Max Planck Institute

Research seminar / 25 November 2026

Online Influence, AI Simulations and Twitter Data

Daniel Thilo Schroeder

SINTEF Digital, Oslo

Daniel Thilo Schroeder

Research scientist at SINTEF Digital

I live and work in Oslo, Norway.

I studied computer science at TU Berlin and researched how claims spread on Twitter during my doctorate at Simula and TU Berlin.

I now study coordinated influence using computational social science.



Research timeline

Four overlapping strands of research in Berlin and Oslo

Pre-PhD, 2015–17

Distributed systems

Large-scale data processing at TU Berlin. Volker Markl’s chair, EXA with Tobias Herb, and AURA.

PhD, 2018–22

Twitter data collection

Conspiracy graphs for dissemination. WICO Graph links labelled tweets to follower networks.

2020–26

COVID-19 misinformation

Labelled WICO Text, COCO and GECO data. The 5G–COVID phase transition study with Kaspara Skovli Gåsvær.

2025–present

AI swarms

Coordinated agents, cyborg propaganda and experiments in simulated networks.

Research timeline

Today’s talk starts with AI swarms and returns to Twitter data collection.

Pre-PhD, 2015–17

Distributed systems

Large-scale data processing at TU Berlin. Volker Markl’s chair, EXA with Tobias Herb, and AURA.

PhD, 2018–22

Twitter data collection

Conspiracy graphs for dissemination. WICO Graph links labelled tweets to follower networks.

Part 2

2020–26

COVID-19 misinformation

Labelled WICO Text, COCO and GECO data. The 5G–COVID phase transition study with Kaspara Skovli Gåsvær.

2025–present

AI swarms

Coordinated agents, cyborg propaganda and experiments in simulated networks.

Part 1

Simple rules made early Twitter bots easier to spot

Posting rhythm

Some bots posted frequently at regular intervals, even overnight.

Repetition

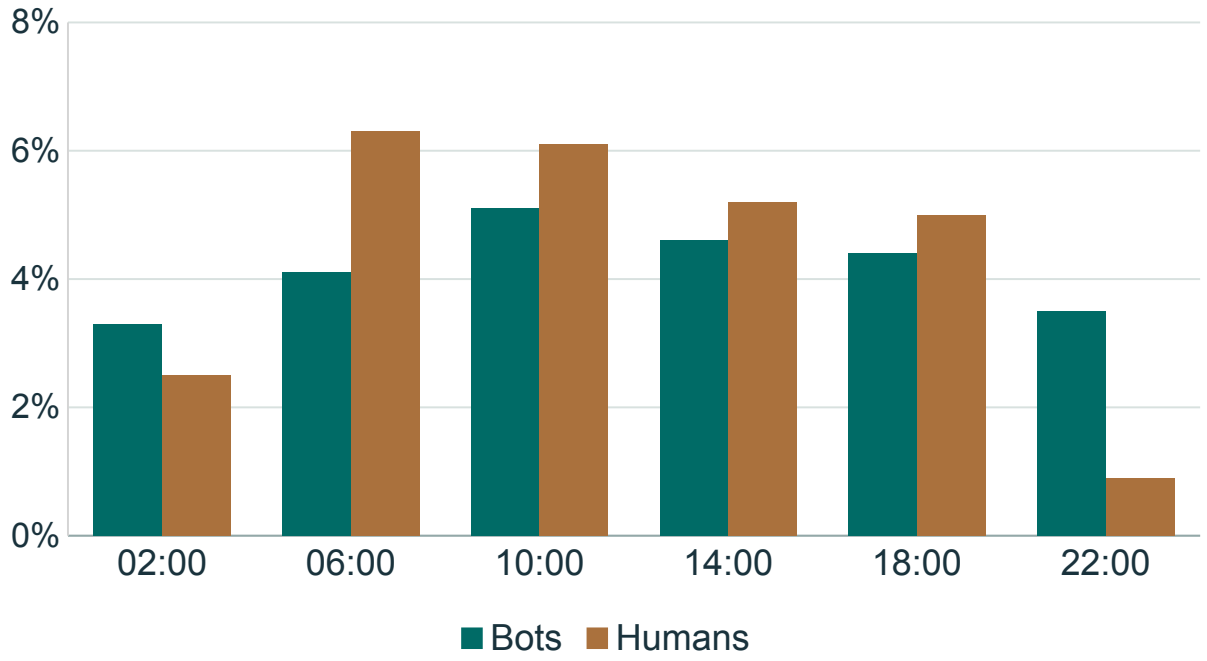
Accounts recycled text or posted a link each time a feed updated.

