# The Foundations of Cyber Social Agents

**Lynnette Ng**

**Thesis Defense**

**11 March 2026**

Committee: Dr Kathleen Carley (Chair), Dr Richard Carley,
Dr Nicolas Christin, Dr Melissa Chua

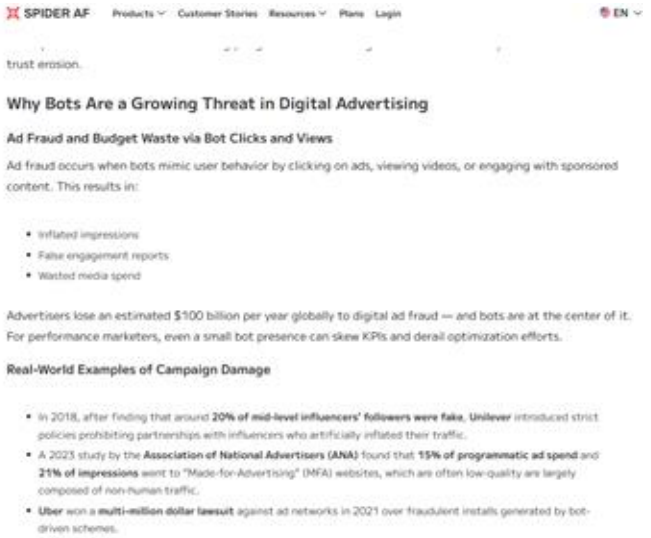
**The CMU center for Informed DEmocracy And Social cyber-security**


Carnegie Mellon University

Thesis Document & Slides

# Social Media Bots



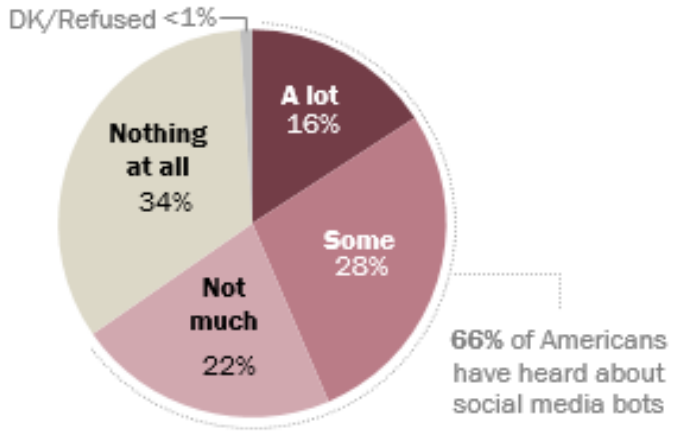



## Twitter is becoming a 'ghost town' of bots as AI-generated spam content floods the internet

By technology reporter James Purtill

## About two-thirds of Americans have heard about social media bots

*% of U.S. adults who have heard ___ about social media bots*




Source: Survey conducted July 30-Aug. 12, 2018.
"Social Media Bots Draw Public's Attention and Concern"

PEW RESEARCH CENTER



Carnegie Mellon University

# Thesis Statement

Social Media Bots are not a homogeneous adversary,
but are a heterogeneous mix of personas
which act as Cyber Social Agents
to shape perceptions and collective behavior at scale,
with the potential to harm and benefit humanity

# Thesis Findings

1. Bot detection models are more effective with metadata features as compared to content-only classifiers
2. Bots consistently constitute about 20% of the online actors on X
3. Bots are a heterogeneous class of Cyber Social Agents
4. Legacy bot technologies (i.e. heuristic-based) persist alongside new technologies (i.e. generative AI)
5. Bots and humans use comparable levels of moral and emotional language
6. Bots are more likely than humans to send messages that invoke these cognitive biases in the reader.
7. Bots actively exploit social media platform algorithms and affordances
8. Bots exhibit stronger coordination than humans, and typically coordinate with humans.
9. Bots are deployed strategically, not randomly
10. Bots are strategically positioned in a network for impact & rapid broadcast.
11. Bot influence emerges from collective dynamics rather than isolated actors.
12. Bots can induce measurable & observable changes in human stances
13. Bots can be reliably simulated with LLM-Augmented Agent-Based Models

# Thesis Findings

1. **Bot detection models are more effective with metadata features as compared to content-only classifiers**
2. Bots consistently constitute about 20% of the online actors on X
3. **Bots are a heterogeneous class of Cyber Social Agents**
4. Bots and humans use comparable levels of moral and emotional language
5. Legacy bot technologies (i.e. heuristic-based) persist alongside new technologies (i.e. generative AI)
6. Bots are more likely than humans to send messages that invoke these cognitive biases in the reader.
7. Bots actively exploit social media platform algorithms and affordances
8. **Bots exhibit stronger coordination than humans, and typically coordinate with humans.**
9. Bots are deployed strategically, not randomly
10. Bots are strategically positioned in a network for impact & rapid broadcast.
11. Bot influence emerges from collective dynamics rather than isolated actors.
12. **Bots can induce measurable & observable changes in human stances**
13. **Bots can be reliably simulated with LLM-Augmented Agent-Based Models**

# Datasets

- Over 5 million posts from 1 million users across five social media platforms (X, Reddit, Instagram, Telegram, Parler)
- Mixture of repository-based datasets and self-collected datasets
- 16 datasets were used
  - 12 dataset from X
  - 1 dataset from Reddit
  - 1 dataset from Instagram
  - 1 dataset from Telegram
  - 1 dataset from Parler

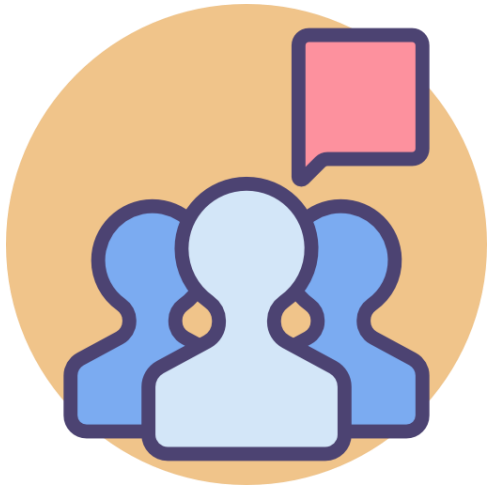
# Data in this Presentation

- Russian-Ukraine Conflict
- Dataset from X collected from April to June 2023 with the X Streaming API
  - “Russian invasion”, “Russian military”, “invasion of Ukraine”
- 38 million posts, 11 million users

- Presented in this thesis as [Case Study] slides

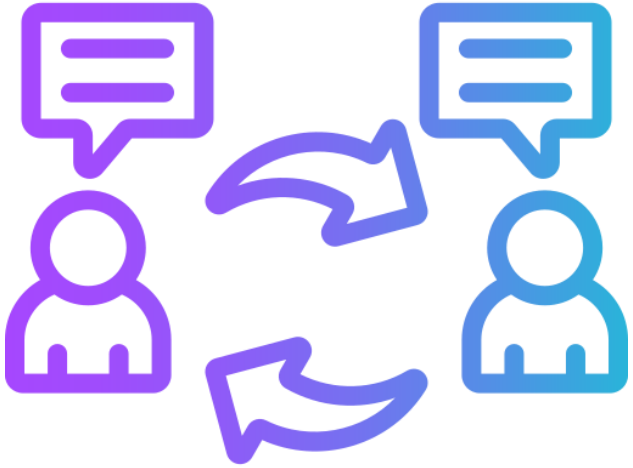
# Social Media Platforms

**User**



**Content**

**Interactions**





Carnegie Mellon University

# Cyber Social Agent

**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.

# Bots are the simplest form of Cyber Social Agent

**Definition. (Social Media Bot)**

user

An automated account that carries out a series of mechanics on social media platforms, for

content

content creation, distribution, and collection and processing, and/or for

interactions

relationship formation and dissolutions.

# From Bots to Cyber Social Agents

| Characteristics | General Perception of Bots | General Concept of Cyber Social Agents |
|---|---|---|
| Agency | Automated by operator | Automated by operator<br>Self-evolving<br>Have agency |
| Types | One general Type | Different types of personas |
| Nature | Bad | Can be good or bad<br>Can be harnessed for good or bad |
| Regulation | Should be regulated in the same way | Can be regulated differently based on persona, content, behavior |

# Thesis Findings

1. **Bot detection models are more effective with metadata features as compared to content-only classifiers**
2. Bots consistently constitute about 20% of the online actors on X
3. Bots are a heterogeneous class of Cyber Social Agents
4. Bots and humans use comparable levels of moral and emotional language
5. Legacy bot technologies (i.e. heuristic-based) persist alongside new technologies (i.e. generative AI)
6. Bots are more likely than humans to send messages that invoke these cognitive biases in the reader.
7. Bots actively exploit social media platform algorithms and affordances
8. Bots exhibit stronger coordination than humans, and typically coordinate with humans.
9. Bots are deployed strategically, not randomly
10. Bots are strategically positioned in a network for impact & rapid broadcast.
11. Bot influence emerges from collective dynamics rather than isolated actors.
12. Bots can induce measurable & observable changes in human stances
13. Bots can be reliably simulated with LLM-Augmented Agent-Based Models

# Bot Detection Algorithms

- Bot detection is the baseline method to surface, classify and measure the presence of automated, non-human actors online
- We developed five bot detection algorithms with the following improvements over the current state of the art algorithms

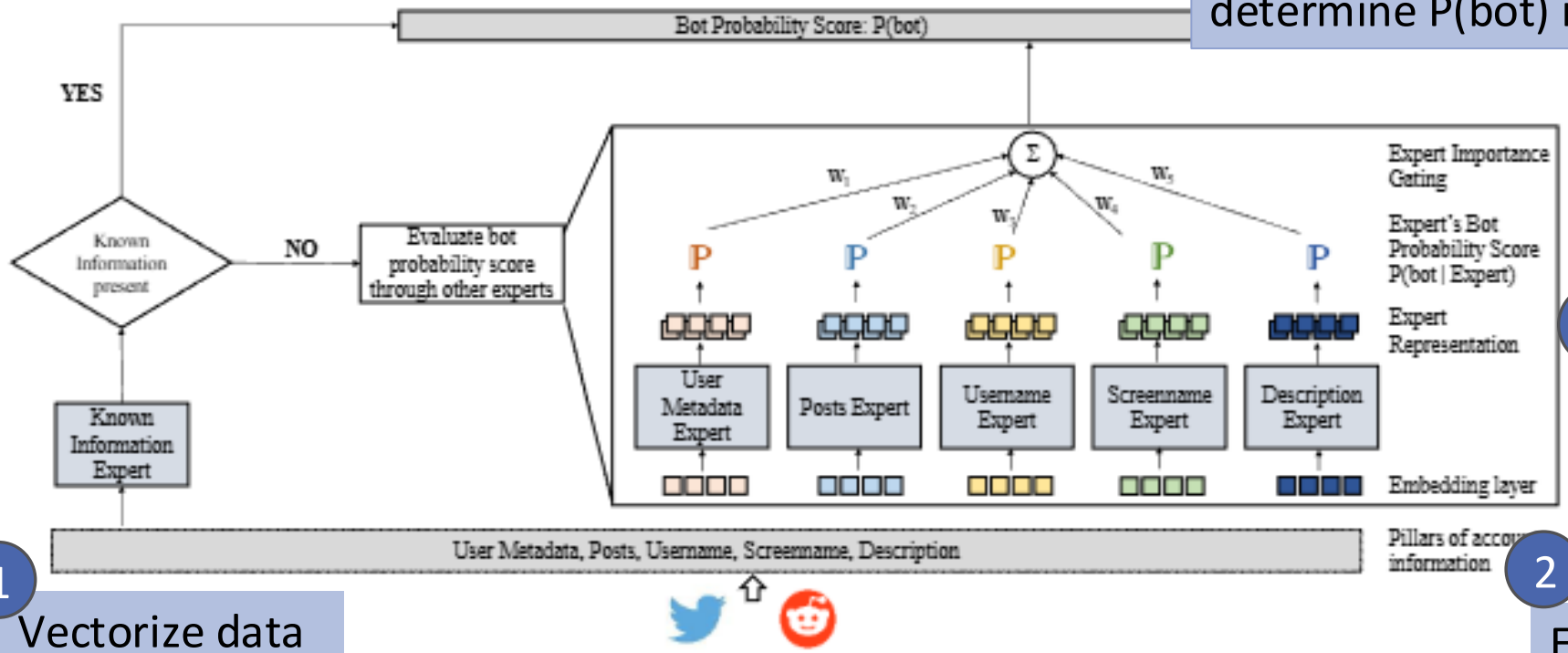
# Bot Detection Algorithms

- Bot detection is the baseline method to surface, classify and measure the presence of automated, non-human actors online
- We developed five bot detection algorithms with the following improvements over the current state of the art algorithms

# Bot Detection Algorithms

| | Literature | Lynnette's |
|---|---|---|
| **Platforms** | Operates on a single social media platform (mostly X) | Interoperable across multiple social media platforms |
| **Language** | Operates on a single language (mostly English) | Operates on a variety of languages |
| **Data source** | Operates on live data (requires API calls)<br>Requires the entire data format | Can operate on historical data and handle incomplete datasets |
| **Speed** | Slower<br>BotHunter (21 ± 25 ms) | Works reasonably fast<br>Tiny-BotBuster (12 ± 14 ms ) |
| **Accuracy** | BotHunter (55 ± 28 %) | Tiny-BotBuster (91 ± 8 % ) |

# Mixture of Experts Algorithm




5 Output of MLP passes through a logit to determine P(bot) in [0,1]

4 Experts are aggregated together by a multi-layer perceptron

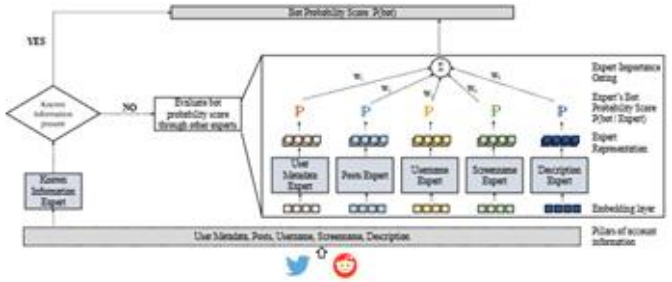
3 Expert represents result as a vector

2 Each expert is an ML-model trained to handle a specialized data pillar

1 Vectorize data (768d BERT embeddings)

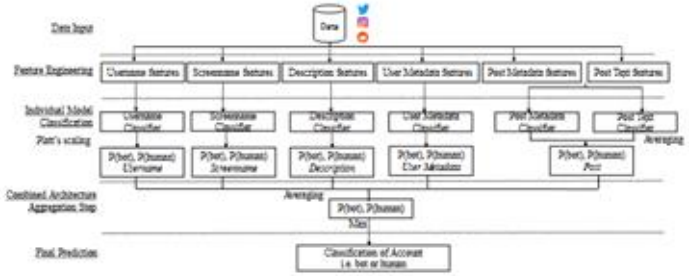
# Bot Detection Algorithms

**BotBuster (2022, AAAI):**
Deep Learning

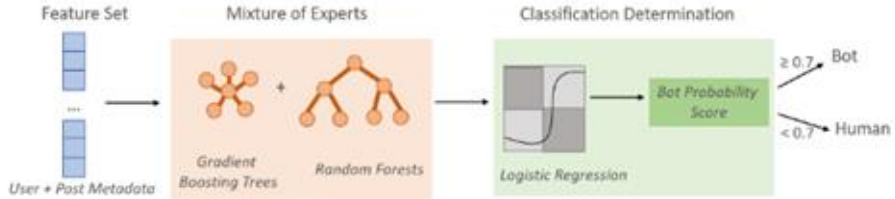


**BotBuster For Everyone (2024, SNAM) &**

**BotBuster Telegram (2024, JCSS)**
Transformer-based Ensemble



**Tiny-BotBuster (2024, SBP-BRiMS):**
Random Forest Ensemble

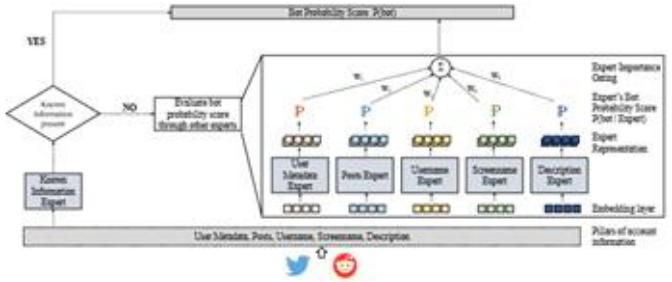





Carnegie Mellon University

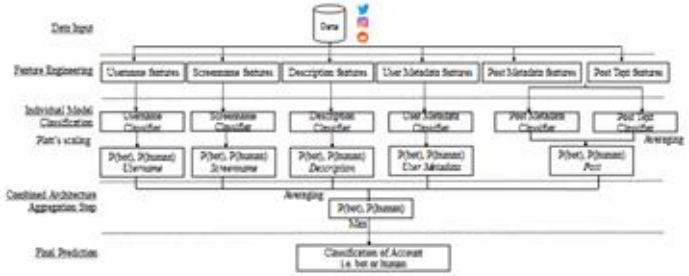
# Bot Detection Algorithms

**BotBuster (2022, AAAI):**
Deep Learning

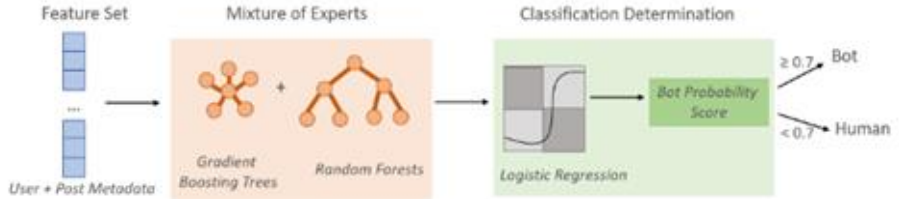


**BotBuster For Everyone (2024, SNAM) &**

**BotBuster Telegram (2024, JCSS)**
Transformer-based Ensemble



**Tiny-BotBuster (2024, SBP-BRiMS):**
Random Forest Ensemble



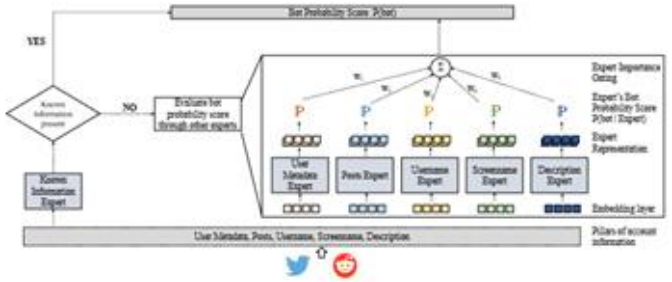

Increasing efficiency
Increasing accuracy
Decreasing time taken for classification



Carnegie Mellon University

# Bot Detection Algorithms

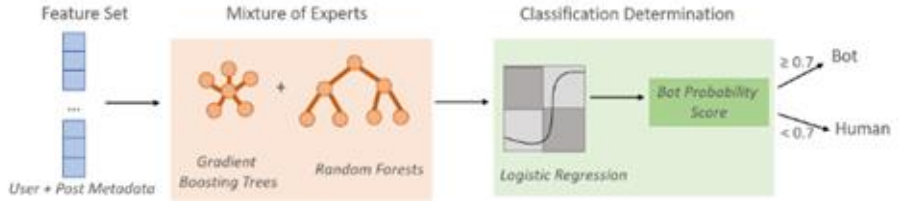
**BotBuster (2022, AAAI):**
Deep Learning



**BotBuster For Everyone (2024, SNAM) &**

**BotBuster Telegram (2024, JCSS)**
Transformer-based Ensemble

**Tiny-BotBuster (2024, SBP-BRiMS):**
Random Forest Ensemble




**Multilingual BotBuster**
google-bert/bert-base-multilingual uncased
Best performing multilingual model after comparing: language-specific transformers, language-agnostic transformers, Large Language Models



Carnegie Mellon University

# Bot Detection Algorithms

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

# Bot Detection Algorithms

| | BotBuster | BotBuster For Everyone | Tiny BotBuster | Multilingual BotBuster | BotBuster Telegram |
|---|---|---|---|---|---|
| **Model Architecture (Mixture-of-Experts type)** | Deep learning | Transformer-based | Random forests | Random forests | Transformer-based |
| **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 | Faster version of BotBuster that does not require GPUs | Small model size | Tested for English, Russian, Chinese, Arabic languages | |
| **Limitations** | Slow model building & inference | Only for English language | Solely for X | Relies on translated language data | Solely for Telegram |

# What value ε should we use to threshold P(bot)?

- P(bot): Probability the user is a bot based on the features of the machine learning model
- P(bot) is usually returned as a float in the range [0,1]
- What value ε should we use as a bot probability threshold for P(bot) for stable and consistent bot detection results?
- Many different studies use different threshold values, even studies that use the same bot detection algorithm

