed on the X platform
			Geographical analysis of how bots affiliate themselves across the world	Gazetter-based location detector [214]	Gazetter needs to be continually updated
		Defined types of information maneuvers in the form of BEND framework	Empirically analyzed the usage and distribution of the BEND maneuvers by different types of bots. [70, 207]	Empirical differences and distribution of the use of BEND maneuvers for bot vs humans. In digital diplomacy, bots use more B's and E's maneuvers. [207] In religious ideology spread, bots use more B's and D's maneuvers. [70]	Only studied general, news and bridging bots.

	Chapter	Other's Work	Thesis Methodological Contributions	Thesis Theoretical Contributions	Thesis Limitations
5	**Network Interactions & Coordinated Profiles of Cyber Social Agents**	Bots coordinate with other Bots	Defined a combined synchronized index to measure coordination across multiple dimensions Defined and studied coordination patterns across several dimensions: semantic, referral, social, visual, cross-platform [70, 203, 216, 217] Profiled the ego-alter coordination of 30 bots and humans [203] Profile political coordination against automation [220]	Designed the Combined Coordination Index implemented as the Coordination Analysis report in the ORA software [206] Observed that humans coordinate with other humans, but Bots coordinate more with humans [203]	Mostly tested coordination on X Coordination Index only fully tested on X, Reddit, Facebook
		Bots form polarized networks	Compared the ego-networks of bots and humans to profile the stereotypical network formations [212]	Political bots have a stereotypical star network, humans have a hierarchal network structure [137, 212]	Only studied political bots, other bots might have slightly different formations
			Analyzed bot-human star motifs used for amplification (retweeting) within COVID-19 discourse [226]	Bots-Humans can form star networks for amplification of information. There are six different types of star network structures, each of them being a strategy. [226]	Only studied undirected networks. Only studied amplification networks.
		Claims that bots can manipulate stances but does not demonstrate empirical evidence. Claims that bot networks can result in polarization.	Observed and modeled one of the effects of bot presence. This effect is the stance flipping behavior, where users can change their stance. Modeled stance flipping as a social influence model by Friedkin-Johnson [204]	Demonstrated that both bots and humans follow the laws of social influence Show the parameters that are more likely to trigger stance flipping: if user is a bot, if neighbors coordinate, if $\geq 80\%$ of neighbors have different stance than ego [204]	Only performed on one event

	Chapter	Other's Work	Thesis Methodological Contributions	Thesis Theoretical Contributions	Thesis Limitations
		Assumed that when bots tweet misinformation, it will be highly engaged with, and does not provide an explanation for what that is	Large scale study of bots spreading misinformation, assessed level of engagement (appeal and scope) [227], and characteristics (cognitive triggers) of the misinformation statements [223]	Bots are used strategically to put out misinformation Bots mostly retweet/share rather than create misinformation [223] Engagement and misinformation usage are not the same thing - engagement can be both positively or negatively correlated with misinformation Provided an explanation for engagement with misinformation - engagement has to do with cognitive bias triggers present in tweet Level of engagement can be measured by appeal/scope metrics constructed based on network science concepts [227]	Only studied one event Only studied 8 cognitive biases
6	**Social Simulation of CSAs & Humans**	Simulating the spread of conspiracy (dis)information in a network of humans only	Simulated the spread of conspiracy information in a synthetic small-world network with conspiracy bots, good bots and information correction bots	Established the need for useful bots for countering conspiracy theories	Only simulated with two types of useful bots. Only simulated with small-world networks

	Chapter	Other's Work	Thesis Methodological Contributions	Thesis Theoretical Contributions	Thesis Limitations
		Modeling of social networks of humans only. Modeling of social networks with either agent-based modeling or LLM-based modeling. Evaluated the network with only content or network characteristics, and no comparison to empirical networks	Modeled a social network with humans and multiple types of bots, evaluated the concent and network characteristics. [225] Modeled a social network with a hybrid approach that uses agent-based and LLM-based modeling. Compared the modeled hybrid approach with empirical networks. [213]	First ever simulated social media environment with multiple types of bots [225]	Major gaps in data realism in terms of content and network characteristics
7	**2023 Russia-Ukraine Conflict**		Gels all concepts together with a cohesive analysis	Demonstrates the application of all concepts of this thesis on a single event	Limited to one event

Table 9.1: Theoretical and Methodological Contributions

9.0.2 Academic

Portions of this thesis has been published in several academic venues, including: Scientific Reports, EPJ Data Science, Online Social Networks and Media and AAAI ICWSM. Some of these papers have won Best Paper Awards, and some have been featured in magazines. Table 9.2 presents the papers that correspond towards each section of the thesis.

	Chapter	CSA Concepts	Papers Associated	Status
2	**Bot Detection**	What is a Bot?	BotBuster:Multi-platform bot detection using a mixture of experts (AAAI ICWSM, 2023) [205] Stabalizing a supervised bot detection algorithm: How much data is needed for consistent predictions? (Online Social Networks and Media, 2022, Best Paper Award) [218] An exploratory analysis of COVID bot vs human disinformation dissemination stemming from the Disinformation Dozen on Telegram (Journal of Computational Social Science, 2023) [221] Tiny-BotBuster: Identifying automated political coordination in digital campaigns (SBP-BRiMS, 2024) [220] Assembling a multi-platform ensemble social bot detector with applications to the US 2020 elections (Social Network Analysis and Mining, 2024) [210]	Published
		Review of bot definitions	A Global Comparison of Social Media Bot and human characteristics (Scientific Reports) [212]	Published
3	**From Bots to Cyber Social Agents**	Bot Personas	AuraSight: Generating Realistic Social Media Data (CMU Technical Report, 2025) [225] Cyborgs for strategic communications on social media (Big Data & Society, 2024) [222] Deflating the Chinese balloon: types of Twitter bots in US-China balloon incident (EPJ Data Science, 2023) [207]	Published
		Good and Bad of Bots	The Dual Personas of Social Media Bots (Book Chapter, 2025)	To appear
4	**Nature of Cyber Social Agents**	Narrative expressions	Bot-Based emotion behavior differences in images during Kashmir Black Day event (SBP-BRiMS, 2020) [202] Active, Aggressive, but to little avail: characterizing bot activity during the 2020 Singaporean elections (SBP-BRiMS, 2020) [294] Deflating the Chinese balloon: types of Twitter bots in US-China balloon incident (EPJ Data Science, 2023, featured in New Scientist) [207]	Published
			Bots exploit cognitive bias triggers to shape misinformation engagement	Under Review
		Motivations & Agencies	Appeal & Scope of Misinformation spread by AI Agents and Humans (AMCIS, 2025) [227] Analyzing social cyber maneuvers for spreading covid-19 pro- and anti-vaccine information (Book chapter, 2022)	Published
		Social Political Representation	Social Cyber Geographical Worldwide Inventory of Bots	In preparation

	Chapter	Concepts	Papers Associated	Paper Status
			Deflating the Chinese Balloon: types of Twitter bots in US-China balloon incident (EPJ Data Science, 2023, featured in New Scientist) [207]	Published
		Semantic Style	A Global Comparison of Social Media Bot and Human Characteristics (Scientific Reports, 2025) [212]	Published
5	**Network Interactions & Coordination Profiles of Cyber Social Agents**	Network Interaction profiles	A Global Comparison of Social Media Bot and Human Characteristics (Scientific Reports, 2025) [202] Star Network Motifs on X during COVID-19 (SBP-BRiMS, 2025) [226]	Published
		Synchronization & Coordination	A combined synchronization index for evaluating collective action on social media (Applied Network Science, 2023) [206] Cross-platform information spread during the January 6th capitol riots (Social Network Analysis and Mining, 2022) [216] Online coordination: methods and comparative case studies of coordinated groups across four events in the united states (ACM WebSci, 2022) [203] Do you hear the people sing? Comparison of synchronized URL and narrative themes in 2020 and 2023 French protests (Frontiers in Big Data) [209] Coordinating Narratives Framework for cross-platform analysis in the 2021 US Capitol Riots (CMOT) [219]	Published
		Network impacts	Pro or anti? A social influence model of online stance Flipping (IEEE TNSE, 2022) [204]	Published
		Content of interaction	Bots exploit cognitive bias triggers to shape misinformation engagement	Under Review
6	**Social Simulations of CSAs-Humans**	Modeling CSAs and Humans	AuraSight: Generating Realistic Social Media Data (CMU Technical Report, 2025) [225] Are LLM-Powered Bots Realistic? (SBP-BRiMS, 2025) [213]	Published
		Simulation of useful CSAs	BotSim: Mitigating the formation of Conspiratorial Societies (JASSS, 2026) [215]	Published

Table 9.2: Summary of Work from this Thesis

9.1 Future Directions

Emerging trends in automation and artificial intelligence suggests that Cyber Social Agents will increasingly function as general purpose systems of digital persuasion to influence social media discourse. They will operate across platforms, modalities and social contexts. Future directions of this thesis must therefore move towards a deeper understanding on how digital persuasion is changing as these artificial agents become more persuasive and more integrated into our digital lives.

From a computational perspective, future work must advance the modeling of Cyber Social Agents. Detection systems must capture not only the content and behavior, but also the contextual information and intent. That is, models should be able to discern how CSAs adapt persuasive strategies to audiences, platforms and moments in time. This includes understanding how CSAs can be used both constructively and harmfully, and developing technologies to mitigate harm while harnessing the good.

Looking ahead, we must also study whether future CSAs will become more persuasive than they are today. Advances in Large Language Models and multimodal generative AI suggest that CSAs may increasingly personalize narratives, exploit social content and coordinate influence campaigns at scale. Computational models must therefore represent the influence of CSAs as a dynamic process that unfolds over time, across platforms and through interaction with both humans and automated agents.

As the use of CSAs become present on nearly every digital platform, and more organizations and individuals harness their power, the scope of computational analysis must widen to study new uses of such agents. For example, new uses of cyborgs or autonomous coordination services for resource allocation. These new uses will reshape the influence ecosystem in ways that the current paradigms, many of which are presented in this thesis, do not capture. Such digital change will raise foundational questions about how agency, autonomy and persuasion should be operationalized in cyber agents that can increasingly act and interact independently without direct human control.

From a sociological perspective, future work must examine how the ubiquity of Cyber Social Agents reshape social influence and persuasion in digital societies. As CSAs become embedded across platforms and social contexts, humans may no longer encounter them as exceptional or suspicious actors, but get used to them being routine participants of online life. This normalization blurs the human-machine boundary, where the difference between organic and engineered influence becomes increasingly difficult to perceive, potentially altering the social construct of information spread.

In the current landscape, CSAs are viewed as subservient tools that work under human direction. Human-machine teaming suggests a future where CSAs may assume quasi-leadership roles for human communities. They may drive conversations, launch campaigns or coordinate behavior across human communities. Sociological research must therefore examine how people interpret influence when it originates from non-human agents and how power shifts when CSAs shape agendas.