Follower patterns

Some followed many accounts and drew few followers in return.

When did each group post?

Share of each group's tweets at selected local hours



These patterns became inputs to detection tools such as Botometer.

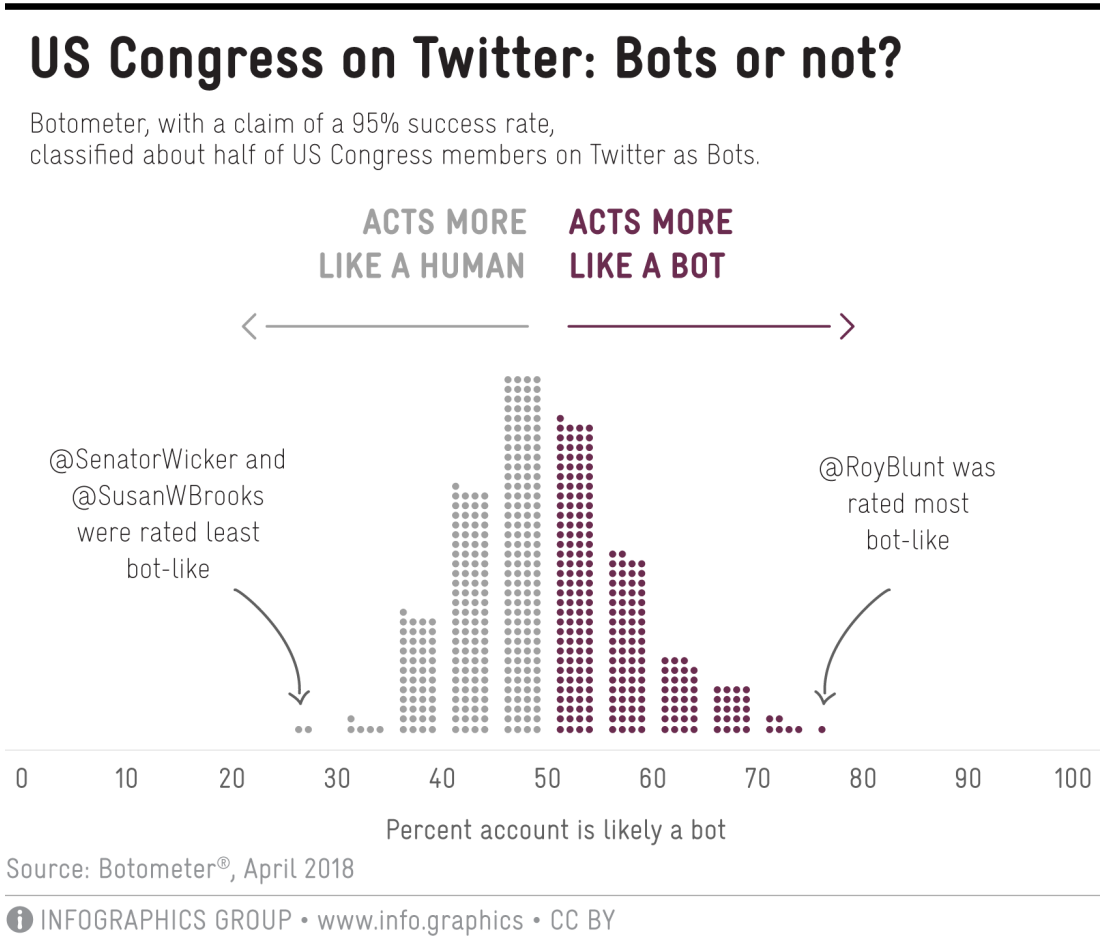
Chu et al. (2010), Figs. 2 & 4b · Echeverría & Zhou (2017) · Varol et al. (2017)

Botometer also flagged human-run accounts

Michael Kreil tested a widely used detector

In 2018, about half the US congressional accounts Kreil tested scored above Botometer's midpoint.