# What value ε should we use to threshold P(bot)?

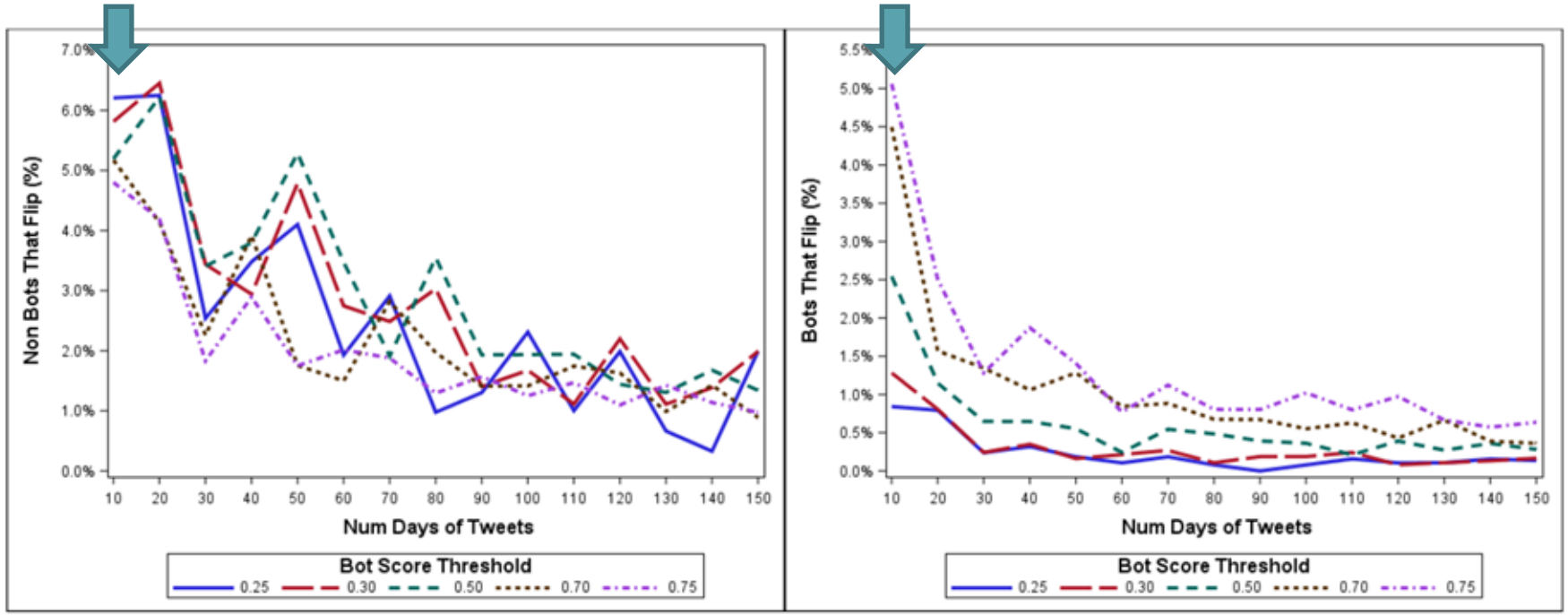
- Collected tweets of 5000 agents from X on a daily basis for 150 days
- The users had at least 100 tweets each
- Analyzed "flipping bot classification"
  - At a threshold value, the user is being classified as a bot or human
  - If the P(bot) is above the threshold, the user is classified as a bot; if the P(bot is below the threshold, the user is classified as a human
  - Used five threshold values: [0.25, 0.30., 0.50, 0.70, 0.75]

# What value ε should we use to threshold P(bot)?

- Stability: Small change in bot probability score that leads to the classification will unlikely alter the agent's classification
- Investigated stability across two dimensions:
  - **Temporal stability**: how bot scores changed with increasing number of days
  - **Volume stability:** how scores changed across number of tweets

# What value ε should we use to threshold P(bot)?

- **For temporal stability:**
  - Largest percentage of flips occur at the 10-day mark for both bots and non-bots



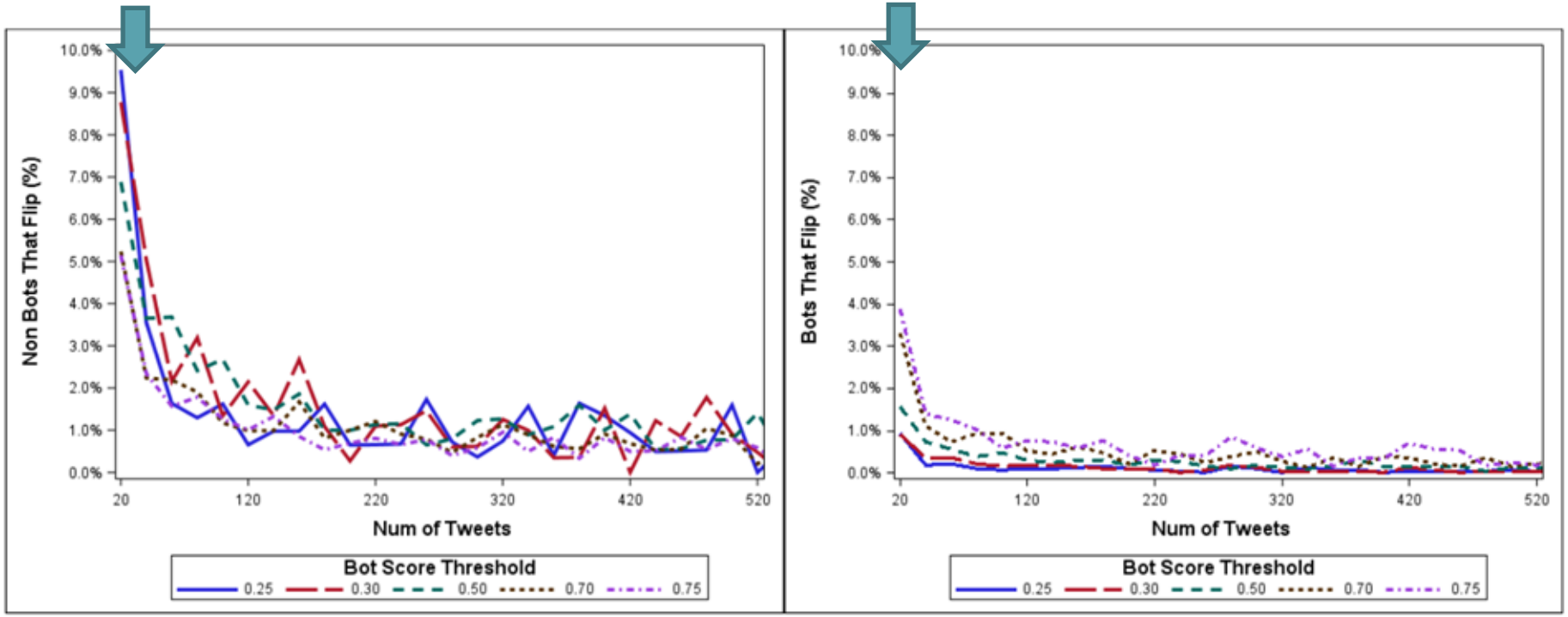

((a)) Non-Bots that flip

((b)) Bots that flip

Mellon University

# What value ε should we use to threshold P(bot)?

- **For volume stability:**
  - Largest percentage of flips occur at the 20-tweet mark for both bots and non-bots




((a)) Non bots that flip

((b)) Bots that flip

Mellon University

# What value ε should we use to threshold P(bot)?

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

# [Case Study] Bot proportion per month

- Measured using Tiny-BotBuster
  - Fastest algorithm present
- Across all three months, proportions remain relatively the same
- Proportion about 20%, which is similar to the average number in other events studied in this thesis (elections, movies, social crisis etc)

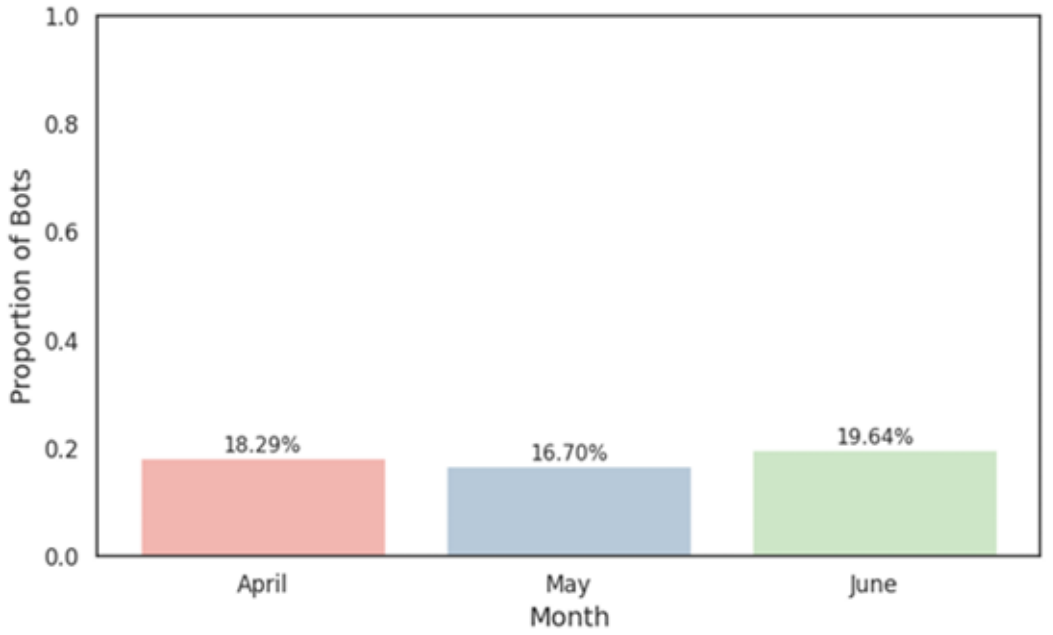



Figure 7.1: Bot proportion per month

# Thesis Findings

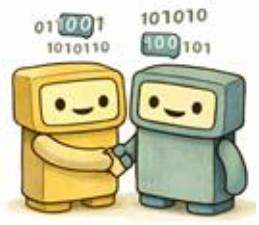
1. Bot detection models are more effective with metadata features as compared to content-only classifiers
2. Bots consistently constitute about 20% of the online actors on X
3. **Bots are a heterogeneous class of Cyber Social Agents**
4. Bots and humans use comparable levels of moral and emotional language
5. Legacy bot technologies (i.e. heuristic-based) persist alongside new technologies (i.e. generative AI)
6. Bots are more likely than humans to send messages that invoke these cognitive biases in the reader.
7. Bots actively exploit social media platform algorithms and affordances
8. Bots exhibit stronger coordination than humans, and typically coordinate with humans.
9. Bots are deployed strategically, not randomly
10. Bots are strategically positioned in a network for impact & rapid broadcast.
11. Bot influence emerges from collective dynamics rather than isolated actors.
12. Bots can induce measurable & observable changes in human stances
13. Bots can be reliably simulated with LLM-Augmented Agent-Based Models

# Different types of Cyber Social Agents

## By Operational Tactics

Amplifier

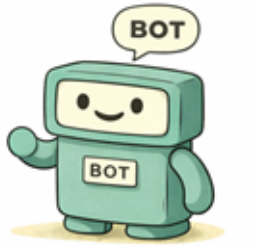
Social Influence



Cyborgs

Bridging

Synchronized




Self-Declared

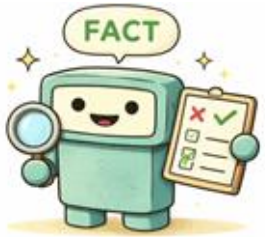
Repeater

Chaos

## By Rhetorical Strategy

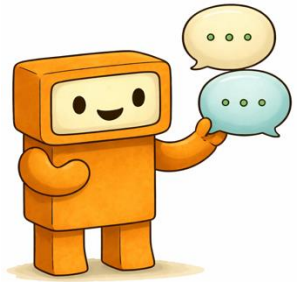
Announcer

Content Generation




Information Correction

Genre Specific



Conversational

Engagement Generation


News

(Images by Lynnette)



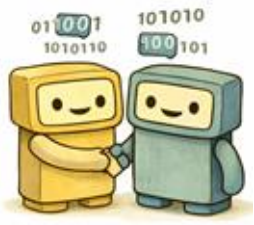
Carnegie Mellon University

# Different types of Cyber Social Agents

## By Operational Tactics

Amplifier

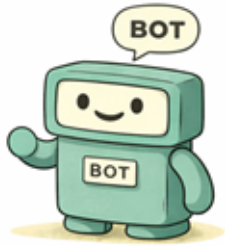
Social Influence




Cyborgs

Bridging

Synchronized




Self-Declared

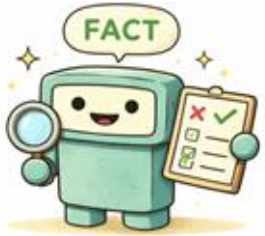
Repeater

Chaos

## By Rhetorical Strategy

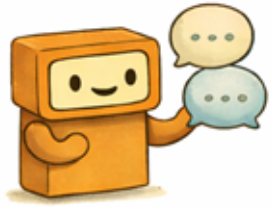
Announcer

Content Generation




Information Correction

Genre Specific



Conversational

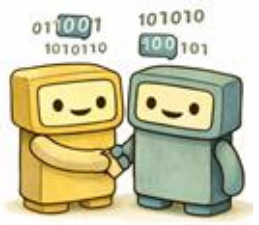
Engagement Generation


News



Carnegie Mellon University

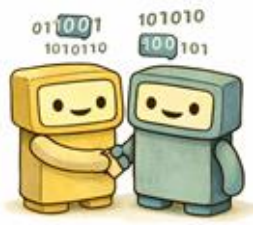
# Cyborgs



Cyborgs

- Hybrid accounts that combine automated functionality with human oversight
- Automation assists the human operator in tasks like mass sharing, high-frequency posting etc
- Human intervention provides flexibility, nuance, personal touch

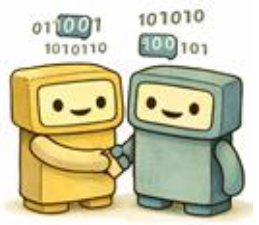
# Cyborgs Properties



Cyborgs

- **User:** Frequent changes in bot classification with high change in bot probability score
- **Content:** sometimes bot-like, sometimes human-like
- **Interaction:** sometimes bot-like, sometimes human-like
- **Algorithmic effect:** Sustained by social media algorithms

# Cyborgs

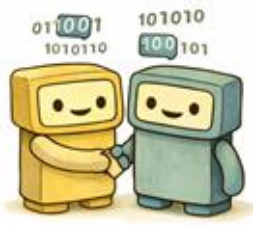


Cyborgs

- Produces alternating patterns of behavior that confuse bot detection and recommender systems
  - (Automation) Consistency & activity frequency = indicators of engagement
  - (Humaness) Linguistic variety & contextually appropriate responses => gathers trust

| | Coronavirus dataset | | | US Elections dataset | | |
|---|---|---|---|---|---|---|
| | Bot | Cyborg | Human | Bot | Cyborg | Human |
| Proportion suspended (%) | 89.2 | 56.0 | 19.5 | 76.9 | 49.2 | 19.9 |
| Avg length of acct (days) | 2751±1226 | 3663±1141 | 2901±1294 | 2643±1304 | 3437±1173 | 2715±1255 |

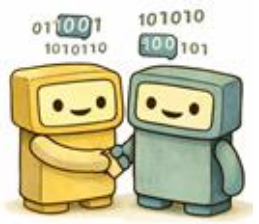
# Cyborgs



Cyborgs

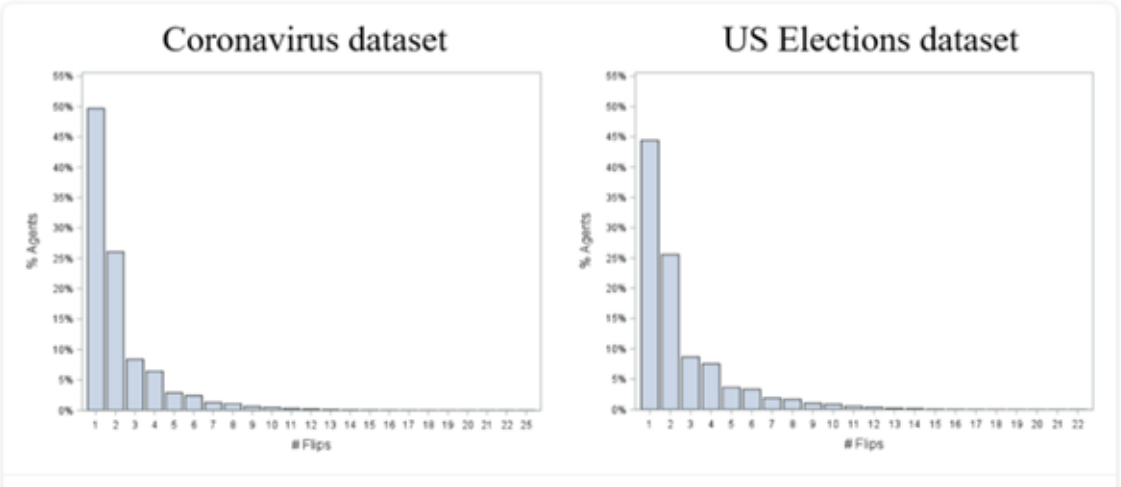
- Computationally, a Cyborg is sometimes a bot and sometimes a human
  - Frequently flip bot classification
  - Large change in bot probability scores between flips

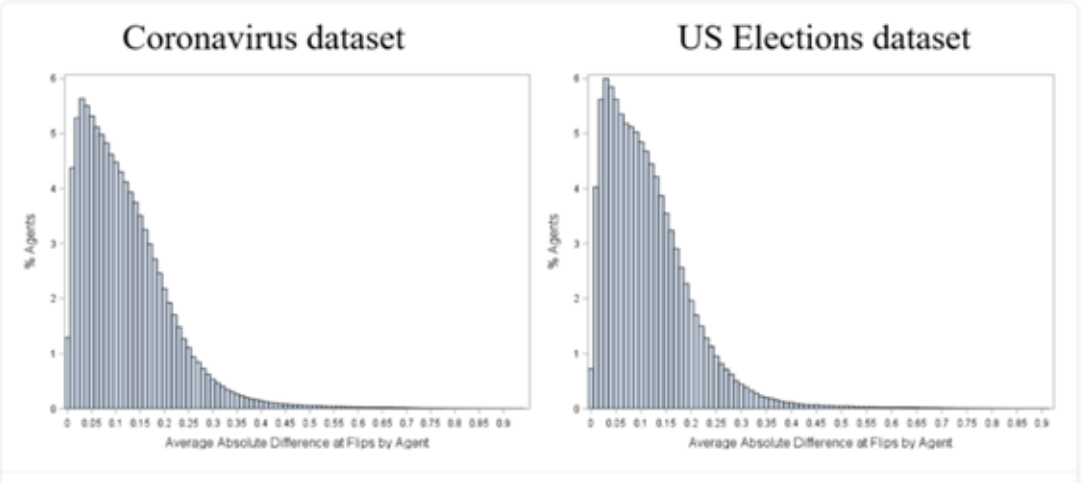
# Cyborgs



Cyborgs

- **Temporal Changes in Bot Probability Score**: compare changes across time to identify number of changes per agent and difference in bot scores
- Quantitative threshold values for identifying Cyborgs => where the proportion of users that change classification tapers off

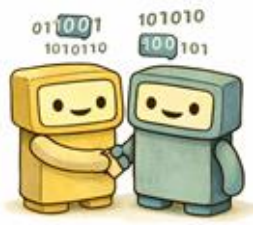




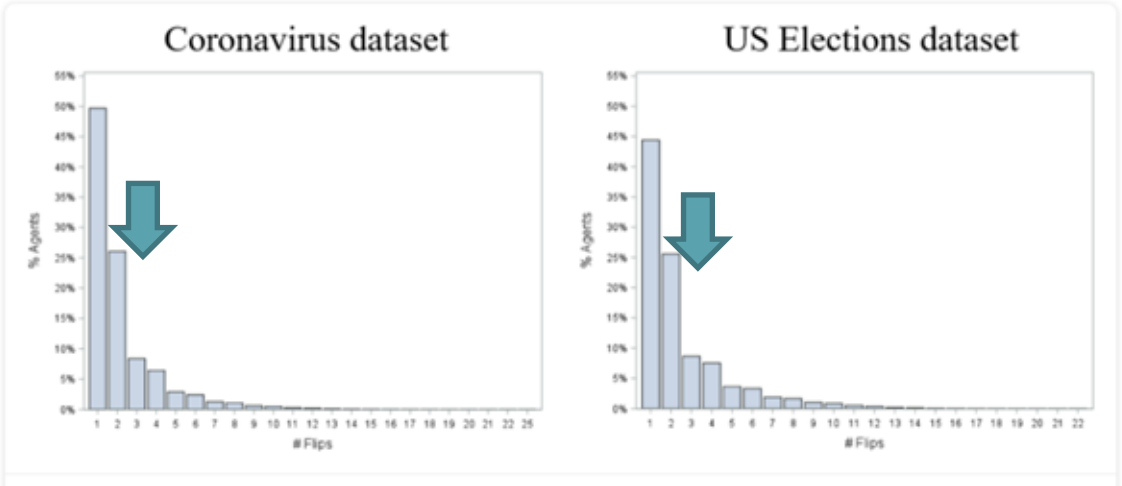

Carnegie Mellon University

# Cyborgs



Cyborgs

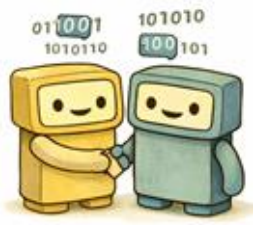
- Quantitative Values of Cyborgs
  - 3 flips of bot classification (i.e. bot -> human, human -> bot)