Finally, governance frameworks must evolve alongside these sociotechnical changes. Effective governance can no longer rely on simplistic “bot” vs “human” distinctions. Instead, it must be grounded in a nuanced understanding of the automation of digital agents, and evaluate CSAs

holistically based on their behavior, intent and impact. By doing so, only can societies constrain harmful manipulation while enabling constructive forms of digital persuasion (i.e., civic engagement, crisis coordination).

Bibliography

[1] Norah Abokhodair, Daisy Yoo, and David W McDonald. Dissecting a social botnet: Growth, content and influence in twitter. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 839–851, 2015. 3.2, 3.4, 3.3, 3.5

[2] Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B Divya. Agentic ai: Autonomous intelligence for complex goals–a comprehensive survey. *IEEe Access*, 2025. 3.2

[3] Nitin Agarwal, Samer Al-Khateeb, Rick Galeano, and Rebecca Goolsby. Examining the use of botnets and their evolution in propaganda dissemination. *Defence Strategic Communications*, 2(1):87–112, 2017. 7.2

[4] Esma Aïmeur, Sabrine Amri, and Gilles Brassard. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30, 2023. 3.2, 3.5

[5] Mohammad Majid Akhtar, Rahat Masood, Muhammad Ikram, and Salil S Kanhere. Sok: False information, bots and malicious campaigns: Demystifying elements of social media manipulations. In *Proceedings of the 19th ACM Asia Conference on Computer and Communications Security*, pages 1784–1800, 2024. 3.3

[6] Raad Al-azawi and O Safaa. Feature extractions and selection of bot detection on twitter a systematic literature review: Feature extractions and selection of bot detection on twitter a systematic literature review. *Inteligencia Artificial*, 25(69):57–86, 2022. 2.2.2

[7] Ahmed Al-Rawi and Vishal Shukla. Bots as active news promoters: A digital analysis of covid-19 tweets. *Information*, 11(10):461, 2020. 3.5

[8] Abeer Aldayel and Walid Magdy. Characterizing the role of bots' in polarized stance on social media. *Social Network Analysis and Mining*, 12(1):30, 2022. 3.5, 4.6

[9] Iuliia Alieva, JD Moffitt, and Kathleen M Carley. How disinformation operations against russian opposition leader alexei navalny influence the international audience on twitter. *Social Network Analysis and Mining*, 12(1):80, 2022. 7.2

[10] Malak Aljabri, Rachid Zagrouba, Afrah Shaahid, Fatima Alnasser, Asalah Saleh, and Dorieh M Alomari. Machine learning-based social media bot detection: a comprehensive literature review. *Social Network Analysis and Mining*, 13(1):20, 2023. 3.3

[11] Panagiotis Andriotis and Atsuhiro Takasu. Emotional bots: content-based spammer detection on social media. In *2018 IEEE international workshop on information forensics*

and security (WIFS), pages 1–8. IEEE, 2018. 3.3

[12] Clio Andris. Integrating social network data into gisystems. *International Journal of Geographical Information Science*, 30(10):2009–2031, 2016. 4.2

[13] Sinan Aral. The future of weak ties. *American Journal of Sociology*, 121(6):1931–1939, 2016. 5.3.1

[14] Marc-André Argentino. Pastel qanon – gnet, 2021. URL `https://gnet-research.org/2021/03/17/pastel-qanon/`. [Accessed 2024-02-12]. 4.2

[15] Valerio Arnaboldi, Marco Conti, Andrea Passarella, and Robin IM Dunbar. Online social networks and information diffusion: The role of ego networks. *Online Social Networks and Media*, 1:44–55, 2017. 5.3.2

[16] Dan Arnaudo. Computational propaganda in brazil: Social bots during elections. 2017. 3.5

[17] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596, 2008. 2.4

[18] Derrik E. Asher, Justine Caylor, Casey Doyle, Alexis R. Neigel, György Korniss, and Boleslaw K. Szymanski. Opinion formation threshold estimates from different combinations of social media data-types. In *Proceedings of the 51st Hawaii International Conference on System Sciences (HICSS 2018)*, pages 2045–2054, 2018. URL `https://arxiv.org/abs/1810.01501`. arXiv preprint arXiv:1810.01501. 6.2

[19] Hadi Askari, Anshuman Chhabra, Bernhard Clemm von Hohenberg, Michael Heseltine, and Magdalena Wojcieszak. Incentivizing news consumption on social media platforms using large language models and realistic bot accounts. *PNAS nexus*, 3(9):pgae368, 2024. 3.2

[20] Dennis Assenmacher, Lena Clever, Lena Frischlich, Thorsten Quandt, Heike Trautmann, and Christian Grimme. Demystifying social bots: On the intelligence of automated social media actors. *Social Media+ Society*, 6(3):2056305120939264, 2020. 1, 2.2.1, **??**, **??**, 3.3

[21] Aldo Averza, Khaled Slhoub, and Siddhartha Bhattacharyya. Evaluating the influence of twitter bots via agent-based social simulation. *IEEE Access*, 10:129394–129407, 2022. 6.2

[22] Marco Avvenuti, Stefano Cresci, Andrea Marchetti, Carlo Meletti, and Maurizio Tesconi. Ears (earthquake alert and report system) a real time decision support system for earthquake crisis management. In *Proceedings of the 20th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1749–1758, 2014. 3.5

[23] Matthew Babcock, David M Beskow, and Kathleen M Carley. Beaten up on twitter? exploring fake news and satirical responses during the black panther movie event. In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*, pages 97–103. Springer, 2018. 1.3.1

[24] Adam Badawy, Emilio Ferrara, and Kristina Lerman. Analyzing the digital traces of po-

litical manipulation: The 2016 russian interference twitter campaign. In *2018 IEEE/ACM international conference on advances in social networks analysis and mining (ASONAM)*, pages 258–265. IEEE, 2018. 3.5

[25] Christopher A Bail, Brian Guay, Emily Maloney, Aidan Combs, D Sunshine Hillygus, Friedolin Merhout, Deen Freelon, and Alexander Volfovsky. Assessing the russian internet research agency's impact on the political attitudes and behaviors of american twitter users in late 2017. *Proceedings of the national academy of sciences*, 117(1):243–250, 2020. 2.4

[26] Peng Bao, Hua-Wei Shen, Junming Huang, and Haiqiang Chen. Mention effect in information diffusion on a micro-blogging network. *PloS one*, 13(3):e0194192–e0194192, Mar 2018. ISSN 1932-6203. doi: 10.1371/journal.pone.0194192. URL `https://pubmed.ncbi.nlm.nih.gov/29558498`. 29558498[pmid]. 5.2.2

[27] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999. 5.2.1, 6.2, 6.6.1

[28] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset, 2020. 1.3.2

[29] Alex Bavelas. Communication patterns in task-oriented groups. *Journal of the Acoustical Society of America*, 22(6):725–730, 1950. 5.2.1

[30] Matthew C Benigni, Kenneth Joseph, and Kathleen M Carley. Online extremism and the communities that sustain it: Detecting the isis supporting community on twitter. *PloS one*, 12(12):e0181405, 2017. 3.5

[31] Matthew C Benigni, Kenneth Joseph, and Kathleen M Carley. Bot-ivistm: assessing information manipulation in social media using network analytics. In *Emerging research challenges and opportunities in computational social network analysis and mining*, pages 19–42. Springer, 2018. 3.3

[32] David M Beskow and Kathleen M Carley. Bot-hunter: a tiered approach to detecting & characterizing automated activity on twitter. In *Conference paper. SBP-BRiMS: International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*, volume 3, 2018. 2.2.1, 2.2.2, 2.4, 2.4.2, **??**, **??**, 4.4, 5.3.2

[33] David M Beskow and Kathleen M Carley. Agent based simulation of bot disinformation maneuvers in twitter. In *2019 Winter simulation conference (WSC)*, pages 750–761. IEEE, 2019. 6.2

[34] David M Beskow and Kathleen M Carley. You are known by your friends: Leveraging network metrics for bot detection in twitter. In *Open Source Intelligence and Cyber Crime: Social Media Analytics*, pages 53–88. Springer, 2020. 2.2.2

[35] Alessandro Bessi and Emilio Ferrara. Social bots distort the 2016 us presidential election online discussion. *First monday*, 21(11-7), 2016. **??**, **??**, 4.3, 5.2.1

[36] John M Betts and Ana-Maria Bliuc. The effect of influencers on societal polarization. In *2022 Winter Simulation Conference (WSC)*, pages 370–381. IEEE, 2022. 6.2

[37] Halil Bisgin, Nitin Agarwal, and Xiaowei Xu. A study of homophily on social media.

World Wide Web, 15(2):213–232, 2012. 5.3.1

[38] Sam Blazek. Scotch: A framework for rapidly assessing influence operations. *Atlantic Council*, 2021. 4.2

[39] Danah Boyd, Scott Golder, and Gilad Lotan. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *2010 43rd Hawaii international conference on system sciences*, pages 1–10. IEEE, 2010. 4.2, 5.3.2

[40] Florian Brachten, Milad Mirbabaie, Stefan Stieglitz, Olivia Berger, Sarah Bludau, and Kristina Schrickel. Threat or opportunity?-examining social bots in social media crisis communication. *arXiv preprint arXiv:1810.09159*, 2018. **??**, **??**

[41] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001. ISSN 0885-6125. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324. 2.4

[42] David A Broniatowski, Amelia M Jamison, SiHua Qi, Lulwah AlKulaib, Tao Chen, Adrian Benton, Sandra C Quinn, and Mark Dredze. Weaponized health communication: Twitter bots and russian trolls amplify the vaccine debate. *American journal of public health*, 108(10):1378–1384, 2018. 3.2, 4.2, 5.3.2

[43] Tom Buchanan. Why do people spread false information online? the effects of message and viewer characteristics on self-reported likelihood of sharing social media disinformation. *Plos one*, 15(10):e0239666, 2020. 6.2

[44] Zijian Cai, Zhaoxuan Tan, Zhenyu Lei, Zifeng Zhu, Hongrui Wang, Qinghua Zheng, and Minnan Luo. Lmbot: distilling graph knowledge into language model for graph-less deployment in twitter bot detection. In *Proceedings of the 17th ACM international conference on web search and data mining*, pages 57–66, 2024. 2.2.2

[45] Gian Maria Campedelli, Iain Cruickshank, and Kathleen M. Carley. A complex networks approach to find latent clusters of terrorist groups. *Applied Network Science*, 4(1):59, 8 2019. ISSN 2364-8228. doi: 10.1007/s41109-019-0184-6. URL https://doi.org/10.1007/s41109-019-0184-6. 5.4

[46] Riccardo Cantini, Fabrizio Marozzo, Domenico Talia, and Paolo Trunfio. Analyzing political polarization on social media by deleting bot spamming. *Big data and cognitive computing*, 6(1):3, 2022. 3.3

[47] Manuel Cargnino. The interplay of online network homogeneity, populist attitudes, and conspiratorial beliefs: Empirical evidence from a survey on german facebook users. *International Journal of Public Opinion Research*, 33(2):337–353, 2021. 4.2