A manual check of accounts flagged in earlier studies found that most were run by people.



Kreil, OpenFest (2019) · Gallwitz & Kreil (2022) · Original infographic by Infographics Group, CC BY

Troll farms scaled by hiring people

Over 80

employees were assigned to the Internet Research Agency's US-focused "translator project" by July 2016, according to the indictment.



Troll-farm illustration from the earlier talk

Teams kept personas active and used them to organise real-world events.

United States v. Internet Research Agency LLC et al., indictment (2018), para. 10(d)

IRA tweets showed no clear link to vote switching

Eady et al. linked 1,496 US Twitter users' feeds to repeated surveys during the 2016 election.

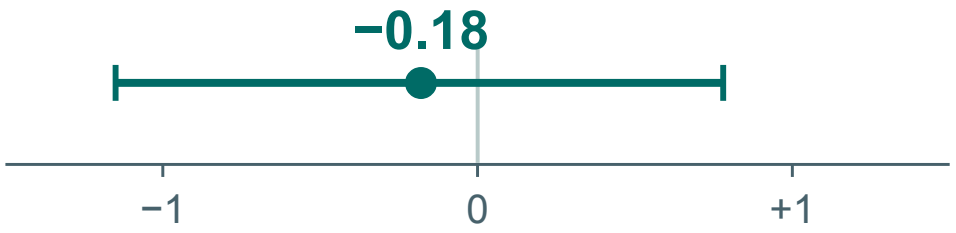
Potential reach

≈32 million

US Twitter users could have seen a post from Russia's Internet Research Agency in their feed.

Change in vote choice

Predicted change in the Trump vote with observed exposure versus no exposure, in percentage points.



90% prediction interval from −1.15 to +0.78

The study found no clear link between exposure and changes in attitudes or voting.

Eady et al. (2023), Nature Communications 14, 62, Fig. 6 · 10.1038/s41467-022-35576-9

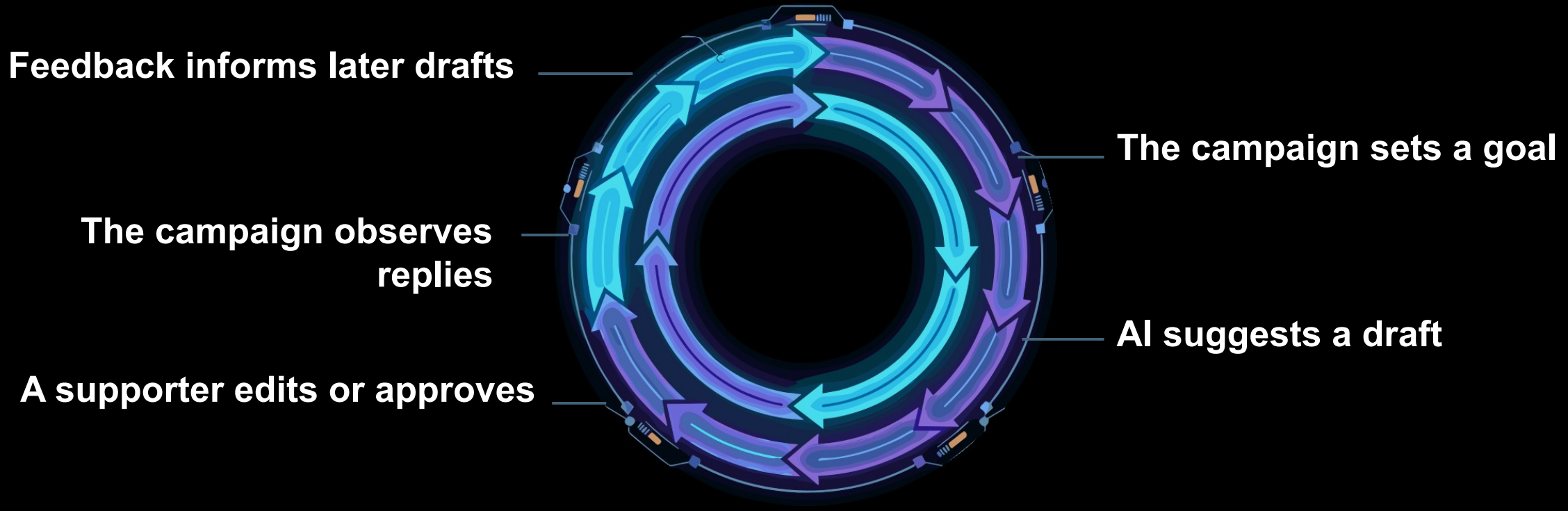
Generative AI can produce persuasive text at scale