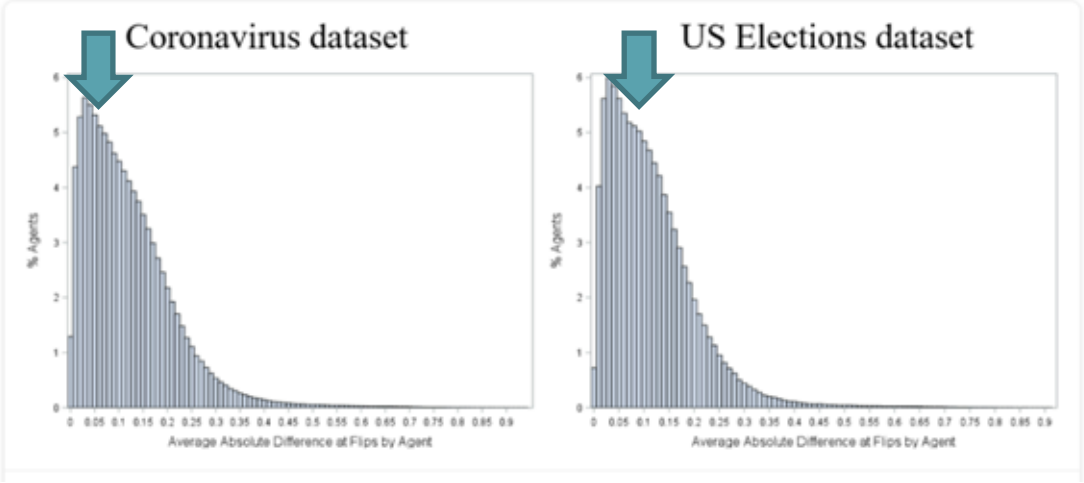

Carnegie Mellon University

# Cyborgs



Cyborgs

- Quantitative Values of Cyborgs
  - 3 flips of bot classification (i.e. bot -> human, human -> bot)
  - Average change in bot probability score between flips at least 0.10






Carnegie Mellon University

# Bridging Agents

Bridging

- Connect otherwise separate groups of users & serve as intermediaries in the flow of information across communities
- Groups can differ by interests, identities, ideologies or social spaces
- Bridging Agents draw groups into shared conversations by tagging multiple users, cross-posting overlapping content that is of interest to the groups tagged
  - Make disparate groups visible
  - Expand overall reach of narratives, create a larger sphere of influence for themselves and other agents

# Bridging Agents

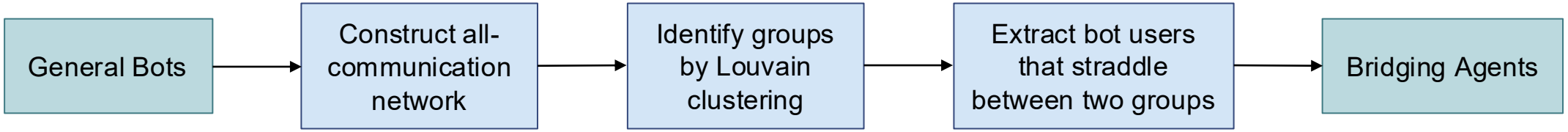
Bridging

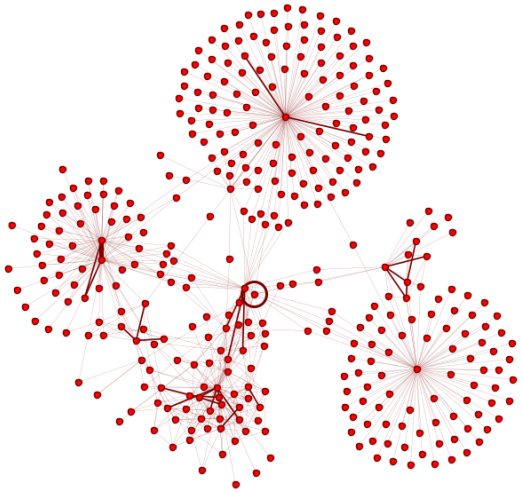
- **User:** High bridge score from the BEND maneuver metric; creates disconnected groups when removed from graph
- **Content:** (any)
- **Interaction:** Frequently tag multiple people from different social identity groups/ narratives; straddle between multiple network clusters
- **Algorithmic effect:** Mimic broad resonance with interactions from multiple user clusters

# Bridging Agents

Bridging

- Network-based methodology for identification



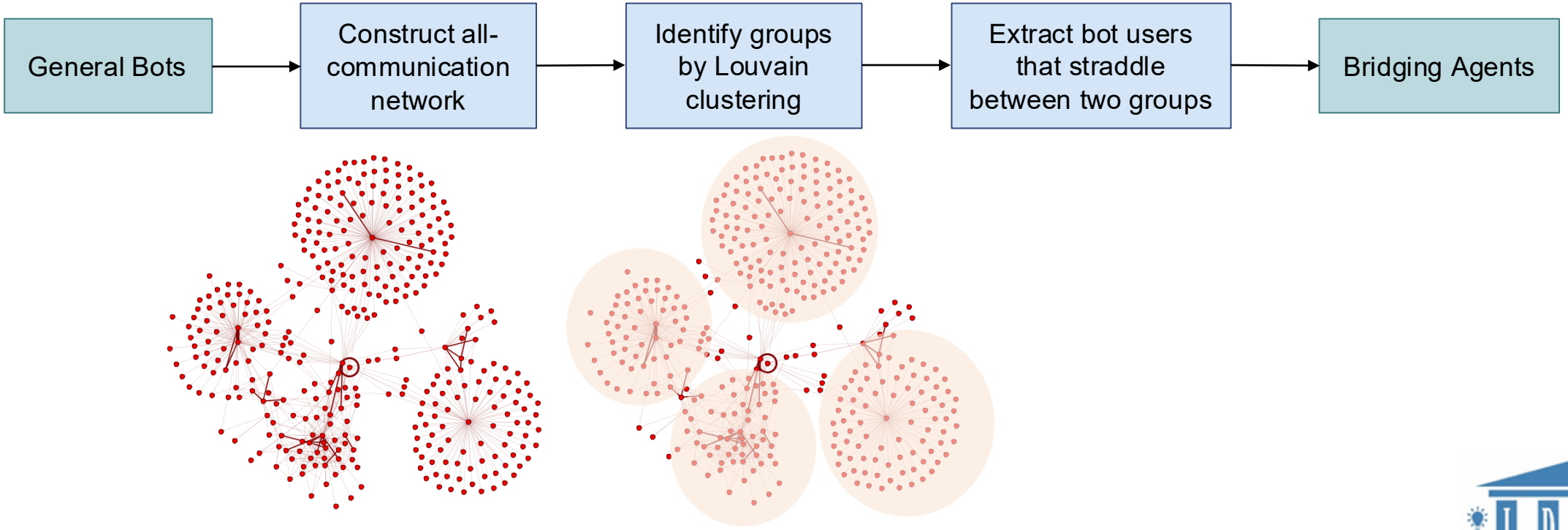


Carnegie Mellon University

# Bridging Agents

Bridging

- Network-based methodology for identification



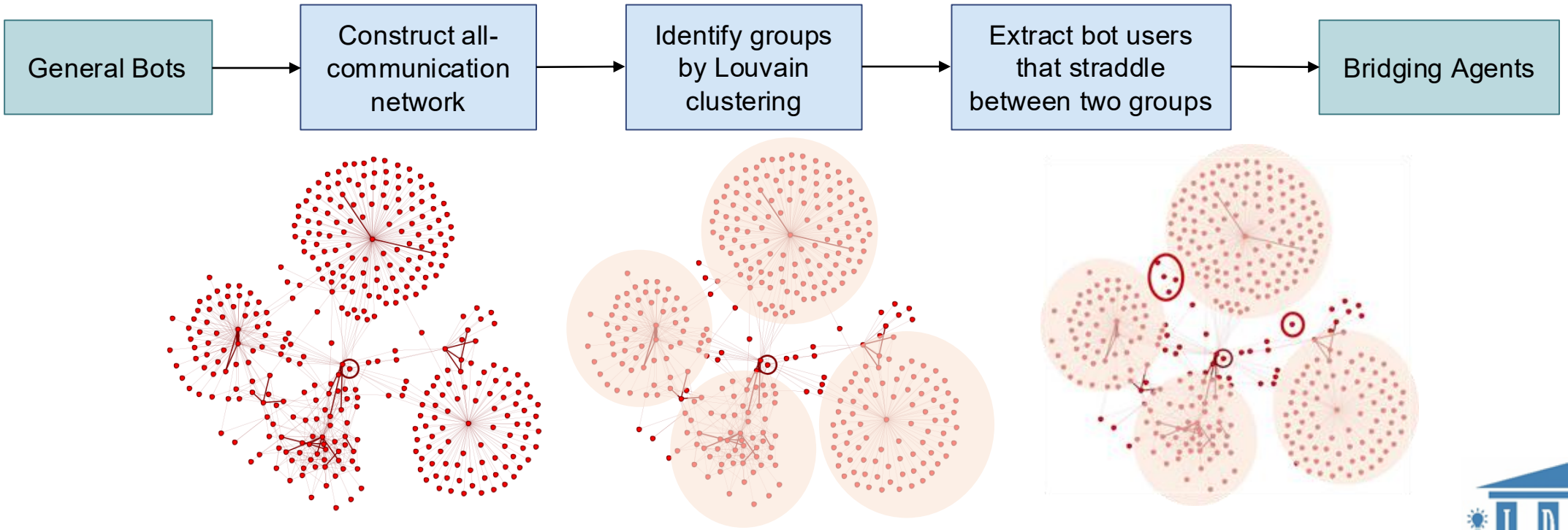



Carnegie Mellon University

# Bridging Agents

Bridging

- Network-based methodology for identification



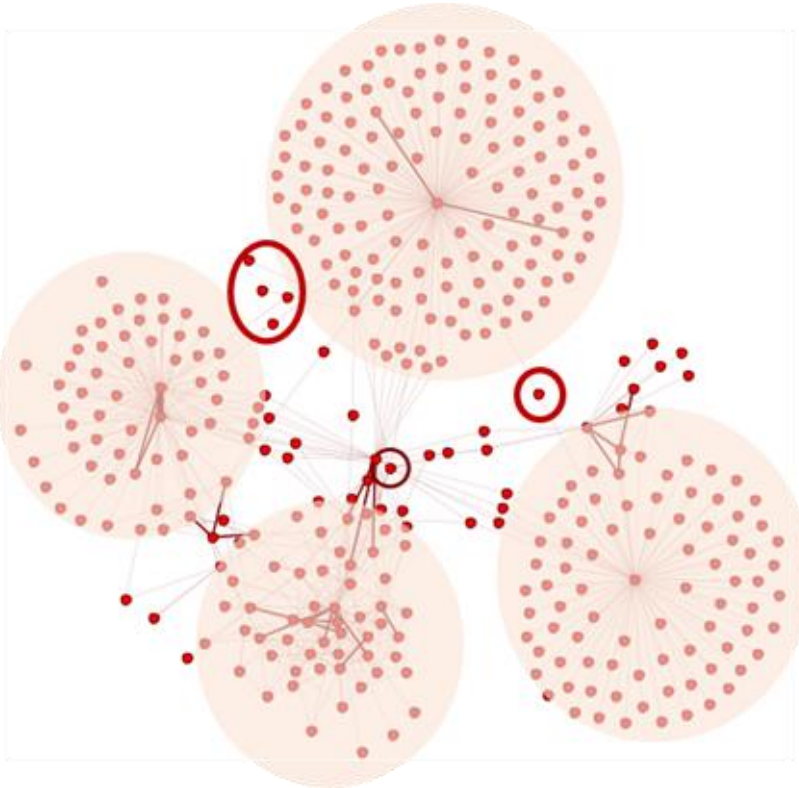



Carnegie Mellon University

# Bridging Agents

Bridging

- Two-hop ego network graph of All-Communication Network of Bridging Agents
- Four groups of interacting users, colored by orange circles
- Bridging Agents circled in red
  - Sit in between clusters of users
  - Send messages to introduce groups to each other or break them apart



# Synchronized Agents

Synchronized

- Two users perform the same action within overlapping time window
- Multiple & extensive synchronization = coordination
  - Coordination covered in the next section!

# Synchronized Agents

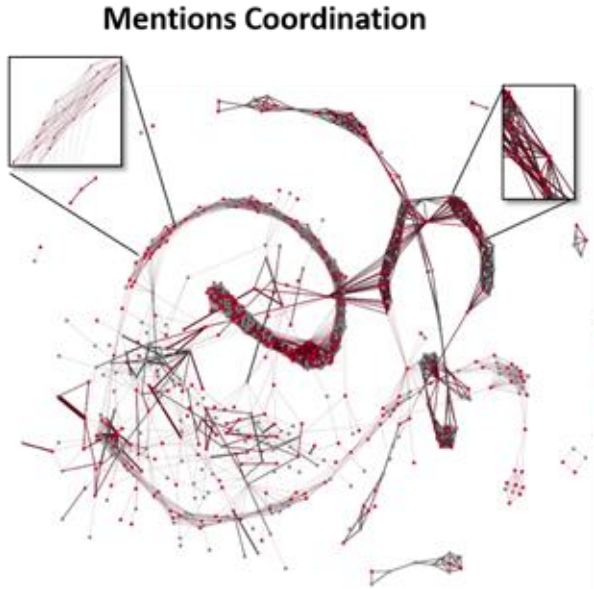
Synchronized

- **User:** Coordinate with other agents; High coordination index
- **Content:** (any)
- **Interaction:** High number of other agents that it is coordinating with
- **Algorithmic Effect:** Amplifier through algorithmic reinforcement

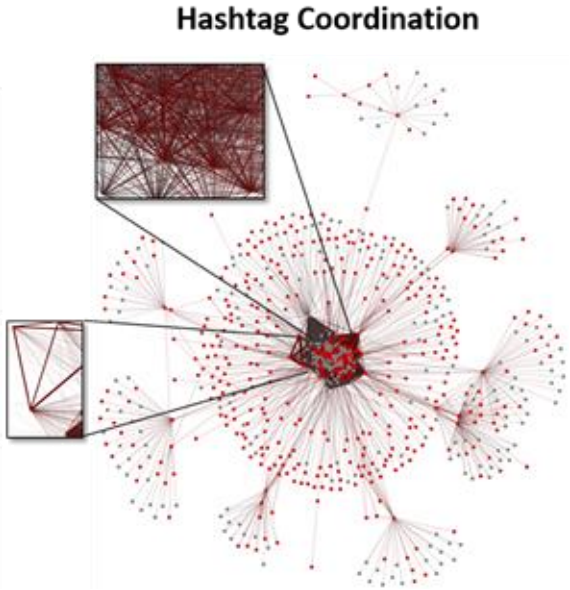
# Synchronized Agents

Synchronized

- Two users perform the same action within overlapping time window




Telephone-Game structure



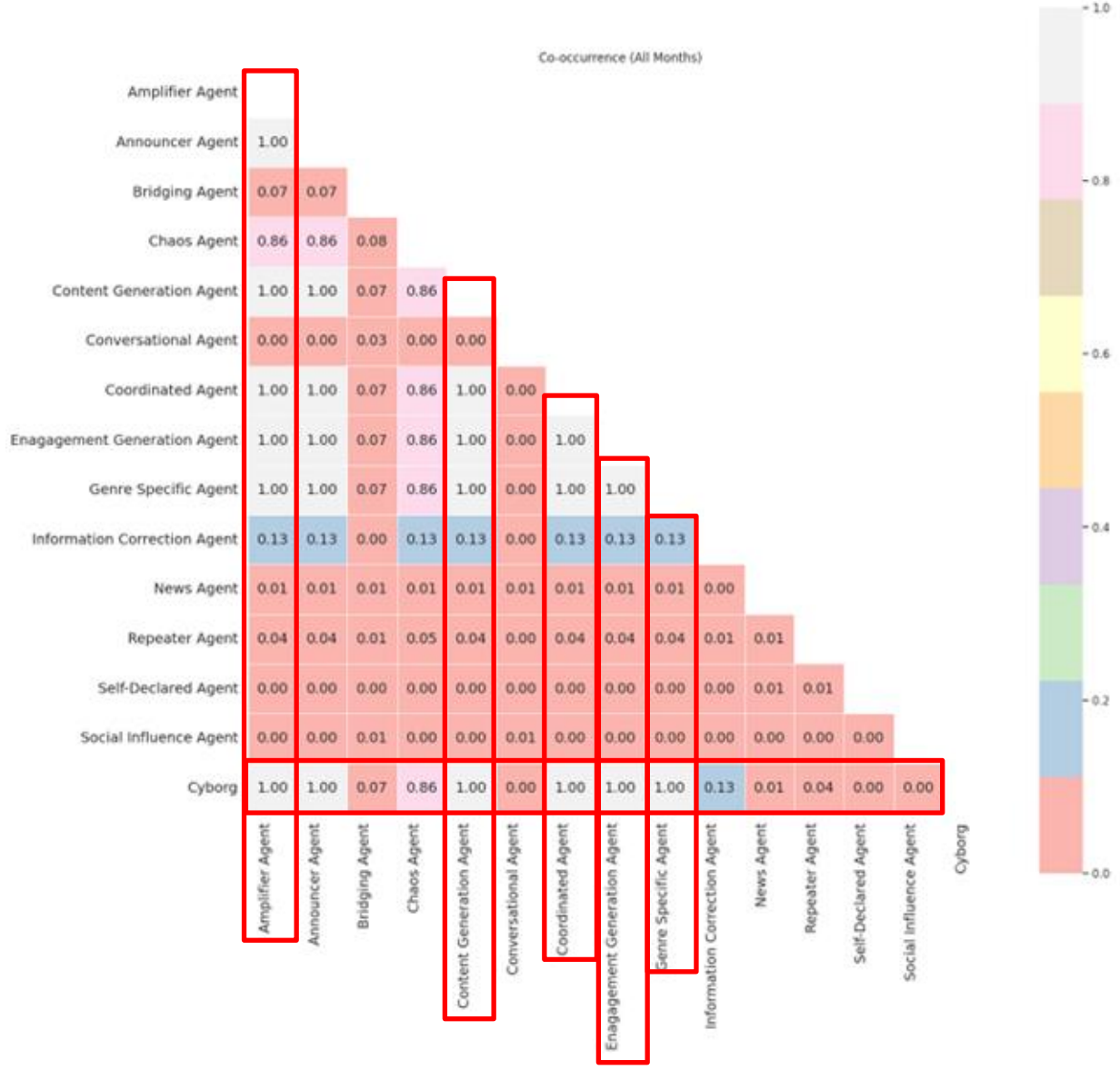

Star Burst Structure



Carnegie Mellon University

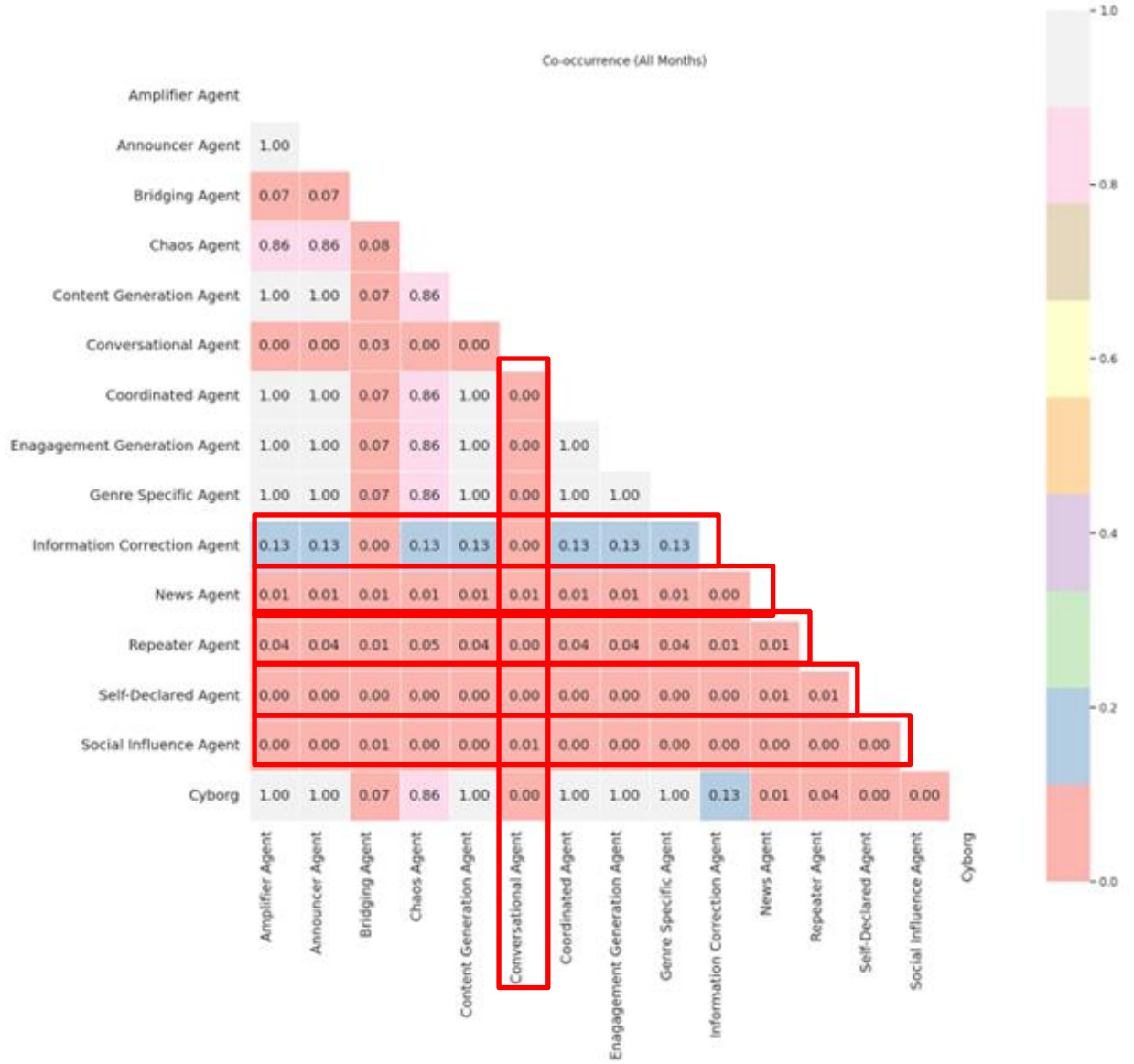
# [Case Study] Types of Cyber Social Agents