[48] Kathleen M Carley. Social cybersecurity: an emerging science. *Computational and mathematical organization theory*, 26(4):365–381, 2020. (document), 3.4, 3.4, 3.4, 4.2, 4.2, 4.5, 4.5, 6.6.3, 7.5.3

[49] Kathleen M Carley, Kenneth Joseph, Mike Kowalchuck, Michael Lanham, and Geoffrey Morgan. Construct user guide. *Available at SSRN 2729263*, 2014. 6.4.1

[50] Peter Carragher, Lynnette Hui Xian Ng, and Kathleen M Carley. Simulation of stance perturbations. In *International Conference on Social Computing, Behavioral-Cultural*

Modeling and Prediction and Behavior Representation in Modeling and Simulation, pages 159–168. Springer, 2023. (document), 6.2, 6.4, 6.3, 6.4, 6.5

[51] Pete Cashmore. Twitter Zombies: 24 `https://mashable.com/archive/twitter-bots`, 2009. [Accessed 17-Jul-2023]. 1

[52] CCDH. The Disinformation Dozen — Center for Countering Digital Hate — CCDH — counterhate.com. `https://counterhate.com/research/the-disinformation-dozen/`, 2021. [Accessed 25-10-2023]. 1.3.4

[53] Damon Centola, Joshua Becker, Devon Brackbill, and Andrea Baronchelli. Experimental evidence for tipping points in social convention. *Science*, 360(6393):1116–1119, June 2018. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aas8827. 6.4.3

[54] Shelly Chaiken. Heuristic versus systematic information processing and the use of source versus message cues in persuasion. *Journal of personality and social psychology*, 39(5): 752, 1980. 4.7

[55] Ho-Chun Herbert Chang and Emilio Ferrara. Comparative analysis of social bots and humans during the covid-19 pandemic. *Journal of Computational Social Science*, 5(2): 1409–1425, 2022. 3.5

[56] Ho-Chun Herbert Chang, Emily Chen, Meiqing Zhang, Goran Muric, and Emilio Ferrara. Social bots and social media manipulation in 2020: The year in review. In *Handbook of Computational Social Science, Volume 1*, pages 304–323. Routledge, 2021. 3.5

[57] Nikan Chavoshi and Abdullah Mueen. Model bots, not humans on social media. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 178–185. IEEE, 2018. 1

[58] Nikan Chavoshi, Hossein Hamooni, and Abdullah Mueen. Debot: Twitter bot detection via warped correlation. In *Icdm*, volume 18, pages 28–65, 2016. 2.2.2, **??**, **??**

[59] Long Chen, Jianguo Chen, and Chunhe Xia. Social network behavior and public opinion manipulation. *Journal of Information Security and Applications*, 64:103060, 2022. 1, 3.5

[60] Siyan Chen and Saul Desiderio. A regression-based calibration method for agent-based models. *Computational Economics*, 59(2):687–700, 2022. 3.2

[61] Wen Chen, Diogo Pacheco, Kai-Cheng Yang, and Filippo Menczer. Neutral bots probe political bias on social media. *Nature communications*, 12(1):5580, 2021. 3.3

[62] Han Shi Jocelyn Chew. The use of artificial intelligence–based conversational agents (chatbots) for weight loss: scoping review and practical recommendations. *JMIR medical informatics*, 10(4):e32578, 2022. 3.2

[63] Eugene Ch'ng. Local interactions and the emergence of a twitter small-world network. *arXiv preprint arXiv:1508.03594*, 2015. 6.2

[64] Zi Chu, Steven Gianvecchio, Haining Wang, and Sushil Jajodia. Detecting automation of twitter accounts: Are you a human, bot, or cyborg? *IEEE Transactions on dependable and secure computing*, 9(6):811–824, 2012. 3.2

[65] Cybersecurity CISA and Infrastructure Security Agency. Social media bots.

https://www.cisa.gov/sites/default/files/publications/social_media_bots_infographic_set_508.pdf. [Accessed 28-10-2024]. **??**

[66] Cloudflare. What is a social media bot? — social media bot definition. https://www.cloudflare.com/learning/bots/what-is-a-social-media-bot/. [Accessed 28-10-2024]. **??**

[67] Andrew J Collins and Gayane Grigoryan. Abmscore: a heuristic algorithm for forming strategic coalitions in agent-based simulation. *Journal of Simulation*, 18(6):1033–1057, 2024. 3.2

[68] Stefano Cresci, Fabrizio Lillo, Daniele Regoli, Serena Tardelli, and Maurizio Tesconi. Cashtag piggybacking: Uncovering spam and bot activity in stock microblogs on twitter. *ACM Transactions on the Web (TWEB)*, 13(2):1–27, 2019. 2.4

[69] Andrew Crooks, Arie Croitoru, Anthony Stefanidis, and Jacek Radzikowski. # earthquake: Twitter as a distributed sensor system. *Transactions in GIS*, 17(1):124–147, 2013. 3.5

[70] Adya Danaditya, Lynnette Hui Xian Ng, and Kathleen M Carley. From curious hashtags to polarized effect: profiling coordinated actions in indonesian twitter discourse. *Social Network Analysis and Mining*, 12(1):105, 2022. (document), 1.3.1, 3.2, 3.4, 3.4, 3.5, 4.2, 4.2, 9.1

[71] Clayton Allen Davis, Onur Varol, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Botornot: A system to evaluate social bots. In *Proceedings of the 25th international conference companion on world wide web*, pages 273–274, 2016. 2.2.2, 3.2

[72] Michael Ann DeVito, Jeremy Birnholtz, and Jeffery T Hancock. Platforms, people, and perception: Using affordances to understand self-presentation on social media. In *Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing*, pages 740–754, 2017. 5.1

[73] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805. 5.4

[74] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 2.4

[75] US Department of Homeland Security DHS. niccs.cisa.gov. https://niccs.cisa.gov/sites/default/files/documents/pdf/ncsam_socialmediabotsoverview_508.pdf?trackDocs=ncsam_socialmediabotsoverview_508.pdf, 2018. [Accessed 29-10-2024]. **??**

[76] Frank Dignum. *Social simulation for a crisis*. Springer, 2021. 3.2

[77] Simone Driessen. Campaign problems: How fans react to taylor swift's controversial political awakening. *American behavioral scientist*, 66(8):1060–1074, 2022. 6.4.2

[78] Zening Duan, Jianing Li, Josephine Lukito, Kai-Cheng Yang, Fan Chen, Dhavan V Shah, and Sijia Yang. Algorithmic agents in the hybrid media system: Social bots, selective amplification, and partisan news about covid-19. *Human Communication Research*, 48 (3):516–542, 2022. 3.5

[79] Tamar Edry, Nason Maani, Martin Sykora, Suzanne Elayan, Yulin Hswen, Markus Wolf, Fabio Rinaldi, Sandro Galea, and Oliver Gruebner. Real-time geospatial surveillance of localized emotional stress responses to covid-19: a proof of concept analysis. *Health & Place*, 70:102598, 2021. 4.2

[80] Zineb Ellaky, Faouzia Benabbou, and Sara Ouahabi. Systematic literature review of social media bots detection systems. *Journal of King Saud University-Computer and Information Sciences*, 35(5):101551, 2023. **??**, **??**

[81] Tuğrulcan Elmas. The role of compromised accounts in social media manipulation. 2022. PhD Thesis. 3.4

[82] Tuğrulcan Elmas, Rebekah Overdorf, and Karl Aberer. Characterizing retweet bots: The case of black market accounts. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 16, pages 171–182, 2022. 3.2

[83] Ela Elsholz, Jon Chamberlain, and Udo Kruschwitz. Exploring language style in chatbots to increase perceived product value and user engagement. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*, pages 301–305, 2019. 3.5

[84] Aviad Elyashar, Michael Fire, Dima Kagan, and Yuval Elovici. Guided socialbots: Infiltrating the social networks of specific organizations' employees. *Ai Communications*, 29 (1):87–106, 2014. 3.5

[85] Jonathan St BT Evans and Keith E Stanovich. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3):223–241, 2013. 4.7

[86] Shangbin Feng, Herun Wan, Ningnan Wang, Jundong Li, and Minnan Luo. Twibot-20: A comprehensive twitter bot detection benchmark. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pages 4485–4494, 2021. 2.2.2

[87] Shangbin Feng, Zhaoxuan Tan, Herun Wan, Ningnan Wang, Zilong Chen, Binchi Zhang, Qinghua Zheng, Wenqian Zhang, Zhenyu Lei, Shujie Yang, et al. Twibot-22: Towards graph-based twitter bot detection. *Advances in Neural Information Processing Systems*, 35:35254–35269, 2022. 2.2.2

[88] Shangbin Feng, Herun Wan, Ningnan Wang, Zhaoxuan Tan, Minnan Luo, and Yulia Tsvetkov. What does the bot say? opportunities and risks of large language models in social media bot detection. *arXiv preprint arXiv:2402.00371*, 2024. 2.2.2

[89] Emilio Ferrara. # covid-19 on twitter: Bots, conspiracies, and social media activism. *arXiv preprint arXiv: 2004.09531*, 2020. 3.5

[90] Emilio Ferrara, Onur Varol, Clayton Davis, Filippo Menczer, and Alessandro Flammini. The rise of social bots. *Communications of the ACM*, 59(7):96–104, 2016. **??**, **??**

[91] Emilio Ferrara, Herbert Chang, Emily Chen, Goran Muric, and Jaimin Patel. Character-

izing social media manipulation in the 2020 us presidential election. *First Monday*, 2020. 3.3, 3.5

[92] Jenny Rose Finkel, Trond Grenager, and Christopher D Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting of the association for computational linguistics (ACL'05)*, pages 363–370, 2005. 4.3

[93] Deen Freelon, Michael Bossetta, Chris Wells, Josephine Lukito, Yiping Xia, and Kirsten Adams. Black trolls matter: Racial and ideological asymmetries in social media disinformation. *Social Science Computer Review*, 40(3):560–578, 2022. 3.5

[94] Linton C. Freeman. Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215–239, 1979. 5.2.1

[95] Noah E Friedkin and Eugene C Johnsen. Social influence and opinions. *Journal of Mathematical Sociology*, 15(3-4):193–206, 1990. Publisher: Taylor & Francis. 5.5, 6.4.1

[96] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232, 2001. 2.4

[97] Wai-Tat Fu and Q Vera Liao. Information and attitude diffusion in networks. In *International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction*, pages 205–213. Springer, 2012. 6.2

[98] Pascale Fung, Yoram Bachrach, Asli Celikyilmaz, Kamalika Chaudhuri, Delong Chen, Willy Chung, Emmanuel Dupoux, Hongyu Gong, Hervé Jégou, Alessandro Lazaric, et al. Embodied ai agents: Modeling the world. *arXiv preprint arXiv:2506.22355*, 2025. 3.3, 3.2