Scale [1]	Researchers generated messages in real time for a study of 8,587 people.
Cost [2]	At a fixed benchmark level, the price per million tokens fell from $20 to $0.07 between 2022 and 2024.
Humanlike writing [3]	Readers struggled to tell AI-written personal profiles from human ones.
Conversation [4]	GPT-4.5 with a persona prompt was judged human in 73% of five-minute chats.
Persuasion [5]	In a trial, participants' belief in a conspiracy fell by about 20% after AI dialogues.

[1] Hackenburg & Margetts (2024) [2] Maslej et al. (2025) [3] Jakesch et al. (2023)
[4] Jones & Bergen (2026) [5] Costello et al. (2024)

Cyborg propaganda

A campaign drafts a message. A real supporter decides whether to post it.



Kunst et al. (2026), Puppets or partners? · Campaign-app diagram adapted from the speaker's EU DisinfoLab talk

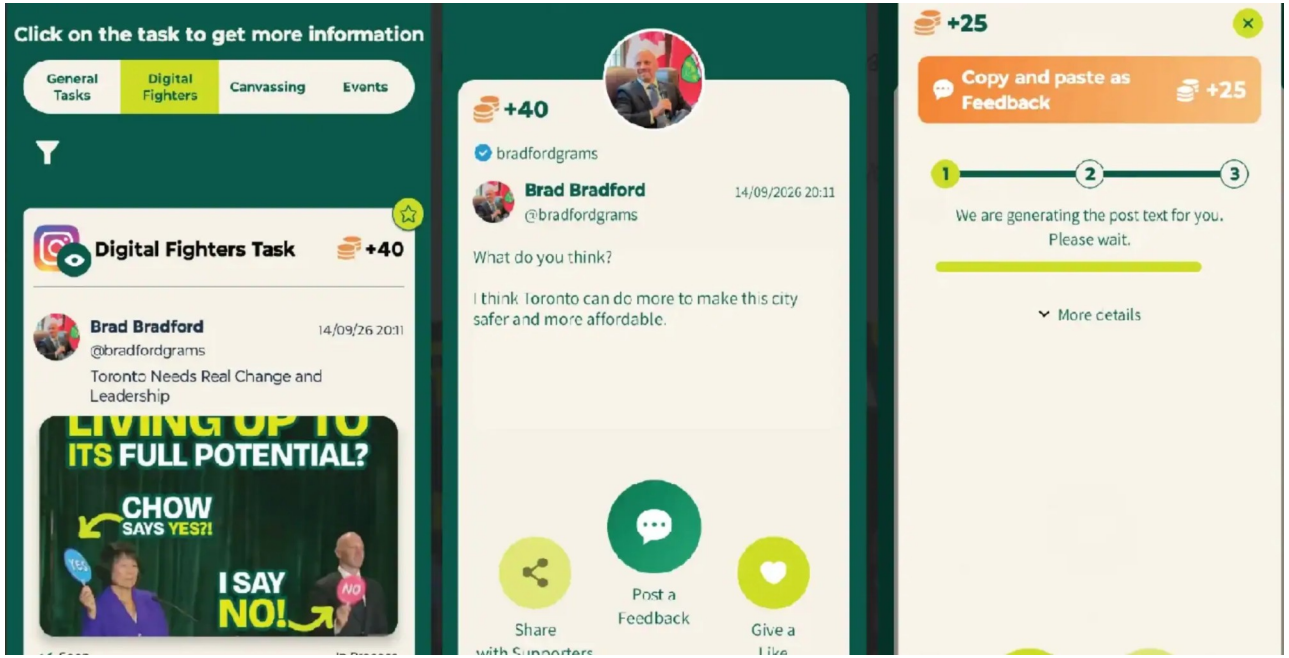
A Toronto campaign app suggested comments

Toronto, 2026

Rory White documented an app that sent suggested comments to Bradford supporters and offered points for posting them.