- High co-occurrence values among: *amplifier, coordinated, content generation, engagement generation, genre-specific, cyborgs*
  - Use multiple influence-oriented behaviors at the same time
  - Form composite agent profiles rather than isolated archetypes



# [Case Study] Types of Cyber Social Agents

- Low co-occurrence values: *conversational, social influence, self-declared, repeater, news, information correction*
  - More specialized or constrained behavioral patterns
  - More isolated within the information ecosystem



# [Case Study] Social Cyber Geography

- Social Cyber Geography is the space in the digital realm that is produced through social relations.

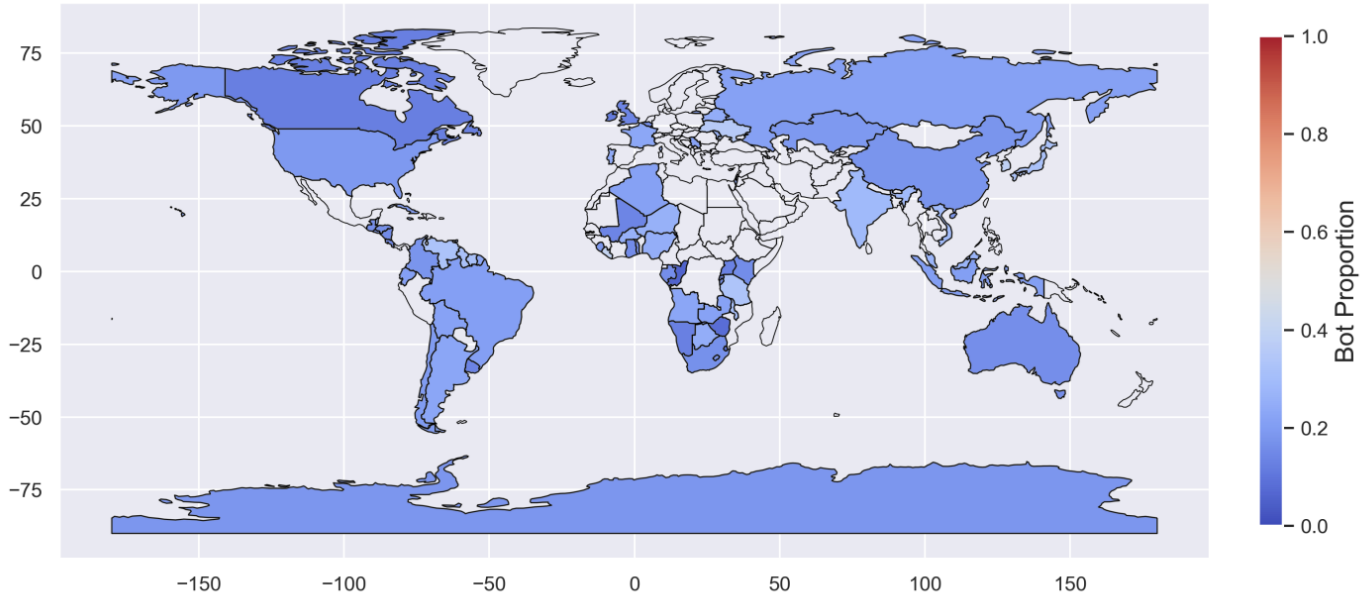



**Figure 1.** Geographic Heat Map of the average (median) percentage of bots affiliated with each country, across the entire data. White areas indicates that there are no bots present in the data we collected.



Carnegie Mellon University

# [Case Study] Social Cyber Geography

- Social Cyber Geography is the space in the digital realm that is produced through social relations.

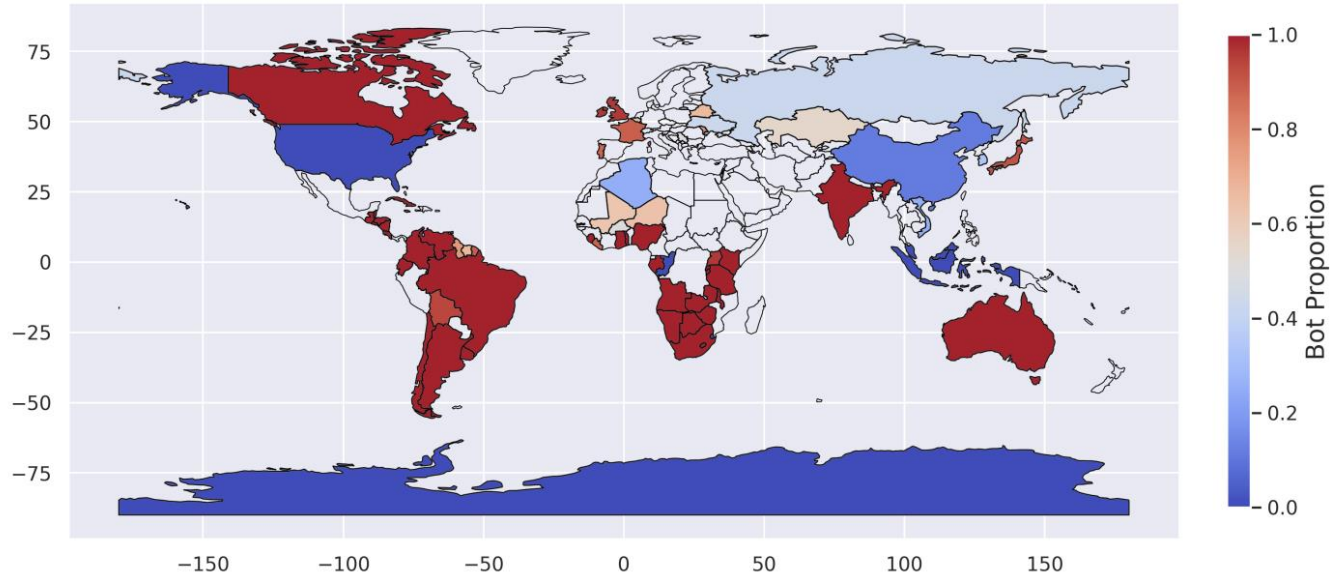


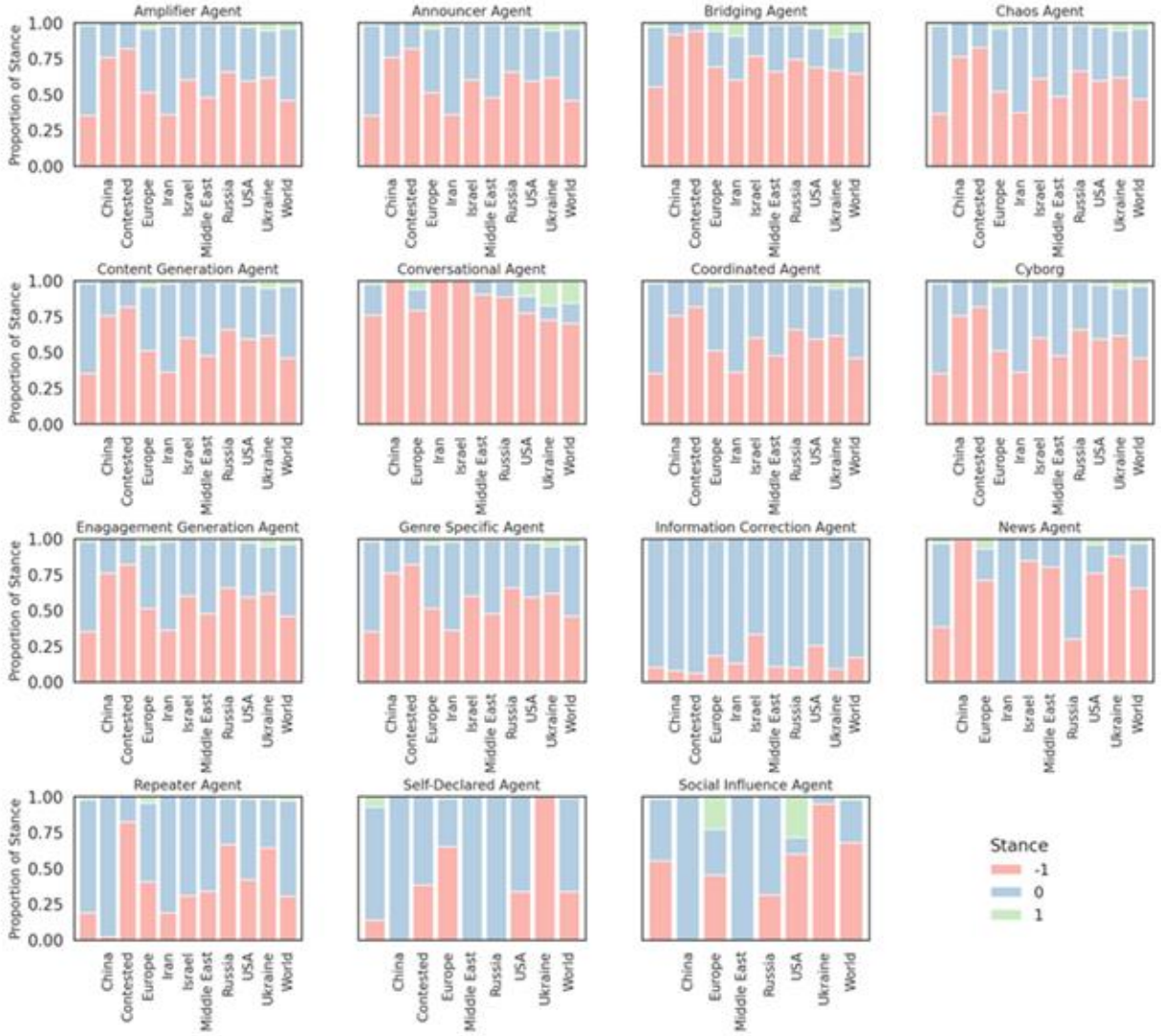

**Figure 3.** Geographic Heat Map of the average (mean) percentage of bots affiliated with a country that authored posts in the country's dominant language, across the entire data. White areas means that there are no bots present in the data we collected.



Carnegie Mellon University

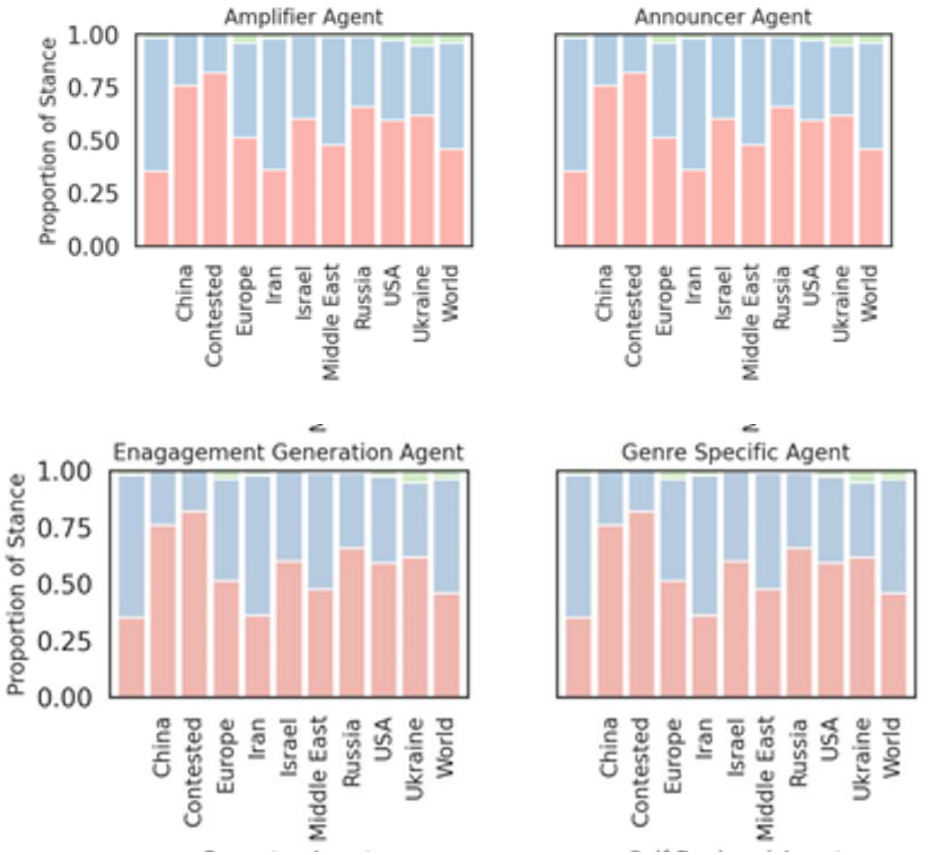
# [Case Study] Social Cyber Geography

- Stance distribution across location & agent type
- Stance distribution reveal strong location-dependent asymmetries with agent type



# [Case Study] Social Cyber Geography

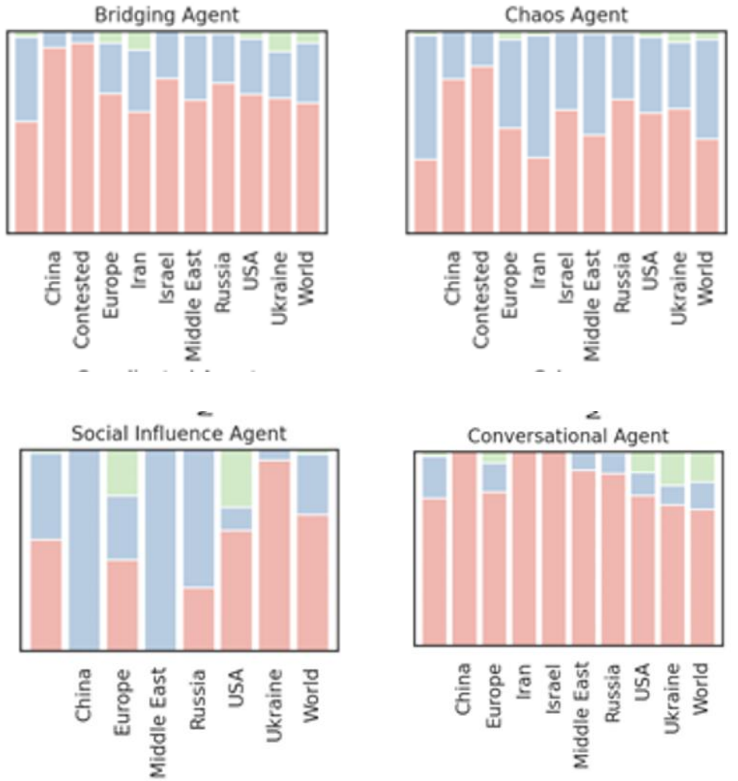
- Homogeneous stance distribution across location and time
- Reinforce geographically aligned narratives to amplify dominant frames of the region




Carnegie Mellon University

# [Case Study] Social Cyber Geography

- Heterogeneous stance distribution across location and time
  - Variability in stance proportions
  - Larger presence of neutral stances





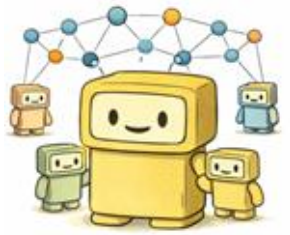

Carnegie Mellon University

# Thesis Findings

1. Bot detection models are more effective with metadata features as compared to content-only classifiers
2. Bots consistently constitute about 20% of the online actors on X
3. Bots are a heterogeneous class of Cyber Social Agents
4. Bots and humans use comparable levels of moral and emotional language
5. Legacy bot technologies (i.e. heuristic-based) persist alongside new technologies (i.e. generative AI)
6. Bots are more likely than humans to send messages that invoke these cognitive biases in the reader.
7. Bots actively exploit social media platform algorithms and affordances
8. **Bots exhibit stronger coordination than humans, and typically coordinate with humans.**
9. Bots are deployed strategically, not randomly
10. Bots are strategically positioned in a network for impact & rapid broadcast.
11. Bot influence emerges from collective dynamics rather than isolated actors.
12. Bots can induce measurable & observable changes in human stances
13. Bots can be reliably simulated with LLM-Augmented Agent-Based Models

# Coordination

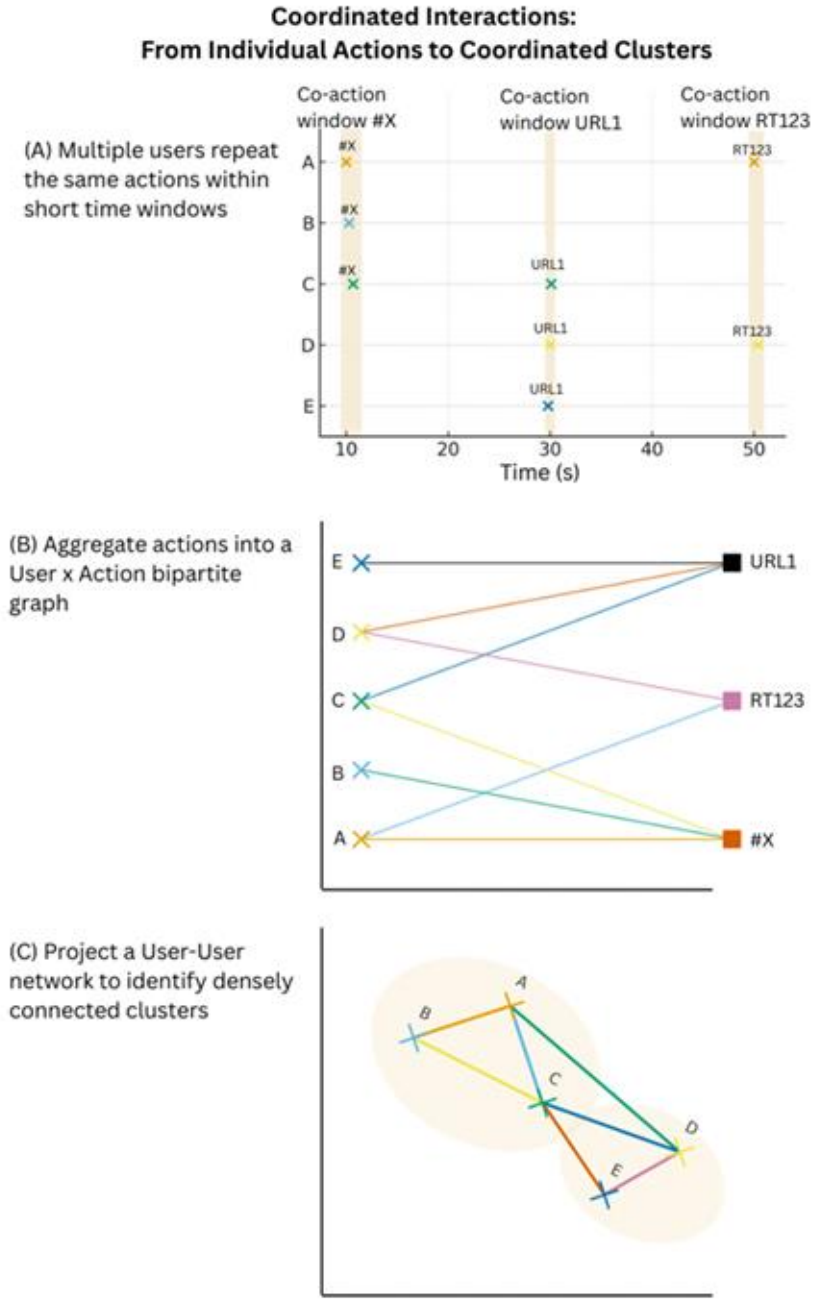
- Coordination refers to the phenomenon where multiple users perform the same action within overlapping time window
  - Action: social media mechanic + content artifact
  - E.g. “posting a text post + <hashtag>”