[99] Cindy Gallois, Tania Ogay, and Howard Giles. Communication accommodation theory. In *Theorizing about Intercultural Communication*, pages 121–148. 2005. 4.7

[100] Margherita Gambini, Serena Tardelli, and Maurizio Tesconi. The anatomy of conspiracy theorists: unveiling traits using a comprehensive twitter dataset. *Computer Communications*, 217:25–40, 2024. 3.2

[101] Venus Garg. Designing the mind: How agentic frameworks are shaping the future of ai behavior. *Journal of Computer Science and Technology Studies*, 7(5):182–193, 2025. 3.2

[102] Anastasia Giachanou, Xiuzhen Zhang, Alberto Barrón-Cedeño, Olessia Koltsova, and Paolo Rosso. Online information disorder: fake news, bots and trolls. *International Journal of Data Science and Analytics*, 13(4):265–269, 2022. 6.2

[103] Fabio Giglietto, Nicola Righetti, Luca Rossi, and Giada Marino. It takes a village to manipulate the media: coordinated link sharing behavior during 2018 and 2019 italian elections. *Information, Communication & Society*, 23(6):867–891, 2020. 5.2.2

[104] Zafar Gilani, Reza Farahbakhsh, Gareth Tyson, Liang Wang, and Jon Crowcroft. Of bots and humans (on twitter). In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017*, pages 349–354, 2017. 3.3

[105] Nigel Gilbert. *Agent-based models*. Sage Publications, 2019. 3.3

[106] Maria Glenski, Svitlana Volkova, and Srijan Kumar. User engagement with digital deception. *Disinformation, Misinformation, and Fake News in Social Media: Emerging Research Challenges and Opportunities*, pages 39–61, 2020. 3.5

[107] Ted Goertzel. Belief in conspiracy theories. *Political psychology*, pages 731–742, 1994. 6.2, 6.3

[108] Robert Gorwa and Douglas Guilbeault. Unpacking the social media bot: A typology to guide research and policy. *Policy & Internet*, 12(2):225–248, 2020. **??**, **??**, 3.2, 3.4, 3.3, 3.5

[109] Timothy Graham, Sam Hames, and Elizabeth Alpert. The coordination network toolkit: a framework for detecting and analysing coordinated behaviour on social media. *Journal of Computational Social Science*, 7(2):1139–1160, 2024. 5.2.2

[110] Christian Grimme, Mike Preuss, Lena Adam, and Heike Trautmann. Social bots: Human-like by means of human control? *Big data*, 5(4):279–293, 2017. 3.3, 1

[111] Christian Grimme, Dennis Assenmacher, and Lena Adam. Changing perspectives: Is it sufficient to detect social bots? In *Social Computing and Social Media. User Experience and Behavior: 10th International Conference, SCSM 2018, Held as Part of HCI International 2018, Las Vegas, NV, USA, July 15–20, 2018, Proceedings, Part I*, pages 445–461. Springer, 2018. 3.3, 4.7

[112] Nir Grinberg, Mor Naaman, Blake Shaw, and Gilad Lotan. Extracting diurnal patterns of real world activity from social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, pages 205–214, 2013. 4.1

[113] Andrew Guess, Jonathan Nagler, and Joshua Tucker. Less than you think: Prevalence and predictors of fake news dissemination on facebook. *Science advances*, 5(1):eaau4586, 2019. 6.2

[114] Jing Guo and Hsuan-Ting Chen. How does multi-platform social media use lead to biased news engagement? examining the role of counter-attitudinal incidental exposure, cognitive elaboration, and network homogeneity. *Social Media & Society*, 8(4): 20563051221129140, 2022. 4.2

[115] Nick Hajli, Usman Saeed, Mina Tajvidi, and Farid Shirazi. Social bots and the spread of disinformation in social media: the challenges of artificial intelligence. *British Journal of Management*, 33(3):1238–1253, 2022. 3.3

[116] Kevin A Hallgren. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23, 2012. 2.4

[117] Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. Cross-domain label-adaptive stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9011–9028, 2021. 2.4

[118] Mallory J. Harris, Ryan Murtfeldt, Shufan Wang, Erin A. Mordecai, and Jevin D. West. Perceived experts are prevalent and influential within an antivaccine community on twitter. *PNAS Nexus*, 3(2):pgae007, 2024. 4.2

[119] William Hart, Dolores Albarracín, Alice H Eagly, Inge Brechan, Matthew J Lindberg, and

Lisa Merrill. Feeling validated versus being correct: a meta-analysis of selective exposure to information. *Psychological bulletin*, 135(4):555, 2009. 6.2

[120] A. Haupt and W. Camber. Conflict prediction through geo-spatial interpolation of radicalization in syrian social media. *Tradoc Analysis Center*, 2015. 4.3

[121] Stefanie Haustein, Timothy D Bowman, Kim Holmberg, Andrew Tsou, Cassidy R Sugimoto, and Vincent Larivière. Tweets as impact indicators: Examining the implications of automated "bot" accounts on t witter. *Journal of the Association for Information Science and Technology*, 67(1):232–238, 2016. 3.5

[122] Kadhim Hayawi, Susmita Saha, Mohammad Mehedy Masud, Sujith Samuel Mathew, and Mohammed Kaosar. Social media bot detection with deep learning methods: a systematic review. *Neural Computing and Applications*, 35(12):8903–8918, 2023. 2.2.2, **??**, **??**

[123] Qinglai He, Yili Hong, and TS Raghu. Platform governance with algorithm-based content moderation: An empirical study on reddit. *Information Systems Research*, 2024. 3.5

[124] Runzhi He, Hao He, Yuxia Zhang, and Minghui Zhou. Automating dependency updates in practice: An exploratory study on github dependabot. *IEEE Transactions on Software Engineering*, 49(8):4004–4022, 2023. 5.3.1

[125] L. Herman. For who page? tiktok creators' algorithmic dependencies. In D. De Sainz Molestina, L. Galluzzo, F. Rizzo, and D. Spallazzo, editors, *IASDR 2023: Life-Changing Design*, Milan, Italy, October 2023. doi: 10.21606/iasdr.2023.576. URL `https://doi.org/10.21606/iasdr.2023.576`. October 9–13. 4.2

[126] Matthew Hicks. *AI-Enabled Social Cyber Maneuver Detection and Creation*. PhD thesis, Carnegie Mellon University, 2024. 6.6.1

[127] Jennifer Hill, W Randolph Ford, and Ingrid G Farreras. Real conversations with artificial intelligence: A comparison between human–human online conversations and human–chatbot conversations. *Computers in human behavior*, 49:245–250, 2015. 3.4

[128] McKenzie Himelein-Wachowiak, Salvatore Giorgi, Amanda Devoto, Muhammad Rahman, Lyle Ungar, H Andrew Schwartz, David H Epstein, Lorenzo Leggio, and Brenda Curtis. Bots and misinformation spread on social media: Implications for covid-19. *Journal of medical Internet research*, 23(5):e26933, 2021. 3.5

[129] Lennart Hofeditz, Christian Ehnis, Deborah Bunker, Florian Brachten, and Stefan Stieglitz. Meaningful use of social bots? possible applications in crisis communication during disasters. 2019. 2.2.1, 3.2, 3.5

[130] Hyehyun Hong and Hyun Jee Oh. Utilizing bots for sustainable news business: Understanding users' perspectives of news bots in the age of social media. *Sustainability*, 12 (16):6515, 2020. 3.3, 3.5

[131] Christy Galletta Horner, Dennis Galletta, Jennifer Crawford, and Abhijeet Shirsat. Emotions: The unexplored fuel of fake news on social media. In *Fake news on the internet*, pages 147–174. Routledge, 2023. 6.6.3

[132] Xu Huang, Jianxun Lian, Yuxuan Lei, Jing Yao, Defu Lian, and Xing Xie. Recommender ai agent: Integrating large language models for interactive recommendations. *ACM Trans-*

actions on Information Systems, 43(4):1–33, 2025. 3.2

[133] Yu Huang. Levels of ai agents: from rules to large language models. *arXiv preprint arXiv:2405.06643*, 2024. 3.2

[134] Sofia Hurtado, Poushali Ray, and Radu Marculescu. Bot detection in reddit political discussion. In *Proceedings of the fourth international workshop on social sensing*, pages 30–35, 2019. 2.2.2

[135] Imperva. What are Bots — Bot Types & Mitigation Techniques — Imperva — imperva.com. `https://www.imperva.com/learn/application-security/what-are-bots/`. [Accessed 29-10-2024]. **??**

[136] Gerardo Iñiguez, Sara Heydari, János Kertész, and Jari Saramäki. Universal patterns in egocentric communication networks. *Nature Communications*, 14(1):5217, 2023. 5.2.1

[137] Charity S Jacobs, Lynnette Hui Xian Ng, and Kathleen M Carley. Tracking china's cross-strait bot networks against taiwan. In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*, pages 115–125. Springer, 2023. (document), 1, 2.4, 3.2, 3.3, 3.4, 3.5, 4.2, 5.2, 5.4, 5.4, 9.1

[138] Charity S Jacobs, Lynnette Hui Xian Ng, and Kathleen M Carley. Det: Detection evasion techniques of state-sponsored accounts. Technical report, Center for Open Science, 2024. 3.4

[139] Shagun Jhaver, Iris Birman, Eric Gilbert, and Amy Bruckman. Human-machine collaboration for content regulation: The case of reddit automoderator. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 26(5):1–35, 2019. 3.5

[140] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with gpus. *arXiv preprint arXiv:1702.08734*, 2017. 5.4

[141] Nathaniel Johnston and Dave Greene. *Conway's Game of Life: Mathematics and Construction*. Nathaniel Johnston, 2022. 3.2

[142] Ewa Kacewicz, James W Pennebaker, Matthew Davis, Moongee Jeon, and Arthur C Graesser. Pronoun use reflects standings in social hierarchies. *Journal of Language and Social Psychology*, 33(2):125–143, 2014. 4.6

[143] Brian L Keeley. Journal of philosophy, inc. *The Journal of Philosophy*, 96(3):109–126, 1999. 6.2

[144] Ryan Kenny, Baruch Fischhoff, Alex Davis, Kathleen M Carley, and Casey Canfield. Duped by bots: why some are better than others at detecting fake social media personas. *Human factors*, 66(1):88–102, 2024. **??**, **??**

[145] Tuja Khaund, Baris Kirdemir, Nitin Agarwal, Huan Liu, and Fred Morstatter. Social bots and their coordination during online campaigns: a survey. *IEEE Transactions on Computational Social Systems*, 9(2):530–545, 2021. 3.5, 5.2.2

[146] Catherine King, Daniele Bellutta, and Kathleen M Carley. Lying about lying on social media: a case study of the 2019 canadian elections. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation*

in Modeling and Simulation, pages 75–85. Springer, 2020. 1.3.1, 2.2.2

[147] Robert H. Knapp. A psychology of rumor. *Public Opinion Quarterly*, 8(1):22–37, 1944. 4.7