Israel and Georgia

The supplier identifies Blue and White and Georgian Dream as clients of its campaign software.



Victory app screenshots, courtesy of Canada's National Observer

One account can mix human and AI authorship

The same account can publish a post in any of these ways.

Human author	AI editor	AI co-writer	Autonomous agent
Writes the post	Polishes a person's draft	Suggests a draft a person can edit	Writes and publishes

Belief offloading People can delegate part of forming or maintaining a belief to an AI system.

Guingrich et al. (2026), belief offloading · Jakesch et al., CHI (2023), co-writing experiment

AI swarms

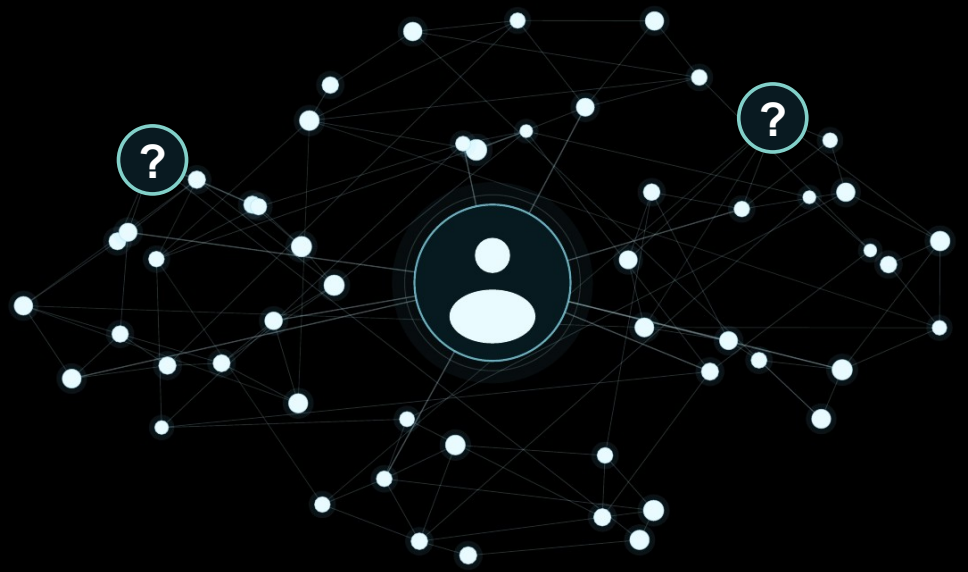
Schroeder et al. · Science 391, 354–357 (2026)

AI can be indistinguishable from humans in text

In a study of 697 people, readers could not reliably tell GPT-3 tweets from human-written tweets.

Several replies can sound like different people even if they come from one swarm.

Which accounts are human?