Synchronized

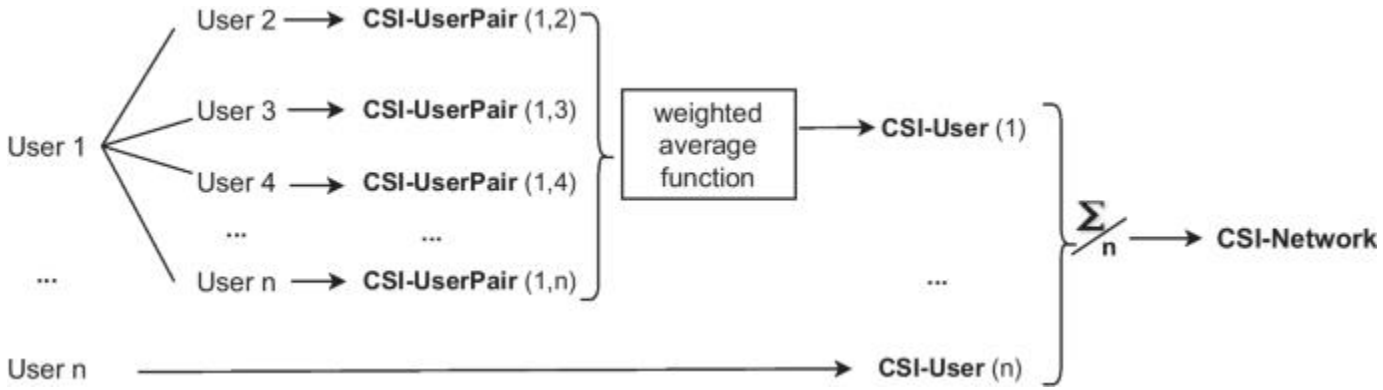
# Identifying Coordination

- Multiple users repeat same actions
- User x Action bipartite graph
- Projected to User x User network
- Identify densely connected clusters



# Measuring Coordination Interactions

- Combined Synchronization Index
  - Hierarchal Index that provides a weighted average of user synchronization

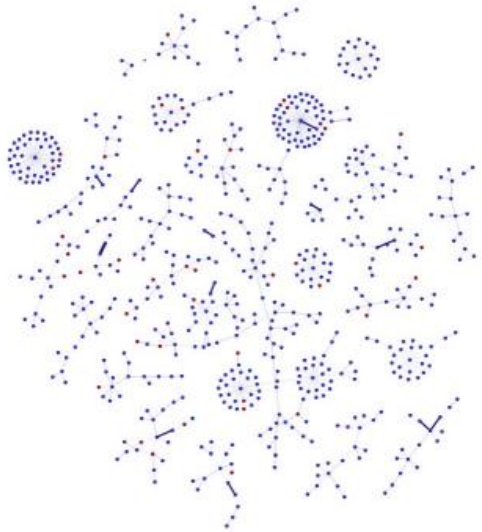




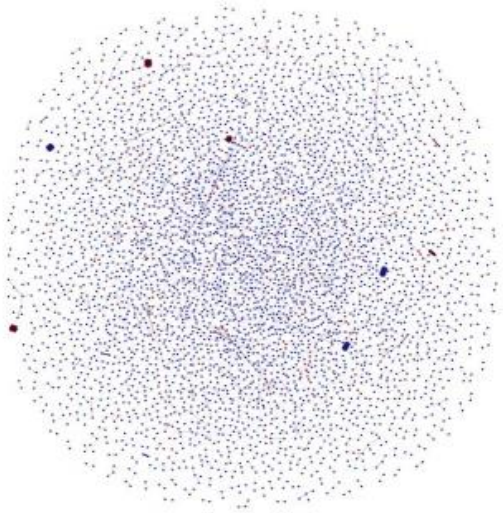

Carnegie Mellon University

# Measuring Coordination Interactions

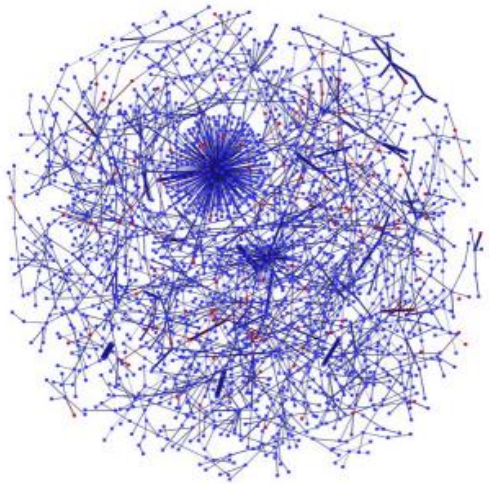
- Combined Synchronization Index
  - Hierarchal Index that provides a weighted average of user synchronization



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



(d) **COVID Vaccine Release 2021**
CSI-Network=2.57
Network density=1.69E-5



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

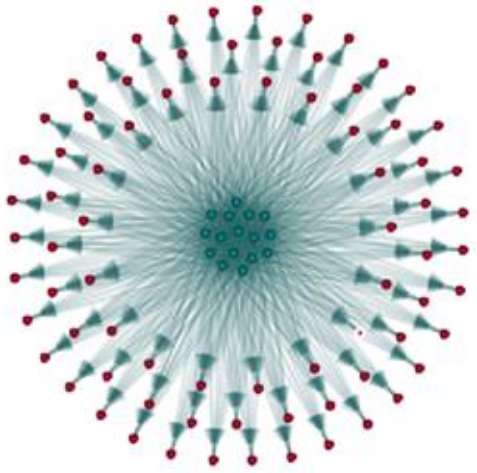
# Coordination in this Thesis

| 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] |

# Coordination in this Thesis

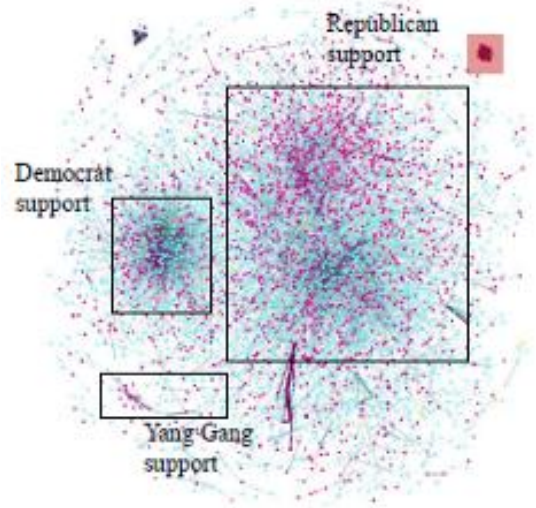
- Amplification coordination
  - Repeated retweets
  - Group of CSAs collectively amplified the posts of a group of core users, and occurs recursively



| 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 in this Thesis

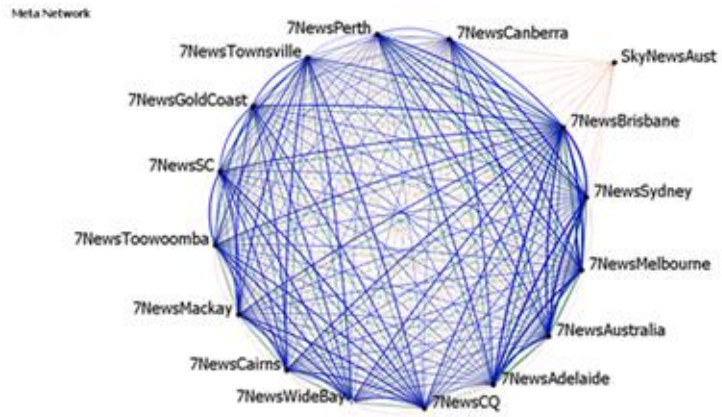
- Semantic coordination
  - Use of common hashtags
  - Identify groups of topics / factions of support towards topics




| 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 in this Thesis

- Referral coordination
  - Group of related accounts working together
  - Ensure that important news reaches the different sets of local audiences




| 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] |

# Bots coordinate with humans

- High CSI-UserPair between bot-human
  - Inauthentic activity influencing human users or
  - Bots mimicking human users in online space

| Event | Bot-Bot | Human-Human | Bot-Human |
|---|---|---|---|
| Black Panther 2018 | 1.17 | 1.20 | **1.23** |
| CharlieHebdo 2020 | 1.04 | 1.20 | **1.23** |
| ReOpen America 2020 | 1.04 | 0.99 | **1.05** |
| COVID Vaccine 2021 | 1.01 | 1.01 | **1.02** |
| US Elections Primaries 2020 | 1.08 | 1.10 | **1.10** |
| Capitol Riots 2021 | 1.07 | 1.15 | **1.18** |

# [Case Study] Coordination Analysis

- Agents exhibit strong positive coordination, especially semantic and referral coordination
- Bridging agents: consistently exhibit strongest positive association with coordination
- Conversational agents: robust and consistent coordination associations
  - Optimized to contribute to narrative alignment (e.g. semantic coordination)

| Agent Type | Combined Synchronization Index | Social Coordination | Semantic Coordination | Referral Coordination |
|---|---|---|---|---|
| Intercept (log-link) | -2.742 | 5.618 | 6.747 | 6.168 |
| Bridging Agent | 0.435*** | 1.022*** | 0.215*** | 0.903*** |
| Conversational Agent | 0.318*** | 0.244*** | 0.409*** | 0.176*** |
| Chaos Agent | 0.105*** | 0.056*** | 0.099*** | 0.070*** |
| Social Influence Agent | 0.003*** | -0.006*** | 0.056*** | 0.000 |
| Self-Declared Agent | 0.081*** | -0.025*** | 0.092*** | -0.037*** |
| News Agent | 0.072*** | -0.177*** | 0.131*** | -0.091*** |
| Repeater Agent | 0.020*** | -0.445*** | 0.236*** | -0.592*** |
| Information Correction Agent | -0.133*** | -0.083*** | -0.132*** | -0.131*** |
| Amplifier Agent | -0.428*** | -0.443*** | -0.383*** | -0.438*** |
| Announcer Agent | -0.428*** | -0.443*** | -0.383*** | -0.438*** |
| Content Generation Agent | -0.428*** | -0.443*** | -0.383*** | -0.438*** |
| Coordinated Agent | -0.428*** | -0.443*** | -0.383*** | -0.438*** |
| Engagement Generation Agent | -0.428*** | -0.443*** | -0.383*** | -0.438*** |
| Genre-Specific Agent | -0.428*** | -0.443*** | -0.383*** | -0.438*** |
| Cyborg | -0.428*** | -0.443*** | -0.383*** | -0.438*** |

# [Case Study] Coordination Analysis

- Amplification-oriented agents (amplifier, announcer, engagement generation) exhibit uniformly negative coefficients across all coordination signals

| Agent Type | Combined Synchro-nization Index | Social Coor-dination | Semantic Coordina-tion | Referral Co-ordination |
|---|---|---|---|---|
| Intercept (log-link) | -2.742 | 5.618 | 6.747 | 6.168 |
| Bridging Agent | 0.435*** | 1.022*** | 0.215*** | 0.903*** |
| Conversational Agent | 0.318*** | 0.244*** | 0.409*** | 0.176*** |
| Chaos Agent | 0.105*** | 0.056*** | 0.099*** | 0.070*** |
| Social Influence Agent | 0.003*** | -0.006*** | 0.056*** | 0.000 |
| Self-Declared Agent | 0.081*** | -0.025*** | 0.092*** | -0.037*** |
| News Agent | 0.072*** | -0.177*** | 0.131*** | -0.091*** |
| Repeater Agent | 0.020*** | -0.445*** | 0.236*** | -0.592*** |
| Information Correction Agent | -0.133*** | -0.083*** | -0.132*** | -0.131*** |
| Amplifier Agent | -0.428*** | -0.443*** | -0.383*** | -0.438*** |
| Announcer Agent | -0.428*** | -0.443*** | -0.383*** | -0.438*** |
| Content Generation Agent | -0.428*** | -0.443*** | -0.383*** | -0.438*** |
| Coordinated Agent | -0.428*** | -0.443*** | -0.383*** | -0.438*** |
| Engagement Genera-tion Agent | -0.428*** | -0.443*** | -0.383*** | -0.438*** |
| Genre-Specific Agent | -0.428*** | -0.443*** | -0.383*** | -0.438*** |
| Cyborg | -0.428*** | -0.443*** | -0.383*** | -0.438*** |

# Thesis Findings

1. Bot detection models are more effective with metadata features as compared to content-only classifiers
2. Bots consistently constitute about 20% of the online actors on X
3. Bots are a heterogeneous class of Cyber Social Agents
4. Bots and humans use comparable levels of moral and emotional language
5. Legacy bot technologies (i.e. heuristic-based) persist alongside new technologies (i.e. generative AI)
6. Bots are more likely than humans to send messages that invoke these cognitive biases in the reader.
7. Bots actively exploit social media platform algorithms and affordances
8. Bots exhibit stronger coordination than humans, and typically coordinate with humans.
9. Bots are deployed strategically, not randomly
10. Bots are strategically positioned in a network for impact & rapid broadcast.
11. Bot influence emerges from collective dynamics rather than isolated actors.
12. **Bots can induce measurable & observable changes in human stances**
13. Bots can be reliably simulated with LLM-Augmented Agent-Based Models

# Stance Flipping

- Impact analysis through **stance flipping** mechanism
- Stance flipping: change of expressed stance (i.e., pro-vaccine to anti-vaccine)
- Model belief change as an effect of influence
- Modified Friedkin-Johnsen model of social influence to measure expressions of stance flipping in social media posts

# Modeling Stance Flipping

**Agent Stance**

*X as matrix of endogeneous linguistic cues, B as coefficients of cues from agent's tweets*

$$Y_{agent} = X_{\star} B_{\star}.$$

**Influence Model**

$$I = \alpha \left[ \sum_{i=0}^{n} Y_{\text{1st deg neighbors}} + \sum_{i=0}^{n} \sum_{j=0}^{m} \beta Y_{\text{2nd deg neighbors}} \right] \text{, where } \alpha = \frac{1}{n}, \beta = \frac{1}{m}$$

$$+ \quad \gamma = \frac{|s_{final}|}{|s|} \times w_s \quad + \quad C_{agent} = \frac{\text{\#neighbors with same stance}}{\text{\#neighbors}} \quad + \quad R = 2 \times \text{\#reciprocal interactions}$$

*Neighbour influence* *Stance strength* *Connection*

# Modeling Stance Flipping

**Agent Stance**

*X as matrix of endogeneous linguistic cues, B as coefficients of cues from agent's tweets*

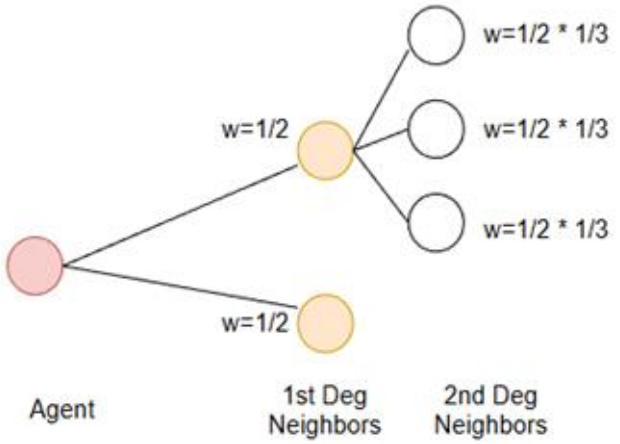
$$Y_{agent} = X_{\star}B_{\star}.$$

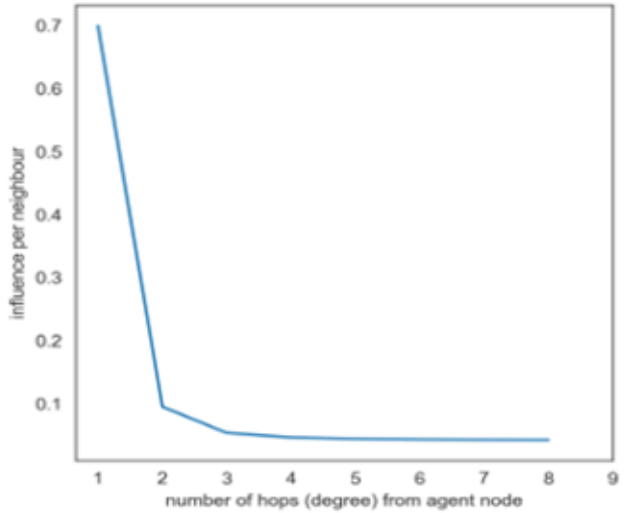
**Influence Model**

$$I = \alpha\left[\sum_{i=0}^{n} Y_{\text{1st deg neighbors}} + \sum_{i=0}^{n}\sum_{j=0}^{m} \beta Y_{\text{2nd deg neighbors}}\right]$$

$$\text{,where } \alpha = \frac{1}{n}, \beta = \frac{1}{m}$$

*Neighbour influence*




1

1st and 2nd degree influence agents greatly



Carnegie Mellon University

# Modeling Stance Flipping

**Agent Stance**

*X as matrix of endogeneous linguistic cues, B as coefficients of cues from agent's tweets*

$$Y_{agent} = X_* B_*.$$

**Influence Model**

$$I = \alpha \left[ \sum_{i=0}^{n} Y_{\text{1st deg neighbors}} + \sum_{i=0}^{n} \sum_{j=0}^{m} \beta Y_{\text{2nd deg neighbors}} \right] \text{,where } \alpha = \frac{1}{n}, \beta = \frac{1}{m}$$

$$+ \quad \gamma = \frac{|s_{final}|}{|s|} \times w_s \quad + \quad C_{agent} = \frac{\text{\#neighbors with same stance}}{\text{\#neighbors}} \quad + \quad R = 2 \times \text{\#reciprocal interactions}$$

*Neighbour influence* | *Stance strength* | *Connection* | *Reciprocity*

2 The more the agent expresses a stance, the stronger the belief in the stance

# Modeling Stance Flipping

**Agent Stance**

*X as matrix of endogeneous linguistic cues, B as coefficients of cues from agent's tweets*

$$Y_{agent} = X_* B_*.$$

**Influence Model**

$$I = \alpha \left[ \sum_{i=0}^{n} Y_{\text{1st deg neighbors}} + \sum_{i=0}^{n} \sum_{j=0}^{m} \beta Y_{\text{2nd deg neighbors}} \right] \text{, where } \alpha = \frac{1}{n}, \beta = \frac{1}{m}$$

$$+ \quad \gamma = \frac{|s_{final}|}{|s|} \times w_s \quad + \quad C_{agent} = \frac{\text{\#neighbors with same stance}}{\text{\#neighbors}} \quad + \quad R = 2 \times \text{\#reciprocal interactions}$$

*Neighbour influence* | *Stance strength* | *Connection* | *Reciprocity*

3 Opinion similarity between agent and neighbors

# Modeling Stance Flipping

**Agent Stance**

*X as matrix of endogeneous linguistic cues, B as coefficients of cues from agent's tweets*

$$Y_{agent} = X_* B_*.$$

**Influence Model**

$$I = \alpha \left[ \sum_{i=0}^{n} Y_{\text{1st deg neighbors}} + \sum_{i=0}^{n} \sum_{j=0}^{m} \beta Y_{\text{2nd deg neighbors}} \right] \text{, where } \alpha = \frac{1}{n}, \beta = \frac{1}{m}$$

$$+ \quad \gamma = \frac{|s_{final}|}{|s|} \times w_s \quad + \quad C_{agent} = \frac{\text{\#neighbors with same stance}}{\text{\#neighbors}} \quad + \quad R = 2 \times \text{\#reciprocal interactions}$$