[148] Maxim Kolomeets and Andrey Chechulin. Analysis of the malicious bots market. In *2021 29th conference of open innovations association (FRUCT)*, pages 199–205. IEEE, 2021. 3.3

[149] Maxim Kolomeets and Andrey Chechulin. Social bot metrics. *Social Network Analysis and Mining*, 13(1):36, 2023. 3.3

[150] Maxim Kolomeets, Olga Tushkanova, Vasily Desnitsky, Lidia Vitkova, and Andrey Chechulin. Experimental evaluation: Can humans recognise social media bots? *Big Data and Cognitive Computing*, 8(3):24, 2024. 1

[151] Joel Krueger and Lucy Osler. Communing with the dead online: chatbots, grief, and continuing bonds. *Journal of Consciousness Studies*, 29(9-10):222–252, 2022. 2.2.1, 3.2, 3.5

[152] Aytalina Kulichkina, Nicola Righetti, and Annie Waldherr. Protest and repression on social media: Pro-navalny and pro-government mobilization dynamics and coordination patterns on russian twitter. *new media & society*, 27(9):5433–5454, 2025. 7.2

[153] Sumeet Kumar. *Social Media Analytics for Stance Mining A Multi-Modal Approach with Weak Supervision.* PhD thesis, Carnegie Mellon University, USA, 2020. 4.5, 7.5.1

[154] Adrienne LaFrance. The Internet Is Mostly Bots — theatlantic.com. `https://www.theatlantic.com/technology/archive/2017/01/bots-bots-bots/515043/`, 2017. [Accessed 16-Jul-2023]. 1

[155] Jiyoung Lee and Kim Bissell. Correcting vaccine misinformation on social media: the inadvertent effects of repeating misinformation within such corrections on COVID-19 vaccine misperceptions. *Current Psychology*, pages 1–13, 2024. 4.2

[156] Jure Leskovec, Lars Backstrom, Ravi Kumar, and Andrew Tomkins. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 462–470, 2008. 5.2.1

[157] Stephan Lewandowsky and Sander Van Der Linden. Countering misinformation and fake news through inoculation and prebunking. *European review of social psychology*, 32(2): 348–384, 2021. 6.3

[158] Linda Li, Orsolya Vasarhelyi, and Balazs Vedres. Social bots sour activist sentiment without eroding engagement, 2024. 4.2

[159] Hause Lin, Marlyn T Savio, Xieyining Huang, Miriah Steiger, Rachel Guevara, Dali Szostak, Gordon Pennycook, and David G Rand. Accuracy prompts protect professional content moderators from the illusory truth effect, March 2024. URL `https://osf.io/preprints/psyarxiv/gswm6`. 4.2

[160] Xialing Lin, Patric R. Spence, and Kenneth A. Lachlan. Social media and credibility indicators: The effect of influence cues. *Computers in Human Behavior*, 63:264–271, 2016. 4.2

[161] Darren L Linvill and Patrick L Warren. Engaging with others: How the ira coordinated information operation made friends. *Harvard Kennedy School Misinformation Review*, 1 (2), 2020. 7.2

[162] Xia Liu, Alvin C Burns, and Yingjian Hou. An investigation of brand-related user-generated content on twitter. *Journal of Advertising*, 46(2):236–247, 2017. 3.5

[163] Tetyana Lokot and Nicholas Diakopoulos. News bots: Automating news and information dissemination on twitter. *Digital journalism*, 4(6):682–699, 2016. 3.2, 3.5

[164] Salvador Lopez-Joya, Jose A Diaz-Garcia, M Dolores Ruiz, and Maria J Martin-Bautista. Dissecting a social bot powered by generative ai: anatomy, new trends and challenges. *Social Network Analysis and Mining*, 15(1):7, 2025. 3.3

[165] Gilad Lotan, Erhardt Graeff, Mike Ananny, Devin Gaffney, Ian Pearce, et al. The arab spring— the revolutions were tweeted: Information flows during the 2011 tunisian and egyptian revolutions. *International journal of communication*, 5:31, 2011. 3.5

[166] Hsiu-Chi Lu and Hsuan-wei Lee. Agents of discord: Modeling the impact of political bots on opinion polarization in social networks. *Social Science Computer Review*, page 08944393241270382, 2024. 3.2, 3.5

[167] Luca Luceri, Felipe Cardoso, and Silvia Giordano. Down the bot hole: actionable insights from a 1-year analysis of bots activity on twitter. *arXiv preprint arXiv:2010.15820*, 2020. 6.2

[168] Josephine Lukito. Coordinating a multi-platform disinformation campaign: Internet research agency activity on three us social media platforms, 2015 to 2017. *Political Communication*, 37(2):238–255, 2020. 7.2

[169] Jianhong Luo and Chaoqi Jin. Fusing content and social relationships: a multi-modal heterogeneous graph transformer approach for social bot detection. *EPJ Data Science*, 14 (1):68, 2025. 2.2.2

[170] Kim Lyons. Parler posts, some with GPS data, have been archived by an independent researcher — theverge.com. https://www.theverge.com/2021/1/11/22224689/parler-posts-gps-data-archived-independent-researchers-amazon-apple-go 2021. [Accessed 17-12-2025]. 1.3.5

[171] Ming-Cheng Ma and John P Lalor. An empirical analysis of human-bot interaction on reddit. In *Proceedings of the Sixth Workshop on Noisy User-generated Text (w-NUT 2020)*, pages 101–106, 2020. 2.2.2

[172] Michael W Macy, James A Kitts, Andreas Flache, and Steve Benard. Polarization in dynamic networks: A Hopfield model of emergent structure. *Dynamic Social Network Modeling and Analysis*, pages 162–173, 2003. Washington DC: National Academies Press. 6.4.1

[173] Thomas Magelinski, David Beskow, and Kathleen M Carley. Graph-hist: Graph classification from latent feature histograms with application to bot detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5134–5141, 2020. 2.2.2

[174] Thomas Magelinski, Lynnette Ng, and Kathleen Carley. A synchronized action framework for detection of coordination on social media. *Journal of Online Trust and Safety*, 1(2), 2022. 5.2.2

[175] Thomas Magelinski, Lynnette Ng, and Kathleen Carley. A synchronized action framework for detection of coordination on social media. *Journal of Online Trust and Safety*, 1(2), 2022. 1.3.1, 3.5, 5.2, 5.4

[176] Markus Maier, Matthias Hein, and Ulrike von Luxburg. Optimal construction of k-nearest neighbor graphs for identifying noisy clusters. *arXiv e-prints*, art. arXiv:0912.3408, 12 2009. 5.4

[177] Markus Maier, Ulrike von Luxburg, and Matthias Hein. How the result of graph clustering methods depends on the construction of the graph. *arXiv e-prints*, art. arXiv:1102.2075, 2 2011. 5.4

[178] Peter Mantello, Tung Manh Ho, and Lena Podoletz. Automating extremism: Mapping the affective roles of artificial agents in online radicalization. In *The Palgrave handbook of malicious use of AI and psychological security*, pages 81–103. Springer, 2023. 3.5

[179] Rebecca Marigliano, Lynnette Hui Xian Ng, and Kathleen M Carley. Analyzing digital propaganda and conflict rhetoric: a study on russia's bot-driven campaigns and counter-narratives during the ukraine crisis. *Social Network Analysis and Mining*, 14(1):170, 2024. (document), 1.3.1, 4.9, 4.10, 7.2, 7.5.1

[180] Franziska Martini, Paul Samula, Tobias R Keller, and Ulrike Klinger. Bot, or not? comparing three methods for detecting social bots in five political discourses. *Big data & society*, 8(2):20539517211033566, 2021. **??**, **??**

[181] Alice E. Marwick and Danah Boyd. I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society*, 13(1):114–133, 2011. 4.7, 4.7

[182] Michele Mazza, Stefano Cresci, Marco Avvenuti, Walter Quattrociocchi, and Maurizio Tesconi. Rtbust: Exploiting temporal patterns for botnet detection on twitter. In *Proceedings of the 10th ACM conference on web science*, pages 183–192, 2019. 2.2.2, 2.4, 4.2

[183] Michele Mazza, Marco Avvenuti, Stefano Cresci, and Maurizio Tesconi. Investigating the difference between trolls, social bots, and humans on twitter. *Computer Communications*, 196:23–36, 2022. 3.3

[184] Megan K McBride, Zack Gold, and Kasey Stricklin. Social media bots: Implications for special operations forces. *Center for Naval Analysis, September*, 2020. 3.2

[185] Fenwick McKelvey and Elizabeth Dubois. Computational propaganda in canada: The use of political bots. 2017. 3.5

[186] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, 27(1):415–444, 2001. 5.3.2

[187] Miriam J. Metzger and Andrew J. Flanagin. Credibility and trust of information in online environments: The use of cognitive heuristics. *Journal of Pragmatics*, 59:210–220, 2013.

4.2

[188] Microsoft. How to spot bots on social media – Microsoft 365 — microsoft.com. https://www.microsoft.com/en-us/microsoft-365-life-hacks/privacy-and-safety/how-to-spot-bots-on-social-media, 2024. [Accessed 29-10-2024]. **??**

[189] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594): 824–827, 2002. 5.2.1

[190] Jae Min Jung, Kawpong Polyorat, and James J. Kellaris. A cultural paradox in authority-based advertising. *International Marketing Review*, 26(6):601–632, 2009. 4.7

[191] Tianjun Mo, Ziyi Jiang, and Qichang Zheng. Interactive ai agent for code refactoring assistance: A study on decision-making strategies and human-agent collaboration effectiveness. *Academia Nexus Journal*, 4(1), 2025. 3.2

[192] Sachin Modgil, Rohit Kumar Singh, Shivam Gupta, and Denis Dennehy. A confirmation bias view on social media induced polarisation during COVID-19. *Information Systems Frontiers*, pages 1–25, 2021. 4.2

[193] Patricia L Moravec, Randall K Minas, and Alan Dennis. Fake news on social media: People believe what they want to believe when it makes no sense at all. *MIS Quarterly*, 43(4):1343–1360, 2019. 6.2

[194] Jacob L. Moreno. *Who Shall Survive? A New Approach to the Problem of Human Interrelations*. Nervous and Mental Disease Publishing Co., Washington, DC, 1934. 5.2.1

[195] Mary S Morgan. Model experiments, virtual experiments, and virtually experiments. *The philosophy of scientific experimentation*, page 216, 2003. 3.2

[196] Emi Moriuchi, V Myles Landers, Deborah Colton, and Neil Hair. Engagement with chatbots versus augmented reality interactive technology in e-commerce. *Journal of Strategic Marketing*, 29(5):375–389, 2021. 3.5

[197] Xinyi Mou, Xuanwen Ding, Qi He, Liang Wang, Jingcong Liang, Xinnong Zhang, Libo Sun, Jiayu Lin, Jie Zhou, Xuanjing Huang, et al. From individual to society: A survey on social simulation driven by large language model-based agents. *arXiv preprint arXiv:2412.03563*, 2024. 6.2