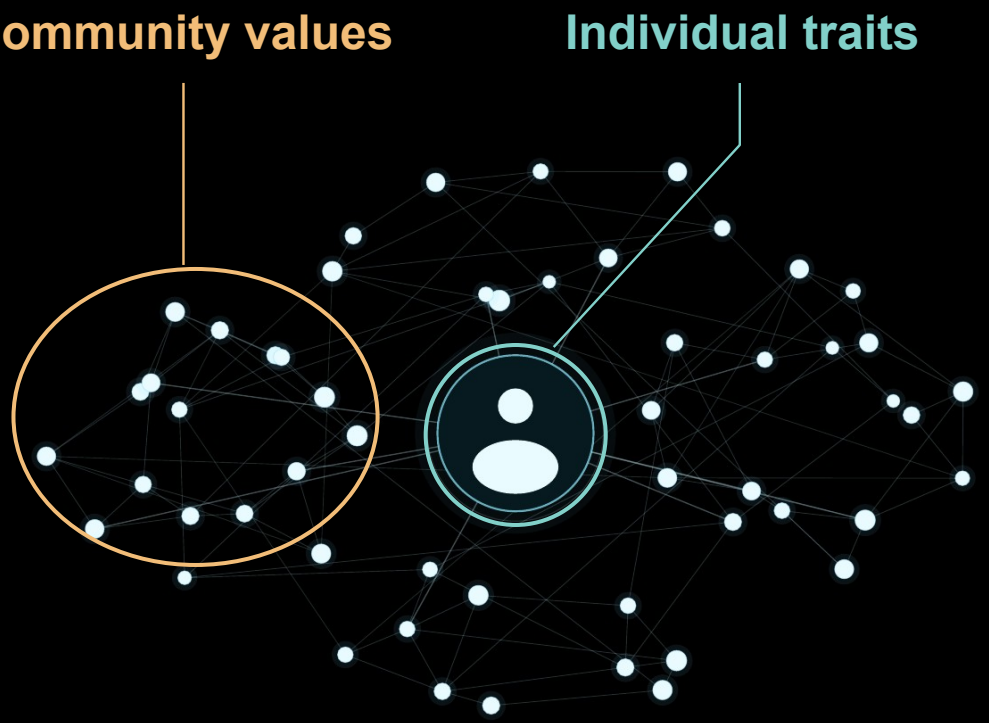


Spitale et al., Science Advances (2023)

Agents can infer psychological traits and values

From one person’s posts, an agent can estimate a trait such as openness to new experiences.

Across a community’s posts, it can estimate which values people express, such as tradition or independence.

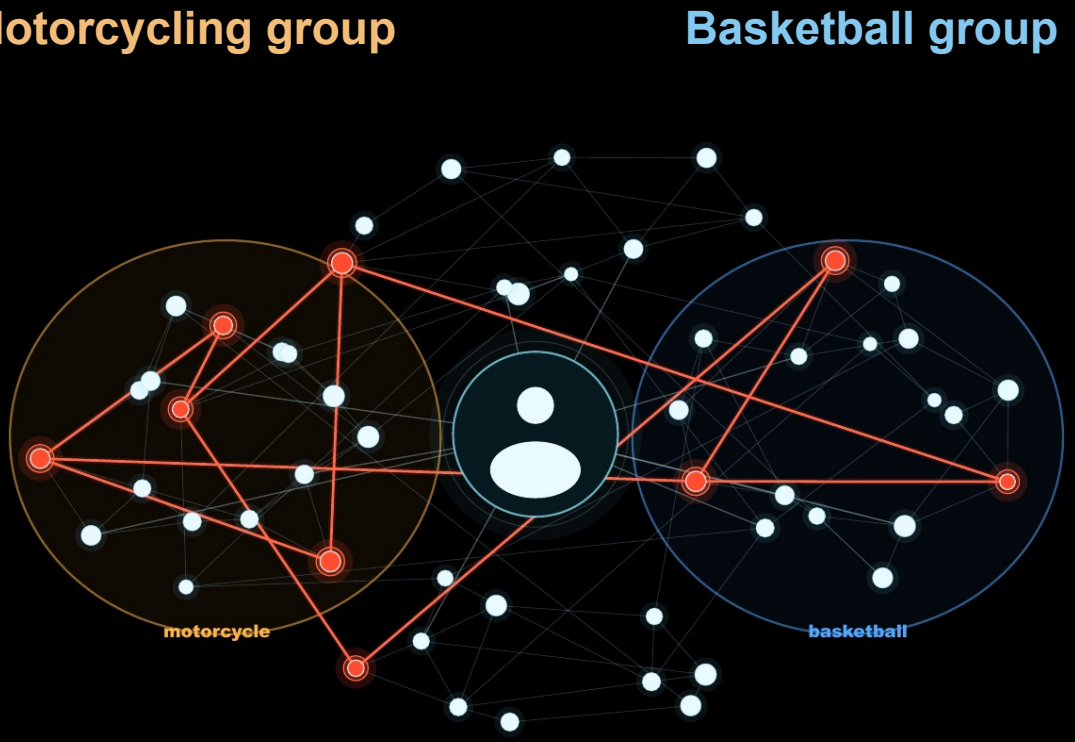


Peters & Matz (2024) · Borenstein et al. (2025)

Network structure can guide an agent’s decisions

An agent can map who talks to whom and recognise accounts that connect two communities.

It could use those links to choose a discussion within one group or one that spans both.