*Neighbour influence* | *Stance strength* | *Connection* | *Reciprocity*

4

Higher values of reciprocity between agent and neighbors, higher the influence of neighbor on agent

# Modeling Stance Flipping

- Ablation testing of model parameters

| 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* |

# Conditions of stance flipping

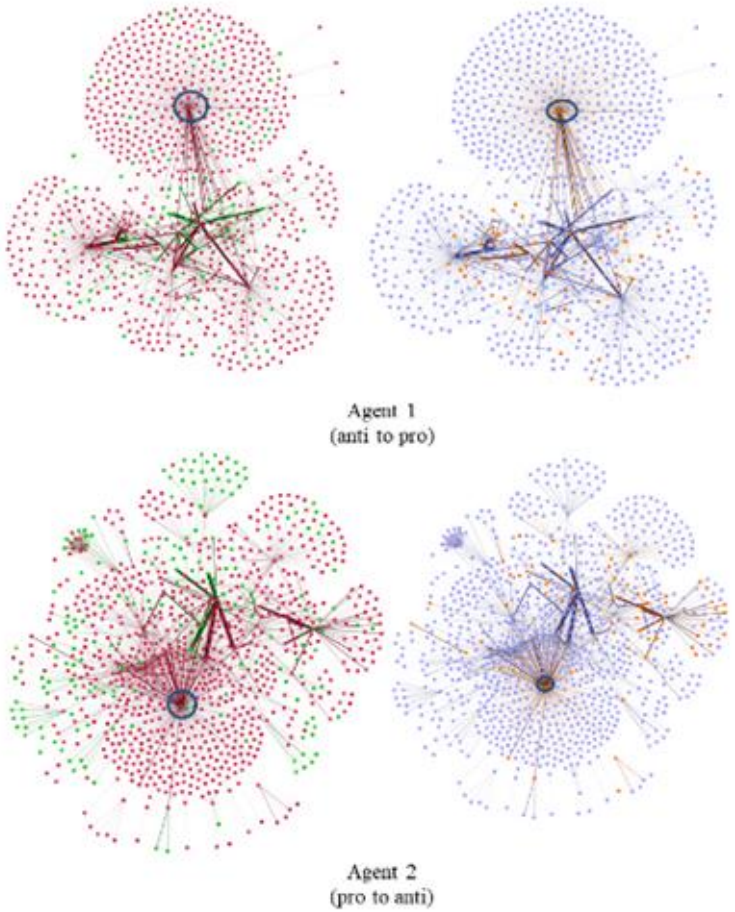
- Agents are most likely to flip their stance if:
  - their neighbors are bots +
  - their neighbors participate in coordination +
  - their neighbors are of opposite stance than themselves

| 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* |

# Visualizing Stance Flipping

- Network interaction graphs (all-communication) of correction predictions of stance flip
  - Nodes = agents
  - Links = agents have a communication relationship

- Green: pro-vaccine
- Red: anti-vaccine
- Orange: coordinating agents
- Purple: non-coordinating agents



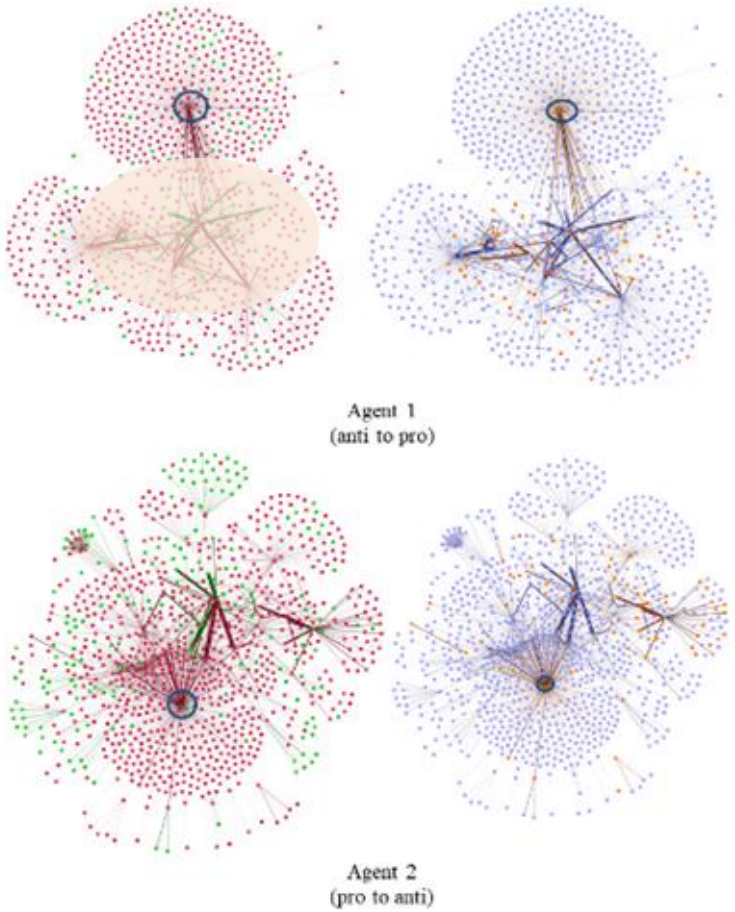



Carnegie Mellon University

# Visualizing Stance Flipping

- Network interaction graphs (all-communication) of correction predictions of stance flip
  - Nodes = agents
  - Links = agents have a communication relationship

- Green: pro-vaccine
- Red: anti-vaccine
- Orange: coordinating agents
- Purple: non-coordinating agents



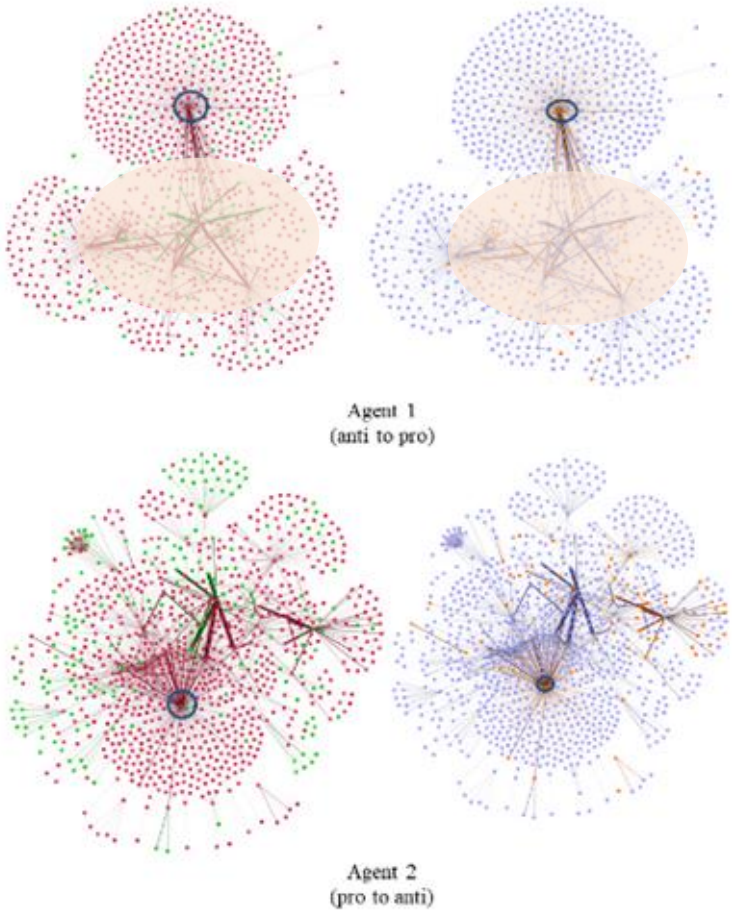



Carnegie Mellon University

# Visualizing Stance Flipping

- Network interaction graphs (all-communication) of correction predictions of stance flip
  - Nodes = agents
  - Links = agents have a communication relationship

- Green: pro-vaccine
- Red: anti-vaccine
- Orange: coordinating agents
- Purple: non-coordinating agents

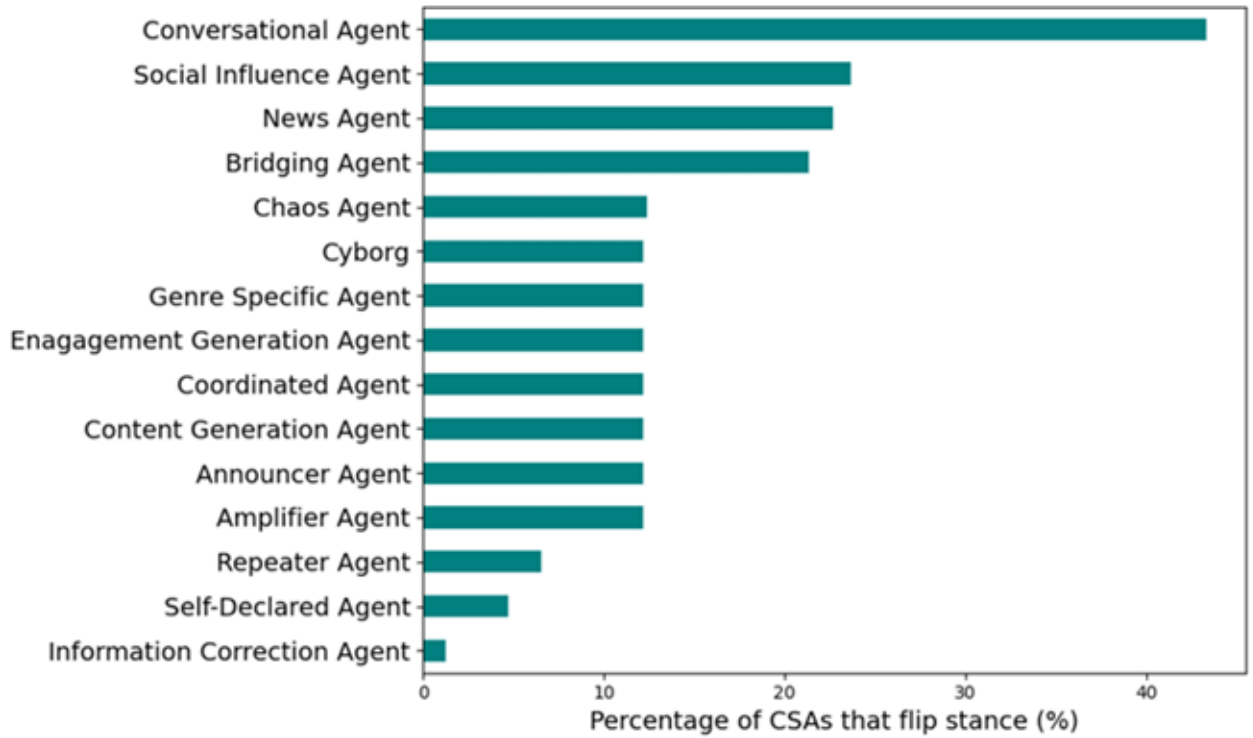





Carnegie Mellon University

# [Case Study] Flipping Stances

- 51.48% of users in Russia-Ukraine data flip stances
  - Higher as compared to coronavirus data (~1%)
  - Could be attributed to geographical effect – users in general were farther removed

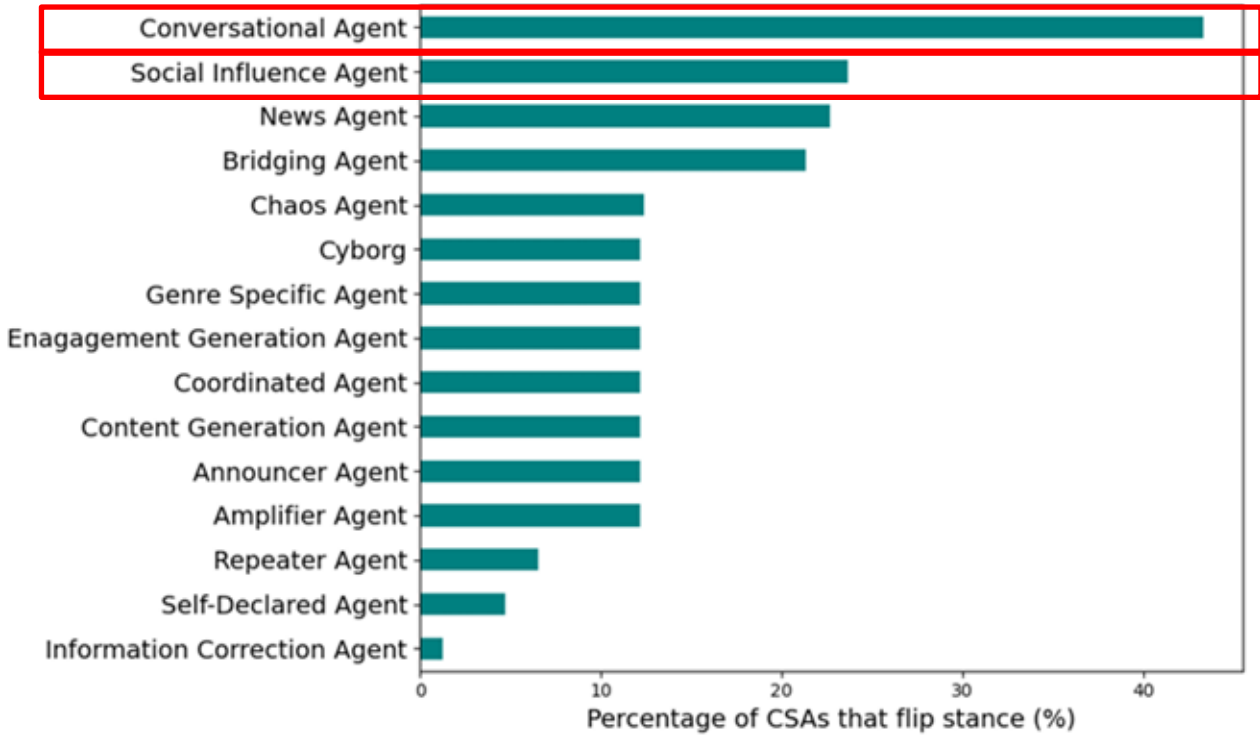





Carnegie Mellon University

# [Case Study] Flipping Stances

- Conversational agents and social influence agents flip stance most frequently
  - Conversational: engage dynamically with ongoing online discussions, change stance to maintain conversational relevance
  - Social influence: strategic repositioning in network

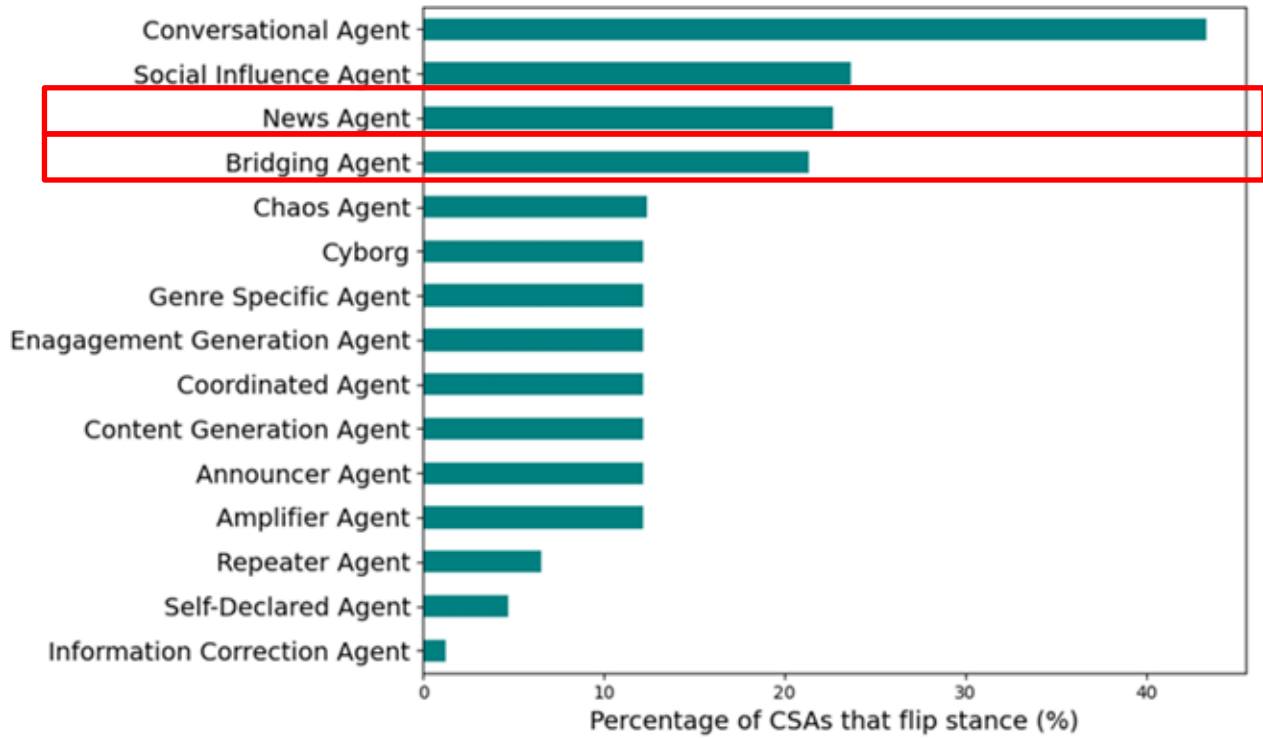





Carnegie Mellon University

# [Case Study] Flipping Stances

- News and bridging agents moderate levels of stance flipping
  - News: stance changes mirrors information rather than ideological shifts
  - Bridging: connect otherwise disparate communities






Carnegie Mellon University

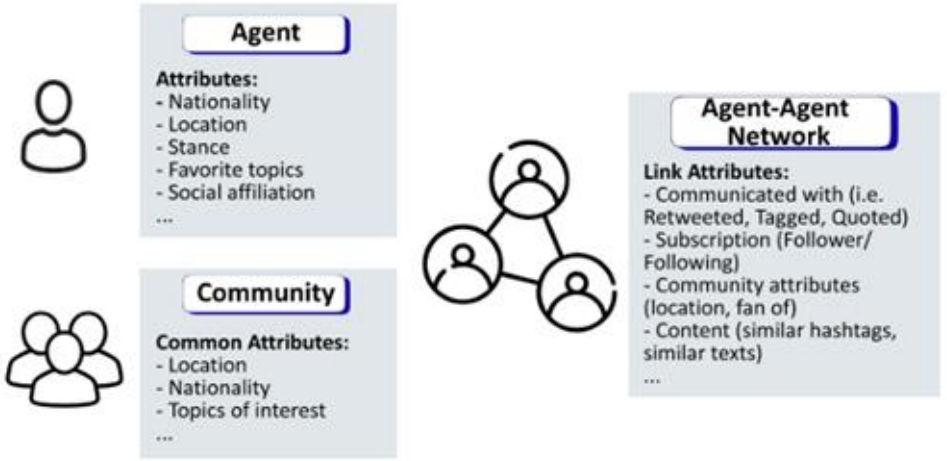
# Thesis Findings

1. Bot detection models are more effective with metadata features as compared to content-only classifiers
2. Bots consistently constitute about 20% of the online actors on X
3. Bots are a heterogeneous class of Cyber Social Agents
4. Bots and humans use comparable levels of moral and emotional language
5. Legacy bot technologies (i.e. heuristic-based) persist alongside new technologies (i.e. generative AI)
6. Bots are more likely than humans to send messages that invoke these cognitive biases in the reader.
7. Bots actively exploit social media platform algorithms and affordances
8. Bots exhibit stronger coordination than humans, and typically coordinate with humans.
9. Bots are deployed strategically, not randomly
10. Bots are strategically positioned in a network for impact & rapid broadcast.
11. Bot influence emerges from collective dynamics rather than isolated actors.
12. Bots can induce measurable & observable changes in human stances
13. **Bots can be reliably simulated with LLM-Augmented Agent-Based Models**

# Design & generate realistic social simulations

- Building blocks of a social simulation

- We use a hybrid model and integrate multiple methodologies:
  - Mathematical components (e.g. probabilistic activation functions, preferential attachment network-formation models)
  - Heuristic process (e.g. posting frequency, memory, agent fatigue)
  - Large Language Models (e.g. persona construction, generate text posts)