[198] Dhiraj Murthy, Alison B Powell, Ramine Tinati, Nick Anstead, Les Carr, Susan Halford, and Mark Weal. Bots and political influence: A sociotechnical investigation of social network capital. *International journal of communication*, 10:4952–4971, 2016. 3.3

[199] Arvind Narayanan. Understanding social media recommendation algorithms. 2023. 3.4

[200] Martin N Ndlela. Social media algorithms, bots and elections in africa. In *Social media and elections in Africa, Volume 1: Theoretical perspectives and election campaigns*, pages 13–37. Springer, 2020. 1

[201] Gina Neff. Talking to bots: Symbiotic agency and the case of tay. *International Journal of Communication*, 2016. 3.5

[202] Lynnette Hui Xian Ng and Kathleen M Carley. Bot-based emotion behavior differences in images during kashmir black day event. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 184–194. Springer, 2021. 3.4, 4.2, 9.1, 9.2

[203] Lynnette Hui Xian Ng and Kathleen M Carley. Online coordination: methods and comparative case studies of coordinated groups across four events in the united states. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 12–21, 2022. 3.4, 5.2, 5.4, 5.4, 6.6.3, 9.1, 9.2

[204] Lynnette Hui Xian Ng and Kathleen M Carley. Pro or anti? a social influence model of online stance flipping. *IEEE Transactions on Network Science and Engineering*, 10(1): 3–19, 2022. (document), 3.2, 3.5, 4.6, 4.6, 4.2, 4.7, 5.3.2, 5.4, 5.11, 5.5, 6.6.2, 9.1, 9.2

[205] Lynnette Hui Xian Ng and Kathleen M Carley. Botbuster: Multi-platform bot detection using a mixture of experts. In *Proceedings of the international AAAI conference on web and social media*, volume 17, pages 686–697, 2023. (document), 1.3.3, 2.4, 2.2, 2.5, **??**, 3.2, 9.1, 9.2

[206] Lynnette Hui Xian Ng and Kathleen M Carley. A combined synchronization index for evaluating collective action social media. *Applied network science*, 8(1):1, 2023. (document), 1.3.1, 1.3.1, 5.3.2, 5.9, 5.3, 9.1, 9.2

[207] Lynnette Hui Xian Ng and Kathleen M Carley. Deflating the chinese balloon: types of twitter bots in us-china balloon incident. *EPJ Data Science*, 12(1):63, 2023. (document), 1.3.1, 3.3, 3.4, 3.4, 3.5, 3.4, 3.6, 3.5, 4.3, 4.11, 4.12, 9.1, 9.2

[208] Lynnette Hui Xian Ng and Kathleen M Carley. Popping the hood on chinese balloons: Examining the discourse between us and china-geotagged accounts. *First Monday*, 2023. (document), 4.6, 4.7, 9.1

[209] Lynnette Hui Xian Ng and Kathleen M Carley. Do you hear the people sing? comparison of synchronized url and narrative themes in 2020 and 2023 french protests. *Frontiers in big Data*, 6:1221744, 2023. 1.3.1, 3.5, 9.2

[210] Lynnette Hui Xian Ng and Kathleen M Carley. Assembling a multi-platform ensemble social bot detector with applications to us 2020 elections. *Social Network Analysis and Mining*, 14(1):45, 2024. (document), 2.3, 2.4, **??**, 2.6, 3.5, 9.1, 9.2

[211] Lynnette Hui Xian Ng and Kathleen M Carley. The dual personas of social media bots. *arXiv preprint arXiv:2504.12498*, 2025. 9.1

[212] Lynnette Hui Xian Ng and Kathleen M Carley. A global comparison of social media bot and human characteristics. *Scientific Reports*, 15(1):10973, 2025. (document), 1, 2.2.2, 2.3, 2.1, 2.1, 2.2, 2.3, 3.3, 4.13, 5.1, 5.1, 6.4, 6.6.2, 9.1, 9.2

[213] Lynnette Hui Xian Ng and Kathleen M Carley. Are llm-powered social media bots realistic? In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 14–23. Springer, 2025. (document), 6.4, 6.5, 6.12, 9.1, 9.2

[214] Lynnette Hui Xian Ng and Kathleen M Carley. Social cyber geographical worldwide

inventory of bots. *arXiv preprint arXiv:2501.18839*, 2025. 2.4, **??**, 4.3, 6.2, 9.1

[215] Lynnette Hui Xian Ng and Kathleen M. Carley. Botsim: Mitigating the formation of conspiratorial societies with useful bots. *Journal of Artificial Societies and Social Simulation*, 29(1):4, 2026. ISSN 1460-7425. doi: 10.18564/jasss.5881. URL `http://jasss.soc.surrey.ac.uk/29/1/4.html`. 9.2

[216] Lynnette Hui Xian Ng, Iain J Cruickshank, and Kathleen M Carley. Cross-platform information spread during the january 6th capitol riots. *Social Network Analysis and Mining*, 12(1):133, 2022. 3.4, 5.2, 9.1, 9.2

[217] Lynnette Hui Xian Ng, JD Moffitt, and Kathleen M Carley. Coordinated through aweb of images: Analysis of image-based influence operations from china, iran, russia, and venezuela. *arXiv preprint arXiv:2206.03576*, 2022. 1, 3.4, 5.2, 9.1

[218] Lynnette Hui Xian Ng, Dawn C Robertson, and Kathleen M Carley. Stabilizing a supervised bot detection algorithm: How much data is needed for consistent predictions? *Online Social Networks and Media*, 28:100198, 2022. (document), 1.3.1, **??**, **??**, 2.7, 2.8, 9.1, 9.2

[219] Lynnette Hui Xian Ng, Iain J Cruickshank, and Kathleen M Carley. Coordinating narratives framework for cross-platform analysis in the 2021 us capitol riots. *Computational and Mathematical Organization Theory*, 29(3):470–486, 2023. 1, 5.2.2, 5.2, 5.4, 9.2

[220] Lynnette Hui Xian Ng, Mihovil Bartulovic, and Kathleen M Carley. Tiny-botbuster: Identifying automated political coordination in digital campaigns. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 25–34. Springer, 2024. (document), 2.4, 2.4, **??**, 3.5, 6.6.2, 9.1, 9.2

[221] Lynnette Hui Xian Ng, Ian Kloo, Samantha Clark, and Kathleen M Carley. An exploratory analysis of covid bot vs human disinformation dissemination stemming from the disinformation dozen on telegram. *Journal of Computational Social Science*, 7(1):695–720, 2024. (document), 1.1, 2.4, **??**, 4.14, 5.3.1, 9.1, 9.2

[222] Lynnette Hui Xian Ng, Dawn C Robertson, and Kathleen M Carley. Cyborgs for strategic communication on social media. *Big Data & Society*, 11(1):20539517241231275, 2024. 3.4, 3.5, 9.1, 9.2

[223] Lynnette Hui Xian Ng, Wenqi Zhou, and Kathleen M Carley. Exploring cognitive bias triggers in covid-19 misinformation tweets: A bot vs. human perspective. *arXiv preprint arXiv:2406.07293*, 2024. 9.1

[224] Lynnette Hui Xian Ng, Iain J Cruickshank, and David Farr. Building bridges between users and content across multiple platforms during natural disasters. In *Proceedings of the 17th ACM Web Science Conference 2025*, pages 499–503, 2025. 3.4

[225] Lynnette Hui Xian Ng, Bianca NY Kang, and Kathleen M Carley. Aurasight: Generating realistic social media data. *arXiv preprint arXiv:2509.08927*, 2025. 6.6, 9.1, 9.2

[226] Lynnette Hui Xian Ng, Divyaansh Sinha, and Kathleen M Carley. Star network motifs on x during covid-19. In *International Conference on Social Computing, Behavioral-Cultural*

Modeling and Prediction and Behavior Representation in Modeling and Simulation, pages 193–202. Springer, 2025. (document), 5.3.1, 5.2, 9.1, 9.2

[227] Lynnette Hui Xian Ng, Wenqi Zhou, and Kathleen M. Carley. Appeal and Scope of Misinformation Spread by AI Agents and Humans. In *AMCIS 2025 Proceedings*, number 6 in Social Computing (SOCCOMP). Americas Conference on Information Systems, 2025. URL https://aisel.aisnet.org/amcis2025/social_comput/social_comput/6/. 9.1, 9.2

[228] Liangbo Ning, Ziran Liang, Zhuohang Jiang, Haohao Qu, Yujuan Ding, Wenqi Fan, Xiaoyong Wei, Shanru Lin, Hui Liu, Philip S Yu, et al. A survey of webagents: Towards next-generation ai agents for web automation with large foundation models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pages 6140–6150, 2025. 3.3, 3.2

[229] NOW. News on the web corpora. URL https://www.english-corpora.org/now/. 3.4

[230] NPR. One of the most influential voices in vaccine misinformation is a doctor, 2021. URL https://www.npr.org/2021/08/08/1025845675/one-of-the-most-influential-voices-in-vaccine-misinformation-is-a-doct [Accessed 2024-02-11]. 4.2

[231] Richard J Oentaryo, Arinto Murdopo, Philips K Prasetyo, and Ee-Peng Lim. On profiling bots in social media. In *Social Informatics: 8th International Conference, SocInfo 2016, Bellevue, WA, USA, November 11-14, 2016, Proceedings, Part I 8*, pages 92–109. Springer, 2016. 3.2, 3.4, 3.3

[232] Ian O'Hara. Automated epistemology: Bots, computational propaganda & information literacy instruction. *The Journal of Academic Librarianship*, 48(4):102540, 2022. 3.5

[233] Carl Öhman, Robert Gorwa, and Luciano Floridi. Prayer-bots and religious worship on twitter: A call for a wider research agenda. *Minds and machines*, 29(2):331–338, 2019. 3.5

[234] Mariam Orabi, Djedjiga Mouheb, Zaher Al Aghbari, and Ibrahim Kamel. Detection of bots in social media: a systematic review. *Information Processing & Management*, 57(4): 102250, 2020. **??**, **??**

[235] Joseph T Ornstein and Ross A Hammond. Agent-based modeling in the social sciences. *New Horizons in Modeling and Simulation for Social Epidemiology and Public Health*, page 22, 2021. 3.2

[236] Diogo Pacheco, Alessandro Flammini, and Filippo Menczer. Unveiling coordinated groups behind white helmets disinformation. In *Companion Proceedings of the Web Conference 2020*, WWW '20, page 611–616, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370240. doi: 10.1145/3366424.3385775. URL https://doi.org/10.1145/3366424.3385775. 5.4

[237] Diogo Pacheco, Pik-Mai Hui, Christopher Torres-Lugo, Bao Tran Truong, Alessandro Flammini, and Filippo Menczer. Uncovering coordinated networks on social media:

Methods and case studies. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):455–466, May 2021. URL https://ojs.aaai.org/index.php/ICWSM/article/view/18075. 5.2.2