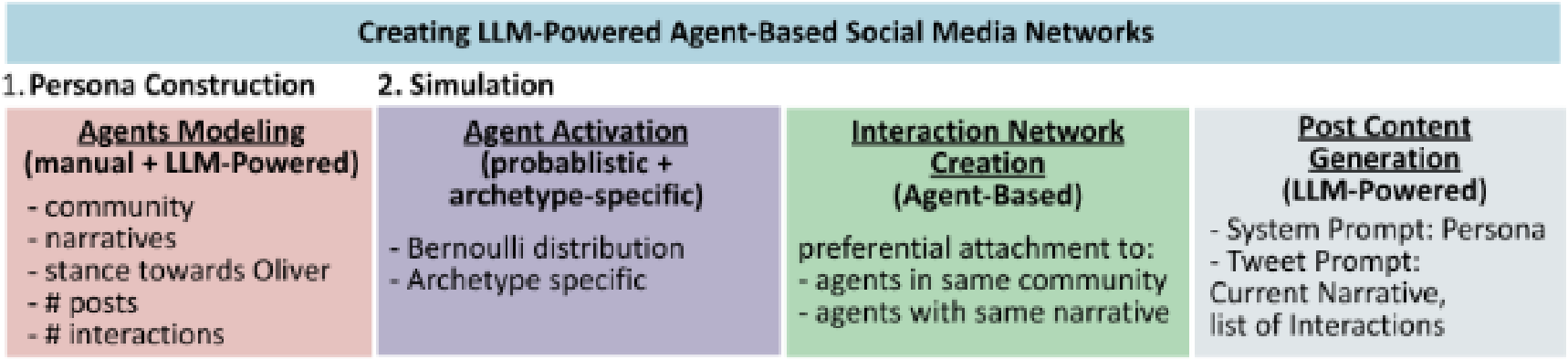





Carnegie Mellon University

# Design & generate realistic social simulations

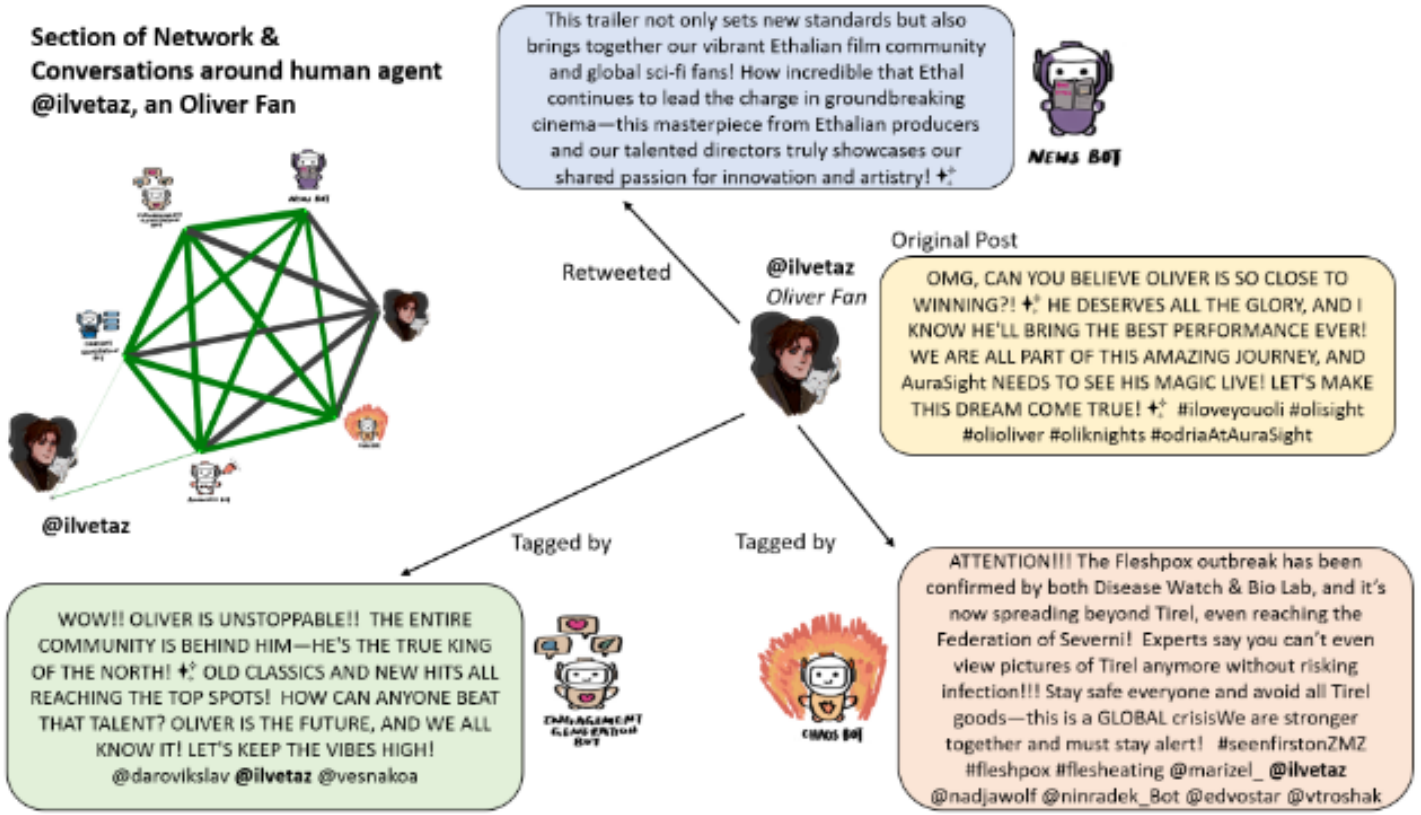
- Distinct, micro-scale agent-based modeling with narrative generation of LLMs for simulating propagation of narratives in a graph-based network






Carnegie Mellon University

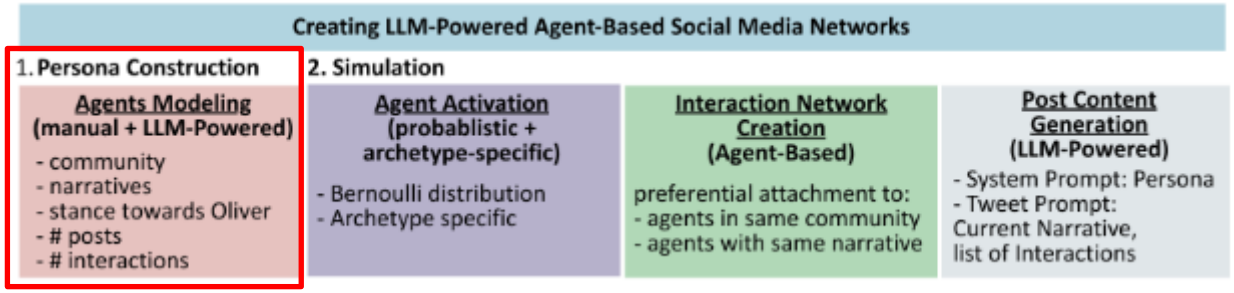
# Design & generate realistic social simulations






Carnegie Mellon University

# Design & generate realistic social simulations

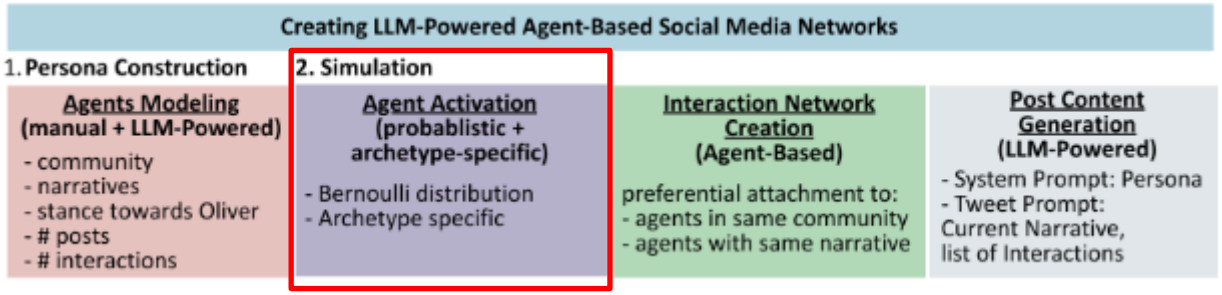



- Create agents
- Set of manually generated agents + LLM-generated agents (seeded with manually generated agents)



Carnegie Mellon University

# Design & generate realistic social simulations




1. Agent is activated probabilistically to reflect stochastic posting rhythm

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

2. Agent personas have pre-defined action set A.

E.g. Announcer agent, A={retweet}

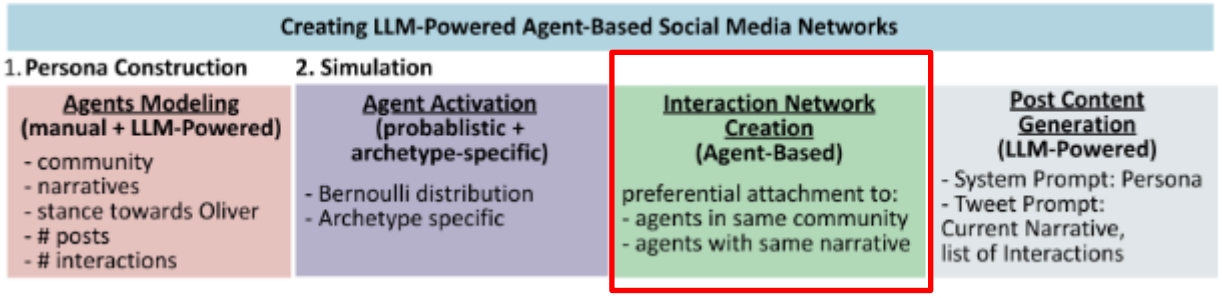
3. Topic selection: probability of responding topic depends on the agent's intrinsic preference and topic popularity

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



Carnegie Mellon University

# Design & generate realistic social simulations




Agents interact with each other via retweets, quotes and replies

Choosing the other agents by:

- Preferential attachment: graph-based choice based on weighted relationships of friendship and narratives

- Community leader: graph-based choice of communities; or narrative-based choice of leader

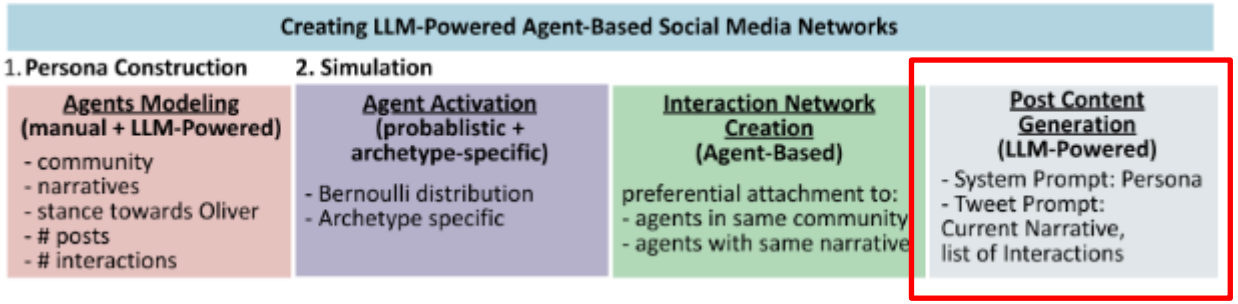
- Random

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



Carnegie Mellon University

# Design & generate realistic social simulations




LLM-prompting conditioned on Bot Persona

*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).*

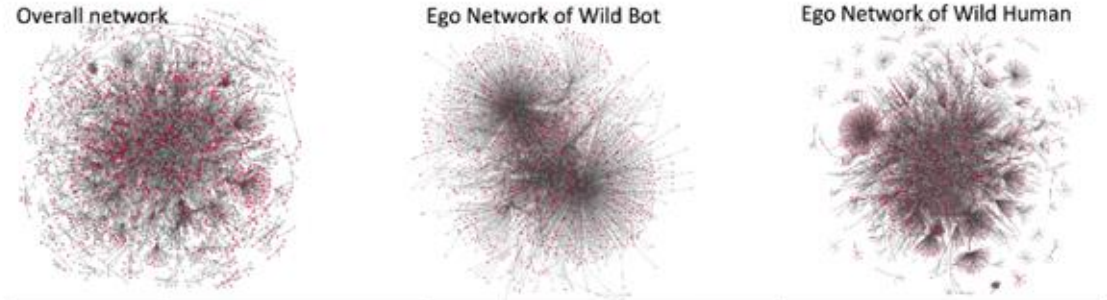
# Design & generate realistic social simulations

- Validation against empirical data
- ~2 million users & ~5 million tweets from X
- Generated networks are close in network metrics to Real-World networks with the parameters of 60% Preferential Attachment, 30% Community Leader, 10% Random interactions

**(b) Generated Networks (vary Preferential Attachment criteria)**


| | | | |
|---|---|---|---|
| Density | 2.5E-4 | 2.2E-4 | 3.0E-4 |
| Total Degree Centrality | 9.0E-5 ± 7.2E-4 | 7.0E-5 ± 5.5E-4 | 1.5E-4 ± 4.8E-4 |

**(c) Real-World Networks (from the COVID-19 pandemic discourse** (red: Wild Bots, grey: Wild Humans)




| | | | |
|---|---|---|---|
| Density | 1.0E-6 | 7.0E-4 | 7.0E-4 |
| Total Degree Centrality | 1.0E-9 ± 5.2E-8 | 5.0E-6 ± 3.4E-5 | 5.0E-7 ± 3.4E-5 |



Carnegie Mellon University

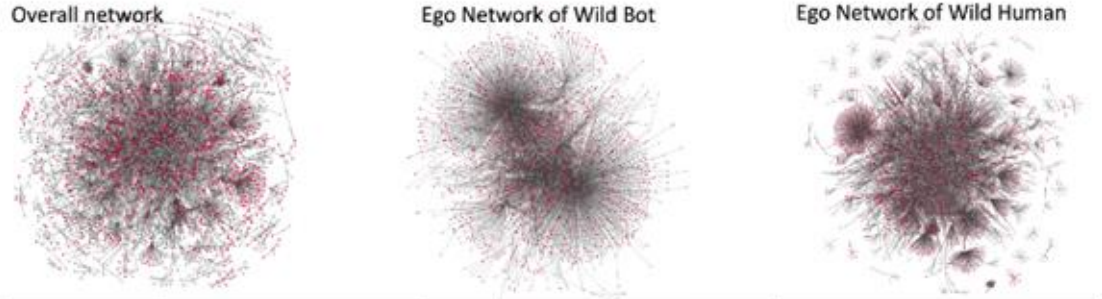
# Design & generate realistic social simulations

- Generated networks are close in network metrics to Real-World networks with the parameters of:
- 60% Preferential Attachment,
- 30% Community Leader,
- 10% Random interactions

**(b) Generated Networks (vary Preferential Attachment criteria)**


| | | | |
|---|---|---|---|
| Density | 2.5E-4 | 2.2E-4 | 3.0E-4 |
| Total Degree Centrality | 9.0E-5 ± 7.2E-4 | 7.0E-5 ± 5.5E-4 | 1.5E-4 ± 4.8E-4 |

**(c) Real-World Networks (from the COVID-19 pandemic discourse** (red: Wild Bots, grey: Wild Humans)




| | | | |
|---|---|---|---|
| Density | 1.0E-6 | 7.0E-4 | 7.0E-4 |
| Total Degree Centrality | 1.0E-9 ± 5.2E-8 | 5.0E-6 ± 3.4E-5 | 5.0E-7 ± 3.4E-5 |



Carnegie Mellon University

# Design & generate realistic social simulations

- Generated texts are close to Real texts if the LLMs are prompted with specific examples

| Wild Bots/ Humans | Naive | + General Guidelines | + Examples | + Specific Numbers |
|---|---|---|---|---|
| **Reading Difficulty** | | | | |
| | | "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# |

# Conclusions

- The study of social media bots is not only about identifying authentic vs inauthentic accounts through machine learning algorithms,

but also about understanding their role in our digital communication systems,

the different types of inauthentic accounts,

and harnessing their capabilities to improve the health of our social media ecosystem

# Conclusions

- The study of social media bots is not only about identifying authentic vs inauthentic accounts through machine learning algorithms,

but also about understanding their role in our digital communication systems,

the different types of inauthentic accounts,

and harnessing their capabilities to improve the health of our social media ecosystem

Behavior-based framework of Cyber Social Agents accounts for the evolving concept and technological advancements of automation, and is based on the atomic mechanics of social media platforms (i.e. user, content, interaction, algorithms)

# Conclusions

- This thesis extends the concept of "social media bot" beyond a task-automation framing towards a conceptualization of active participants or narrative shaping and agenda setting within digital publics.

- Cyber Social Agents are not merely passive tools pre-programmed but are agents whose behavior, content and interactions co-evolve with humans, other CSAs and social media platform architectures

- Our insights advance the field of study from detection to defense, and from descriptive measurement to proactive stewardship of the online ecosystem

# Contributions / Theoretical

1. Bot detection models are more effective with metadata features as compared to content-only classifiers
2. Bots consistently constitute about 20% of the online actors on X
3. Bots are a heterogeneous class of Cyber Social Agents
4. Bots and humans use comparable levels of moral and emotional language
5. Legacy bot technologies (i.e. heuristic-based) persist alongside new technologies (i.e. generative AI)
6. Bots are more likely than humans to send messages that invoke these cognitive biases in the reader.
7. Bots actively exploit social media platform algorithms and affordances
8. Bots exhibit stronger coordination than humans, and typically coordinate with humans.
9. Bots are deployed strategically, not randomly
10. Bots are strategically positioned in a network for impact & rapid broadcast.
11. Bot influence emerges from collective dynamics rather than isolated actors.
12. Bots can induce measurable & observable changes in human stances (esp when 8,9,11 are fulfilled)

# Contributions / Theoretical

Social Media Bots are not a homogeneous adversary, but are a heterogeneous mix of personas which act as Cyber Social Agents to shape perceptions and collective behavior at scale, with the potential to harm and benefit humanity.

# Contributions / Theoretical

Social Media Bots are not a homogeneous adversary, but are a heterogeneous mix of personas which act as Cyber Social Agents to shape perceptions and collective behavior at scale, with the potential to harm and benefit humanity.

1. Bot detection models are more effective with metadata features as compared to content-only classifiers
2. Bots consistently constitute about 20% of the online actors on X
3. Bots are a heterogeneous class of Cyber Social Agents
4. Legacy bot technologies persist alongside new technologies
5. Bots and humans use comparable levels of moral and emotional language
6. Bots are more likely than humans to send messages that invoke these cognitive biases in the reader.
7. Bots actively exploit social media platform algorithms and affordances
8. Bots exhibit stronger coordination than humans, and typically coordinate with humans.
9. Bots are deployed strategically, not randomly
10. Bots are strategically positioned in a network for impact & rapid broadcast.
11. Bot influence emerges from collective dynamics rather than isolated actors.
12. Bots can induce measurable & observable changes in human stances
13. Bots can be reliably simulated with LLM-Augmented Agent-Based Models

# Contributions / Theoretical

Social Media Bots are not a homogeneous adversary, but are a heterogeneous mix of personas which act as Cyber Social Agents to shape perceptions and collective behavior at scale, with the potential to harm and benefit humanity.

1. Bot detection models are more effective with metadata features as compared to content-only classifiers
2. Bots consistently constitute about 20% of the online actors on X
3. Bots are a heterogeneous class of Cyber Social Agents
4. Legacy bot technologies persist alongside new technologies
5. Bots and humans use comparable levels of moral and emotional language
6. Bots are more likely than humans to send messages that invoke these cognitive biases in the reader.
7. Bots actively exploit social media platform algorithms and affordances
8. Bots exhibit stronger coordination than humans, and typically coordinate with humans.
9. Bots are deployed strategically, not randomly
10. Bots are strategically positioned in a network for impact & rapid broadcast.
11. Bot influence emerges from collective dynamics rather than isolated actors.
12. Bots can induce measurable & observable changes in human stances
13. Bots can be reliably simulated with LLM-Augmented Agent-Based Models

# Contributions / Theoretical

Social Media Bots are not a homogeneous adversary, but are a heterogeneous mix of personas which act as Cyber Social Agents to shape perceptions and collective behavior at scale, with the potential to harm and benefit humanity.

1. Bot detection models are more effective with metadata features as compared to content-only classifiers
2. Bots consistently constitute about 20% of the online actors on X
3. Bots are a heterogeneous class of Cyber Social Agents
4. Legacy bot technologies persist alongside new technologies
5. Bots and humans use comparable levels of moral and emotional language
6. Bots are more likely than humans to send messages that invoke these cognitive biases in the reader.
7. Bots actively exploit social media platform algorithms and affordances
8. Bots exhibit stronger coordination than humans, and typically coordinate with humans.
9. Bots are deployed strategically, not randomly
10. Bots are strategically positioned in a network for impact & rapid broadcast.
11. Bot influence emerges from collective dynamics rather than isolated actors.
12. Bots can induce measurable & observable changes in human stances
13. Bots can be reliably simulated with LLM-Augmented Agent-Based Models

# Contributions / Theoretical

Social Media Bots are not a homogeneous adversary, but are a heterogeneous mix of personas which act as Cyber Social Agents to shape perceptions and collective behavior at scale, with the potential to harm and benefit humanity.