[238] Susannah BF Paletz, Michael A Johns, Egle E Murauskaite, Ewa M Golonka, Nick B Pandža, C Anton Rytting, Cody Buntain, and Devin Ellis. Emotional content and sharing on facebook: A theory cage match. *Science Advances*, 9(39):eade9231, 2023. 6.6.2

[239] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pages 1–22, 2023. 6.2

[240] James W. Pennebaker, Matthias R. Mehl, and Kate G. Niederhoffer. Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1):547–577, 2003. 4.7

[241] Gordon Pennycook and David G. Rand. Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proceedings of the National Academy of Sciences*, 116(7):2521–2526, 2019. 4.7

[242] Gordon Pennycook and David G Rand. The psychology of fake news. *Trends in cognitive sciences*, 25(5):388–402, 2021. 6.2

[243] Richard E. Petty and Pablo Brinol. Persuasion: From single to multiple to metacognitive processes. *Perspectives on Psychological Science*, 3(2):137–147, 2008. 4.2

[244] Samantha C Phillips, Lynnette Hui Xian Ng, Wenqi Zhou, and Kathleen M Carley. Moral and emotional influences on attitude stability towards covid-19 vaccines on social media. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*, pages 216–225. Springer, 2024. 9.1

[245] Lara Schibelsky Godoy Piccolo, Pinelopi Troullinou, and Harith Alani. Chatbots to support children in coping with online threats: Socio-technical requirements. In *Proceedings of the 2021 ACM designing interactive systems conference*, pages 1504–1517, 2021. 2.2.1, 3.2, 3.5

[246] Marius-Constantin Popescu, Valentina E. Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. Multilayer perceptron and neural networks. *WSEAS Trans. Cir. and Sys.*, 8(7): 579–588, July 2009. ISSN 1109-2734. 2.4

[247] Nicolas Pröllochs and Stefan Feuerriegel. Mechanisms of true and false rumor sharing in social media: collective intelligence or herd behavior? *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–38, 2023. 4.7

[248] Umair Qazi, Muhammad Imran, and Ferda Ofli. Geocov19: a dataset of hundreds of millions of multilingual covid-19 tweets with location information. *SIGSPATIAL Special*, 12(1):6–15, 2020. 4.2

[249] Steve Rathje and Jay J Van Bavel. The psychology of virality. *Trends in Cognitive Sciences*, 2025. 4.2, 4.7

[250] Carlo Ratti, Stanislav Sobolevsky, Francesco Calabrese, Clio Andris, Jonathan Reades, Mauro Martino, Rob Claxton, and Steven H Strogatz. Redrawing the map of great britain from a network of human interactions. *PloS one*, 5(12):e14248, 2010. 4.2

[251] Adrian Rauchfleisch and Jonas Kaiser. The false positive problem of automatic bot detection in social science research. *PLOS ONE*, 15(10):1–20, 10 2020. doi: 10.1371/journal.pone.0241045. URL `https://doi.org/10.1371/journal.pone.0241045`. 2.4.2

[252] Mathieu Ravaut, Shafiq Joty, and Nancy Chen. Summareranker: A multi-task mixture-of-experts re-ranking framework for abstractive summarization. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4504–4524, 2022. 2.4

[253] Björn Ross, Laura Pilz, Benjamin Cabrera, Florian Brachten, German Neubaum, and Stefan Stieglitz. Are social bots a real threat? An agent-based model of the spiral of silence to analyse the impact of manipulative actors in social networks. *European Journal of Information Systems*, 28(4):394–412, July 2019. ISSN 0960-085X. doi: 10.1080/0960085X.2018.1560920. 6.4.3

[254] Michael Rossetti and Tauhid Zaman. Bots, disinformation, and the first impeachment of us president donald trump. *Plos one*, 18(5):e0283971, 2023. 4.3

[255] J. Ruan. A fully automated method for discovering community structures in high dimensional data. In *2009 Ninth IEEE International Conference on Data Mining*, pages 968–973, 12 2009. doi: 10.1109/ICDM.2009.141. 5.4

[256] Stuart J. Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach (4th Edition)*. Pearson, 2020. ISBN 9781292401133. 3.2

[257] LUIS E Santana and Gonzalo Huerta. Are they bots? social media automation during chile's 2017 presidential campaign. *Cuadernos. info*, 44:61–77, 2019. 3.2

[258] Rose Marie Santini, Debora Salles, Giulia Tucci, Fernando Ferreira, and Felipe Grael. Making up audience: Media bots and the falsification of the public sphere. *Communication Studies*, 71(3):466–487, 2020. 5.3.2

[259] Rose Marie Santini, Débora Salles, Charbelly Estrella Estrella, Carlos Eduardo Barros, and Daniela Orofino. Bots as online impersonators: automated manipulators and their different roles on social media. *The International Review of Information Ethics*, 30(1), 2021. 3.3

[260] Mohsen Sayyadiharikandeh, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. Detection of novel social bots by ensembles of specialized classifiers. In *Proceedings of the 29th ACM international conference on information & knowledge management*, pages 2725–2732, 2020. 2.4, **??**, **??**, 3.3

[261] David Schoch, Franziska B Keller, Sebastian Stier, and JungHwan Yang. Coordination patterns reveal online political astroturfing across the world. *Scientific reports*, 12(1): 4572, 2022. 7.2

[262] Ozgur Can Seckin, Aybuke Atalay, Ege Otenen, Umut Duygu, and Onur Varol. Mech-

anisms driving online vaccine debate during the COVID-19 pandemic. *Social Media & Society*, 10(1):20563051241229657, 2024. 4.2

[263] Chengcheng Shao, Giovanni Luca Ciampaglia, Onur Varol, Kai-Cheng Yang, Alessandro Flammini, and Filippo Menczer. The spread of low-credibility content by social bots. *Nature communications*, 9(1):1–9, 2018. 3.2, 3.5, 4.2, 4.7

[264] Sam Shead. Elon Musk says Twitter deal 'cannot move forward' until he has clarity on fake account numbers — cnbc.com. https://www.cnbc.com/2022/05/17/elon-musk-says-twitter-deal-cannot-move-forward-until-he-has-clarity-o html, 2022. [Accessed 02-05-2025]. 4.3

[265] Fei Shen, Erkun Zhang, Wujiong Ren, Yuan He, Quanxin Jia, and Hongzhong Zhang. Examining the differences between human and bot social media accounts: A case study of the russia-ukraine war. *First Monday*, 28(2), 2023. 4.2, 7.2

[266] Muhammad Hammad Fahim Siddiqui, Iqra Ameer, Alexander F Gelbukh, and Grigori Sidorov. Bots and gender profiling on twitter. In *CLEF (Working Notes)*, 2019. 3.3

[267] Daniel Silverman, Karl Kaltenthaler, and Munqith Dagher. Seeing is disbelieving: The depths and limits of factual misinformation in war. *International Studies Quarterly*, 65 (3):798–810, 2021. 7.6.2

[268] Almog Simchon, William J Brady, and Jay J Van Bavel. Troll and divide: the language of online polarization. *PNAS nexus*, 1(1):pgac019, 2022. 2.4

[269] Mike Simpson. Social Media Bots 101 - All You Need to Know — meltwater.com. https://www.meltwater.com/en/blog/social-media-bots, 2024. [Accessed 29-10-2024]. **??**

[270] Pranay Sindhu and Kumkum Bharti. Influence of chatbots on purchase intention in social commerce. *Behaviour & Information Technology*, 43(2):331–352, 2024. 3.5

[271] Laura Slechten, Cédric Courtois, Lennert Coenen, and Bieke Zaman. Adapting the selective exposure perspective to algorithmically governed platforms: The case of google search. *Communication Research*, 49(8):1039–1065, 2022. 4.2

[272] Bridget Smart, Joshua Watt, Sara Benedetti, Lewis Mitchell, and Matthew Roughan. # istandwithputin versus# istandwithukraine: the interaction of bots and humans in discussion of the russia/ukraine war. In *International Conference on Social Informatics*, pages 34–53. Springer, 2022. 3.3, 7.2

[273] Public Sphere. Making up audience: Media bots and the falsification of the. *Communicating Artificial Intelligence (AI): Theory, Research, and Practice*, page 98, 2020. 3.3

[274] Mamidala Sruthi, Sravanthi Thota, P Kumaraswamy, and G Mahesh. Demonstrate the performance of social bots in identifying the malicious bots based during communication in social media. In *AIP Conference Proceedings*, volume 2418, page 020052. AIP Publishing LLC, 2022. 3.3

[275] Zachary C. Steinert-Threlkeld, Delia Mocanu, Alessandro Vespignani, and James Fowler. Online social networks and offline protest. *EPJ Data Science*, 4(1):19, Nov 2015. ISSN 2193-1127. doi: 10.1140/epjds/s13688-015-0056-y. URL https://doi.org/10.

1140/epjds/s13688-015-0056-y. 5.2.2

[276] Christine Steinmetz, Homa Rahmat, Nancy Marshall, Kate Bishop, Susan Thompson, Miles Park, Linda Corkery, and Christian Tietz. Liking, tweeting and posting: an analysis of community engagement through social media platforms. *Urban Policy and Research*, 39(1):85–105, 2021. 4.7

[277] Massimo Stella, Emilio Ferrara, and Manlio De Domenico. Bots increase exposure to negative and inflammatory content in online social systems. *Proceedings of the National Academy of Sciences*, 115(49):12435–12440, 2018. 4.2, 4.6, 4.2, 4.7

[278] Dahlia Patricia Sterling. A new era in cultural diplomacy: Promoting the image of china's "belt and road" initiative in asia. *Open Journal of Social Sciences*, 6(2):102–116, 2018. 4.5

[279] Stefan Stieglitz, Florian Brachten, Björn Ross, and Anna-Katharina Jung. Do social bots dream of electric sheep? a categorisation of social media bot accounts. *arXiv preprint arXiv:1710.04044*, 2017. 3.2, 3.3, 4.1

[280] Victor Suarez-Lledo and Javier Alvarez-Galvez. Assessing the role of social bots during the covid-19 pandemic: infodemic, disagreement, and criticism. *Journal of Medical Internet Research*, 24(8):e36085, 2022. 3.3

[281] Michael Szell and Stefan Thurner. Measuring social dynamics in a massive multiplayer online game. *Social networks*, 32(4):313–329, 2010. 5.2.1

[282] Zhaoxuan Tan, Shangbin Feng, Melanie Sclar, Herun Wan, Minnan Luo, Yejin Choi, and Yulia Tsvetkov. BotPercent: Estimating bot populations in Twitter communities. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14295–14312, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.954. URL https://aclanthology.org/2023.findings-emnlp.954. **??**, **??**

[283] Sara-Jayne Terp and Pablo Breuer. Disarm: a framework for analysis of disinformation campaigns. In *2022 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*, pages 1–8. IEEE, 2022. 4.2