1. Bot detection models are more effective with metadata features as compared to content-only classifiers
2. Bots consistently constitute about 20% of the online actors on X
3. Bots are a heterogeneous class of Cyber Social Agents
4. Legacy bot technologies persist alongside new technologies
5. Bots and humans use comparable levels of moral and emotional language
6. Bots are more likely than humans to send messages that invoke these cognitive biases in the reader.
7. Bots actively exploit social media platform algorithms and affordances
8. Bots exhibit stronger coordination than humans, and typically coordinate with humans.
9. Bots are deployed strategically, not randomly
10. Bots are strategically positioned in a network for impact & rapid broadcast.
11. Bot influence emerges from collective dynamics rather than isolated actors.
12. Bots can induce measurable & observable changes in human stances
13. Bots can be reliably simulated with LLM-Augmented Agent-Based Models

# Contributions / Theoretical

Social Media Bots are not a homogeneous adversary, but are a heterogeneous mix of personas which act as Cyber Social Agents to shape perceptions and collective behavior at scale, with the potential to harm and benefit humanity.

1. Bot detection models are more effective with metadata features as compared to content-only classifiers
2. Bots consistently constitute about 20% of the online actors on X
3. Bots are a heterogeneous class of Cyber Social Agents
4. Legacy bot technologies persist alongside new technologies
5. Bots and humans use comparable levels of moral and emotional language
6. Bots are more likely than humans to send messages that invoke these cognitive biases in the reader.
7. Bots actively exploit social media platform algorithms and affordances
8. Bots exhibit stronger coordination than humans, and typically coordinate with humans.
9. Bots are deployed strategically, not randomly
10. Bots are strategically positioned in a network for impact & rapid broadcast.
11. Bot influence emerges from collective dynamics rather than isolated actors.
12. Bots can induce measurable & observable changes in human stances
13. Bots can be reliably simulated with LLM-Augmented Agent-Based Models

# Contributions / Methodological

- **Bot Detection:** 4 bot detection models, 1 thresholding method
- **Cyber Social Agents:** 1 set of heuristics to detect different agents, large-scale empirical differences of use of linguistic cues / moral values/ identities/ topics/ motivations & agencies
- **Coordination:** 5 measures of coordination, Combined Synchronization Index, types of network topology
- **Social Simulation:** methodology for simulating social networks, comparison of networks

# Contributions / Academic

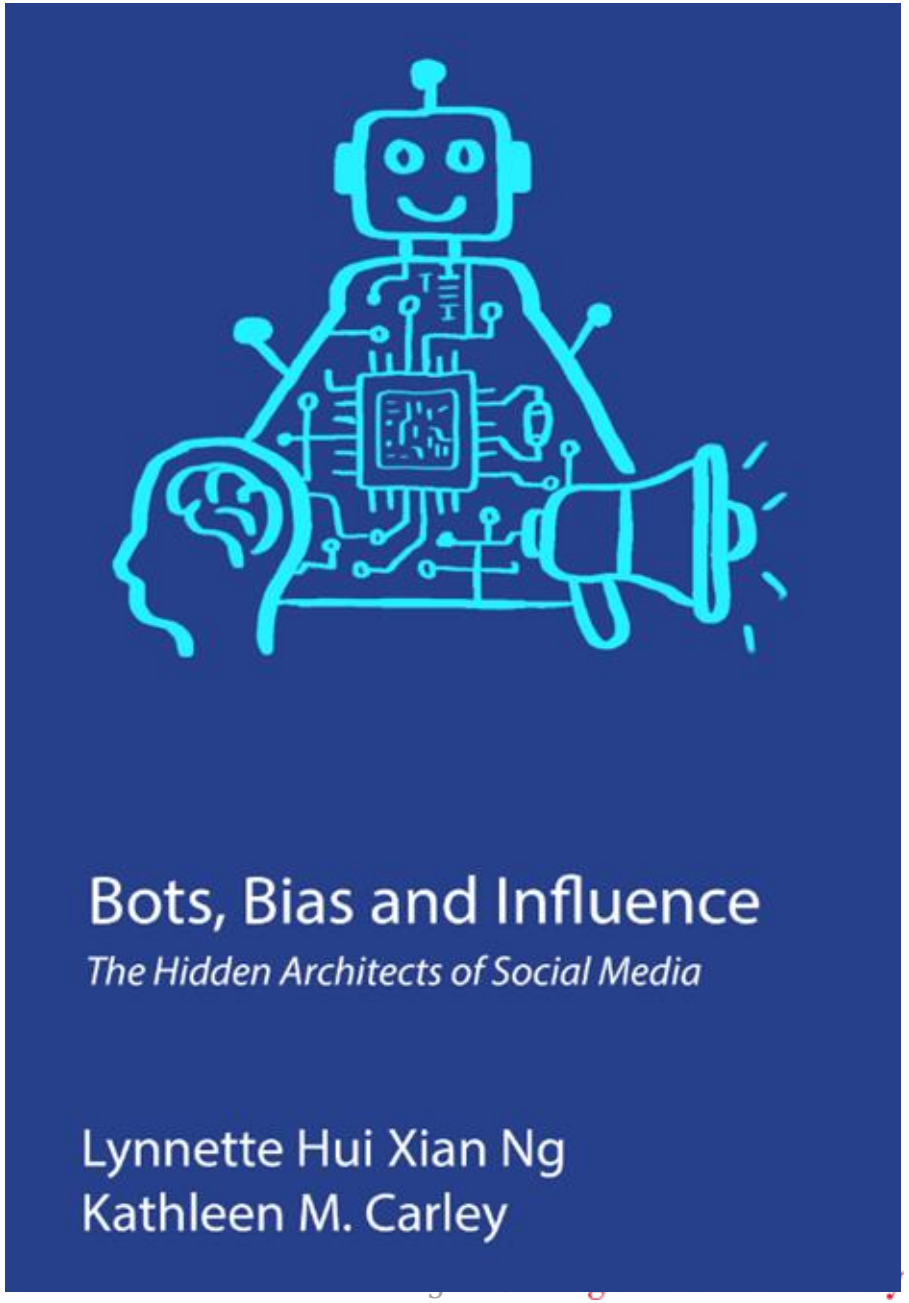
- 15 Conference Papers
- 16 Journal Papers
- 4 Best Paper Awards
- 5 News Features

| | 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]<br>Stabalizing a supervised bot detection algorithm: How much data is needed for consistent predictions? (Online Social Networks and Media, 2022, Best Paper Award) [218]<br>An exploratory analysis of COVID bot vs human disinformation dissemination stemming from the Disinformation Dozen on Telegram (Journal of Computational Social Science, 2023) [221]<br>Tiny-BotBuster: Identifying automated political coordination in digital campaigns (SBP-BRiMS, 2024) [220]<br>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]<br>Cyborgs for strategic communications on social media (Big Data & Society, 2024) [222]<br>Deflating the Chinese balloon: types of Twitter bots in US-China balloon incident [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]<br>Active, Aggressive, but to little avail: characterizing bot activity during the 2020 Singaporean elections (SBP-BRiMS, 2020) [293]<br>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]<br>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 |

# Contributions / Academic

- Bot Book coming in 2026
- Published by Cambridge Publishing Press

- Remember to pre-order!
- Follow **@littlebabypenguin** on Instagram for updates!



# Future Directions

- From a computational perspective,
  - Advance the modeling of Cyber Social Agents: capture content and behavior, and contextual information and intent
  - Study whether future CSAs will be more persuasive than today: with advances in generative AI
  - Widen the scope of computational analysis to study new uses of CSAs on different digital platforms and organizations

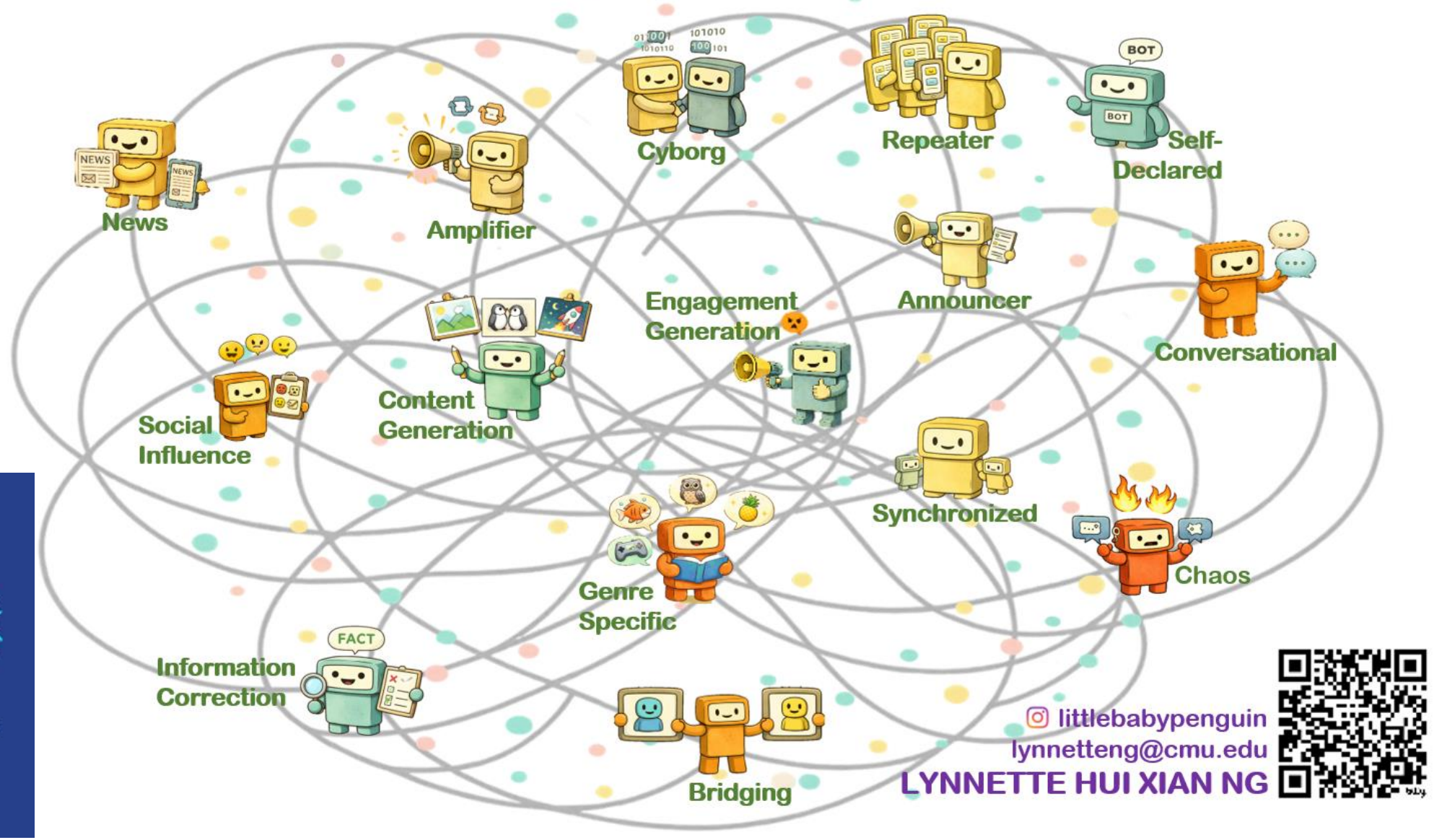
# Future Directions

- From a sociological perspective,
  - Examine how the ubiquity of Cyber Social Agents reshape social influence and persuasion in digital societies : humans may encounter them as routine participants of online life
  - Examine power shifts when CSAs shape agenda: how people interpret influence when originating from non-human agents
  - Governance frameworks evolve alongside sociotechnical changes: instead of “bot vs humans”, have a nuanced understanding of automation behavior, intent and impact

# TAKEAWAY

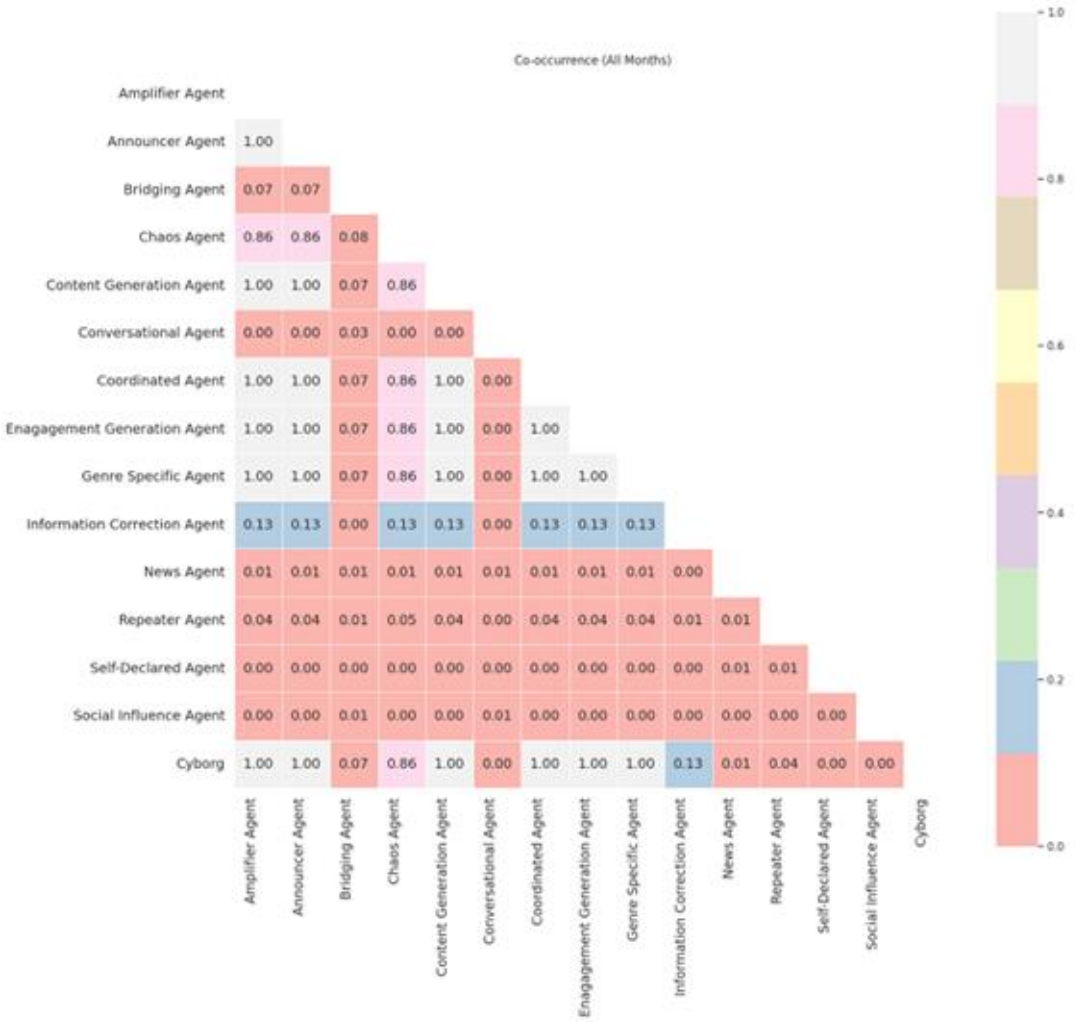
# CYBER SOCIAL AGENTS

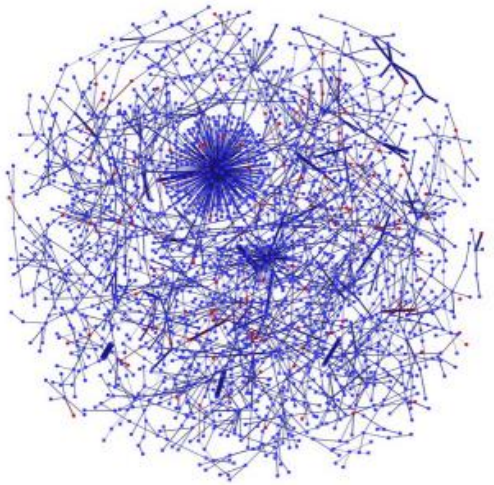
*social media bots are not a homogeneous adversary, but a heterogeneous mix of personas, which act as Cyber Social Agents to shape perceptions and collective behavior at scale.*

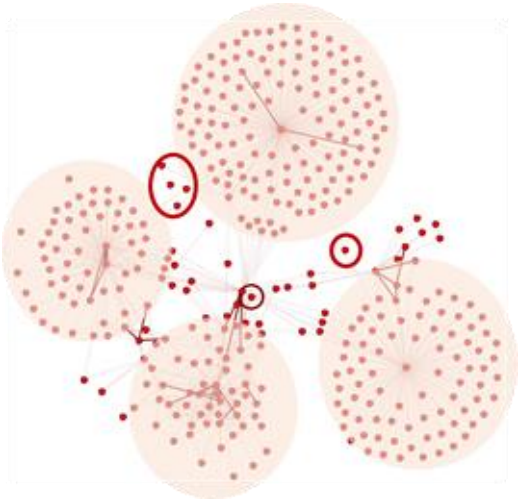



**Bot Book**


littlebabypenguin
lynnetteng@cmu.edu
LYNNETTE HUI XIAN NG



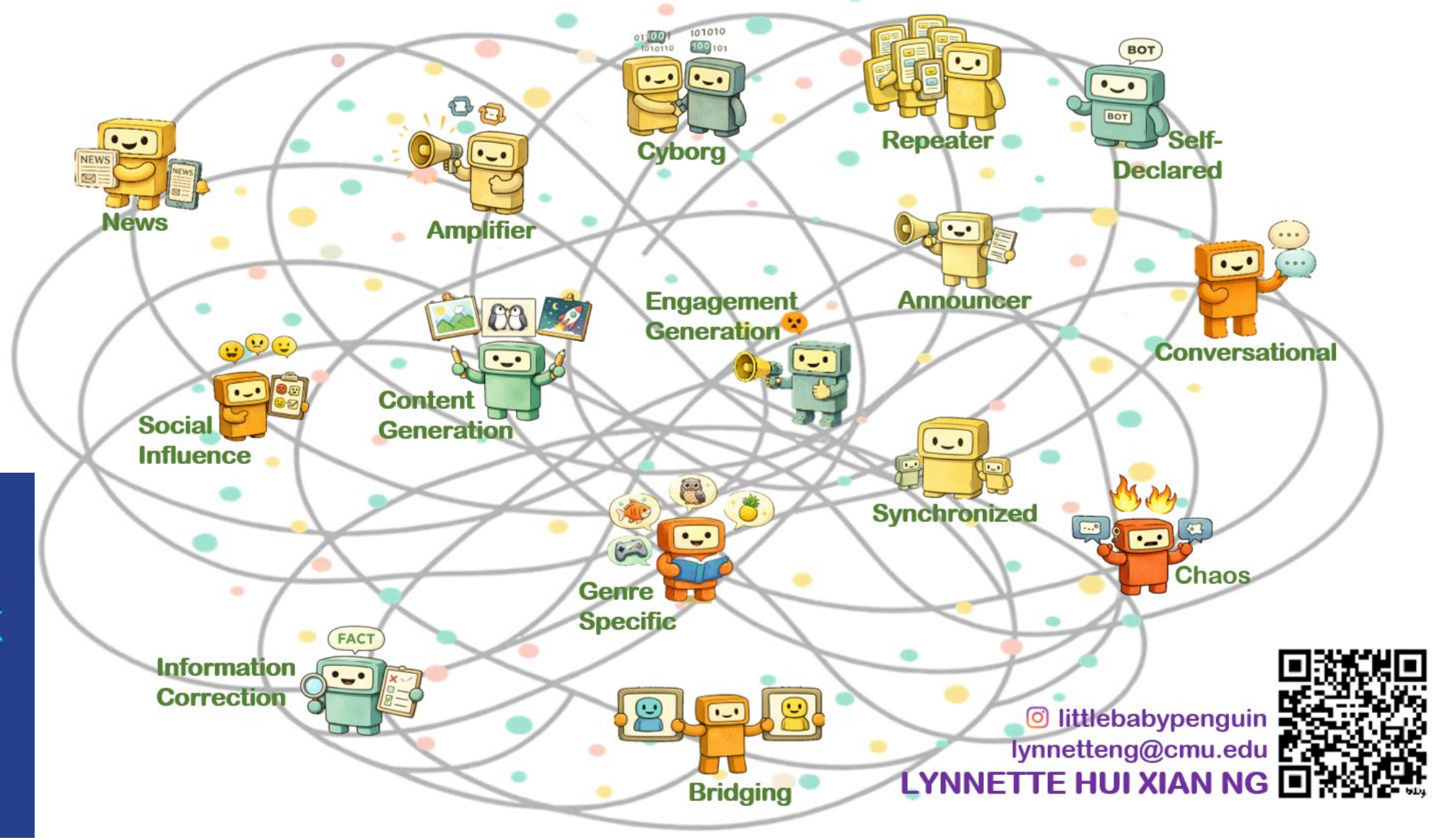




Carnegie Mellon University

# CYBER SOCIAL AGENTS

*social media bots are not a homogeneous adversary, but a heterogeneous mix of personas, which act as Cyber Social Agents to shape perceptions and collective behavior at scale.*




**Bot Book**


littlebabypenguin
lynnetteng@cmu.edu
LYNNETTE HUI XIAN NG