[284] Seth Tisue, Uri Wilensky, et al. Netlogo: A simple environment for modeling complexity. In *International conference on complex systems*, volume 21, pages 16–21. Boston, MA, 2004. 6.5

[285] Alexandru Topirceanu, Mihai Udrescu, and Radu Marculescu. Weighted betweenness preferential attachment: A new mechanism explaining social network formation and evolution. *Scientific reports*, 8(1):10871, 2018. 6.6.1

[286] Petter Törnberg. Echo chambers and viral misinformation: Modeling fake news as complex contagion. *PLoS one*, 13(9):e0203958, 2018. 6.3

[287] Mallory Trent, Holly Seale, Abrar Ahmad Chughtai, Daniel Salmon, and C. Raina MacIntyre. Trust in government, intention to vaccinate and COVID-19 vaccine hesitancy: A comparative survey of five large cities in the united states, united kingdom, and australia. *Vaccine*, 40(17):2498–2505, 2022. 4.7

[288] Milo Trujillo, Sam Rosenblatt, Guillermo de Anda Jáuregui, Emily Moog, Briane Paul V. Samson, Laurent Hébert-Dufresne, and Allison M. Roth. When the echo chamber shatters: Examining the use of community-specific language post-subreddit ban. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 164–178, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.woah-1.18. URL https://aclanthology.org/2021.woah-1.18. 1.3.2

[289] Milena Tsvetkova, Ruth García-Gavilanes, Luciano Floridi, and Taha Yasseri. Even good bots fight: The case of wikipedia. *PloS one*, 12(2):e0171774, 2017. 3.2, 3.5

[290] Amos Tversky and Daniel Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974. 4.2

[291] Seungha Um and Samrachana Adhikari. Considerations in bayesian agent-based modeling for the analysis of covid-19 data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 17(1):e11655, 2024. 3.2

[292] Joshua Uyheng and Kathleen M Carley. Bot impacts on public sentiment and community structures: Comparative analysis of three elections in the asia-pacific. In *International conference on social computing, behavioral-cultural modeling and prediction and behavior representation in modeling and simulation*, pages 12–22. Springer, 2020. 1.3.1, 3.5

[293] Joshua Uyheng and Kathleen M Carley. Bots and online hate during the covid-19 pandemic: case studies in the united states and the philippines. *Journal of computational social science*, 3(2):445–468, 2020. 2.2.2

[294] Joshua Uyheng, Lynnette Hui Xian Ng, and Kathleen M Carley. Active, aggressive, but to little avail: characterizing bot activity during the 2020 singaporean elections. *Computational and Mathematical Organization Theory*, 27(3):324–342, 2021. (document), 1.3.1, 4.8, 9.1, 9.2

[295] Onur Varol, Emilio Ferrara, Clayton Davis, Filippo Menczer, and Alessandro Flammini. Online human-bot interactions: Detection, estimation, and characterization. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 280–289, 2017. 3.2, 3.3, 4.3, 4.7, 5.2.1

[296] Onur Varol, Clayton A Davis, Filippo Menczer, and Alessandro Flammini. Feature engineering for social bot detection. In *Feature engineering for machine learning and data analytics*, pages 311–334. CRC Press, 2018. 2.2.2

[297] Tony Veale, Alessandro Valitutti, and Guofu Li. Twitter: The best of bot worlds for automated wit. In *Distributed, Ambient, and Pervasive Interactions: Third International Conference, DAPI 2015, Held as Part of HCI International 2015, Los Angeles, CA, USA, August 2-7, 2015, Proceedings 3*, pages 689–699. Springer, 2015. 3.5

[298] Otavio R Venâncio, Carlos HG Ferreira, Jussara M Almeida, and Ana Paula C da Silva. Unraveling user coordination on telegram: A comprehensive analysis of political mobilization during the 2022 brazilian presidential election. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 1545–1556, 2024. 5.2.2

[299] Lidia Vitkova, Maxim Kolomeec, and Andrey Chechulin. Taxonomy and bot threats in

social networks. In *2022 International Russian Automation Conference (RusAutoCon)*, pages 814–819. IEEE, 2022. 3.3

[300] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *science*, 359(6380):1146–1151, 2018. 4.7, 6.3

[301] Tiange Wang, Fengkai Wu, and Richard O. Sinnott. A case study in twitter bot identification: Are they still a problem? In *2020 Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, pages 1–8. IEEE, 2020. 4.2

[302] Stanley Wasserman and Katherine Faust. *Triads*, page 556–602. Structural Analysis in the Social Sciences. Cambridge University Press, 1994. 5.2.1

[303] Dominik Wawrzuta, Mariusz Jaworski, Joanna Gotlib, and Mariusz Panczyk. Characteristics of antivaccine messages on social media: systematic review. *Journal of Medical Internet Research*, 23(6):e24564, 2021. 4.2, 4.2

[304] Derek Weber and Lucia Falzon. Temporal nuances of coordination network semantics. *arXiv preprint arXiv:2107.02588*, 2021. 5.2.2

[305] Derek Weber and Frank Neumann. Who's in the gang? revealing coordinating communities in social media. In *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pages 89–93, 2020. doi: 10.1109/ASONAM49781.2020.9381418. 5.2.2

[306] Derek Weber and Frank Neumann. Amplifying influence through coordinated behaviour in social networks. *Social Network Analysis and Mining*, 11(1):111, 2021. 5.4

[307] Barry Wellman and Caroline Haythornthwaite, editors. *The Internet in Everyday Life*. Blackwell Publishing, Malden, MA, 2002. 5.2.1, 5.3.2

[308] Zixuan Weng and Aijun Lin. Public opinion manipulation on social media: Social network analysis of twitter bots during the covid-19 pandemic. *International journal of environmental research and public health*, 19(24):16376, 2022. 1

[309] Andrew Westbrook and Todd S Braver. Cognitive effort: A neuroeconomic approach. *Cognitive, Affective, & Behavioral Neuroscience*, 15(2):395–415, 2015. 6.5.2

[310] Stefan Wojcik. Bots in the Twittersphere — pewresearch.org. `https://www.pewresearch.org/internet/2018/04/09/bots-in-the-twittersphere/`, 2018. [Accessed 17-Jul-2023]. 1

[311] Marty J Wolf, K Miller, and Frances S Grodzinsky. Why we should have seen that coming: comments on microsoft's tay" experiment," and wider implications. *Acm Sigcas Computers and Society*, 47(3):54–64, 2017. 3.5

[312] Samuel C Woolley. Automating power: Social bot interference in global politics. *First Monday*, 2016. 3.3

[313] Samuel C Woolley and Philip N Howard. Political communication, computational propaganda, and autonomous agents: Introduction. *International journal of Communication*, 10, 2016. 1

[314] Samuel C Woolley and Philip N Howard. Social media, revolution, and the rise of the

political bot. In *Routledge handbook of media, conflict and security*, pages 302–312. Routledge, 2016. 3.3

[315] Jun Wu, Xuesong Ye, and Chengjie Mou. Botshape: A novel social bots detection approach via behavioral patterns. In *12th International Conference on Data Mining & Knowledge Management Process*, 2023. 2.2.2

[316] Wentao Xu and Kazutoshi Sasahara. Characterizing the roles of bots on twitter during the COVID-19 infodemic. *Journal of Computational Social Science*, 5(1):591–609, 2022. 4.2, 4.2

[317] Wentao Xu, Kazutoshi Sasahara, Jianxun Chu, Bin Wang, Wenlu Fan, and Zhiwen Hu. Social media warfare: investigating human-bot engagement in english, japanese and german during the russo-ukrainian war on twitter and reddit. *EPJ Data Science*, 14(1):10, 2025. 2.2.2, 7.2

[318] Harry Yaojun Yan and Kai-Cheng Yang. The landscape of social bot research: A critical appraisal. In *Handbook of Critical Studies of Artificial Intelligence*, pages 716–725. Edward Elgar Publishing, 2023. **??**, **??**

[319] Harry Yaojun Yan, Kai-Cheng Yang, James Shanahan, and Filippo Menczer. Exposure to social bots amplifies perceptual biases and regulation propensity. *Scientific Reports*, 13 (1):20707, 2023. 2.2.1, **??**, **??**

[320] Kai-Cheng Yang and Filippo Menczer. Anatomy of an ai-powered malicious social botnet. *Journal of Quantitative Description: Digital Media*, 4, 2024. **??**, **??**

[321] Kai-Cheng Yang, Onur Varol, Clayton A Davis, Emilio Ferrara, Alessandro Flammini, and Filippo Menczer. Arming the public with artificial intelligence to counter social bots. *Human Behavior and Emerging Technologies*, 1(1):48–61, 2019. 2.4

[322] Kai-Cheng Yang, Emilio Ferrara, and Filippo Menczer. Botometer 101: Social bot practicum for computational social scientists. *Journal of computational social science*, 5(2):1511–1528, 2022. 2.2.2, 2.4, 3.2, 3.3

[323] Ziyi Yang, Zaibin Zhang, Zirui Zheng, Yuxian Jiang, Ziyue Gan, Zhiyu Wang, Zijian Ling, Martin Ma, Bowen Dong, Prateek Gupta, et al. Oasis: Open agents social interaction simulations on one million agents. In *NeurIPS 2024 Workshop on Open-World Agents*, 2024. 6.2

[324] Tien Ee Dominic Yeo. Models of viral propagation in digital contexts: How messages and ideas—from internet memes to fake news—created by consumers, bots, and marketers spread. In *The Routledge Handbook of Digital Consumption*, pages 489–501. Routledge, 2022. 3.5

[325] Changseung Yoo, Eunae Yoo, Lu Yan, and Alfonso Pedraza-Martinez. Speak with one voice? examining content coordination and social media engagement during disasters. *Information Systems Research*, 35(2):551–569, 2024. 4.7

[326] Xiaoyi Yuan, Ross J. Schuchard, and Andrew T. Crooks. Examining emergent communities and social bots within the polarized online vaccination debate in twitter. *Social Media & Society*, 5(3):2056305119865465, 2019. 4.2

[327] Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012. 2.4

[328] Hammad Zafar, Fakhre Alam Siddiqui, and Midhat Arif. The digital duo: Exploring the impact of ai chatbots and digital marketing strategies on consumer purchase intentions in pakistan's e-commerce sector. *Journal for Social Science Archives*, 3(1):265–286, 2025. 3.5

[329] Damián H Zanette. Dynamics of rumor propagation on small-world networks. *Physical review E*, 65(4):041908, 2002. 6.2

[330] Bei Zhao, Wujiong Ren, Yicheng Zhu, and Hongzhong Zhang. Manufacturing conflict or advocating peace? a study of social bots agenda building in the twitter discussion of the russia-ukraine war. *Journal of Information Technology & Politics*, 21(2):176–194, 2024. 7.2

[331] Shaolin Zhu, Leiyu Pan, Dong Jian, and Deyi Xiong. Overcoming language barriers via machine translation with sparse mixture-of-experts fusion of large language models. *Information Processing & Management*, 62(3):104078, 2025. 2.4