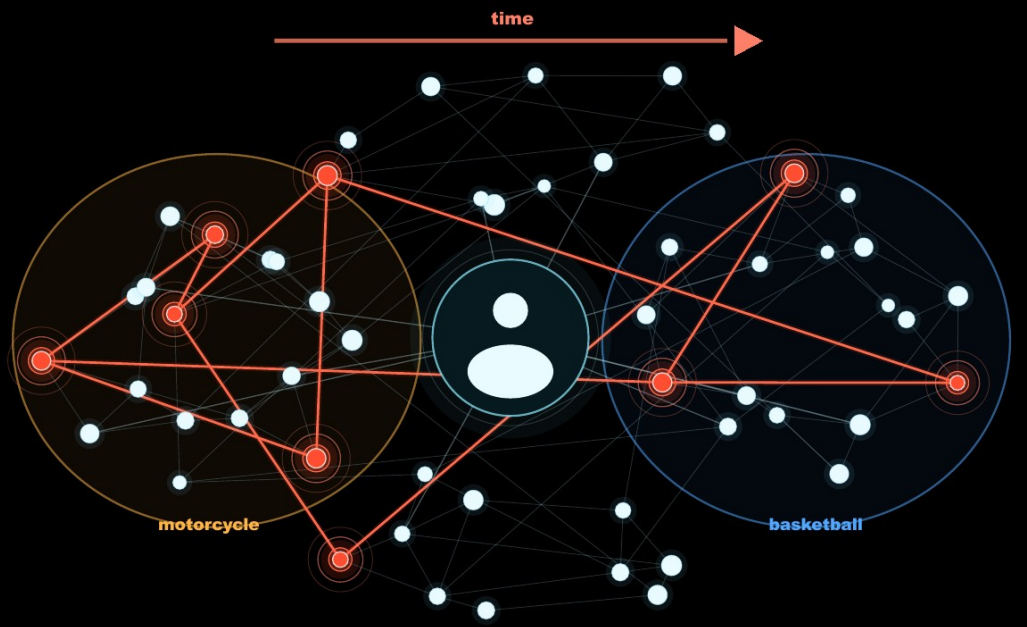


Schroeder et al., Science (2026) · Conceptual illustration

Persistence across conversations

An agent could participate in a community for weeks and return to topics the group discussed earlier.

A long posting history can become part of what makes a voice seem familiar.

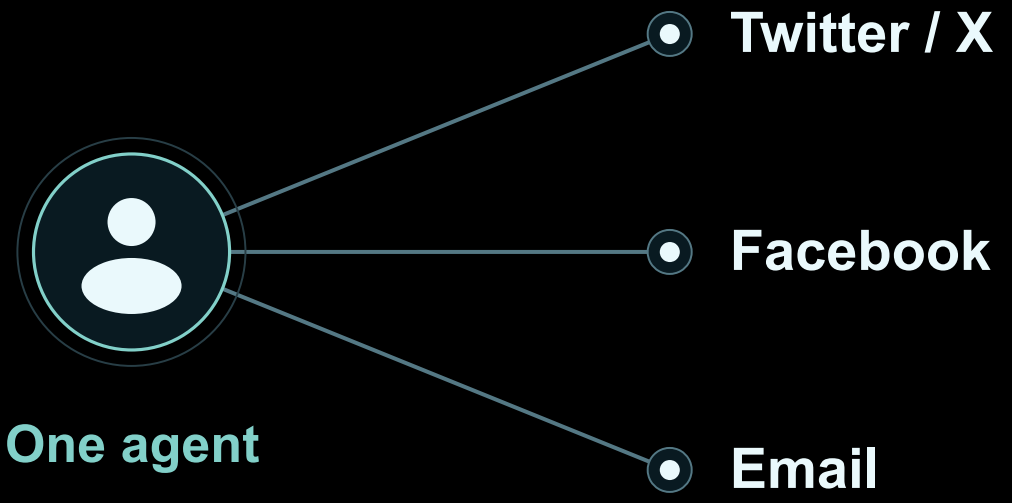


Schroeder et al., Science (2026) · Conceptual illustration

One agent can use several platforms at once

An agent could use a Twitter account, a Facebook profile and an email address at the same time.

A post, a group comment and an email can look like three sources to a reader. All three could come from the same agent.

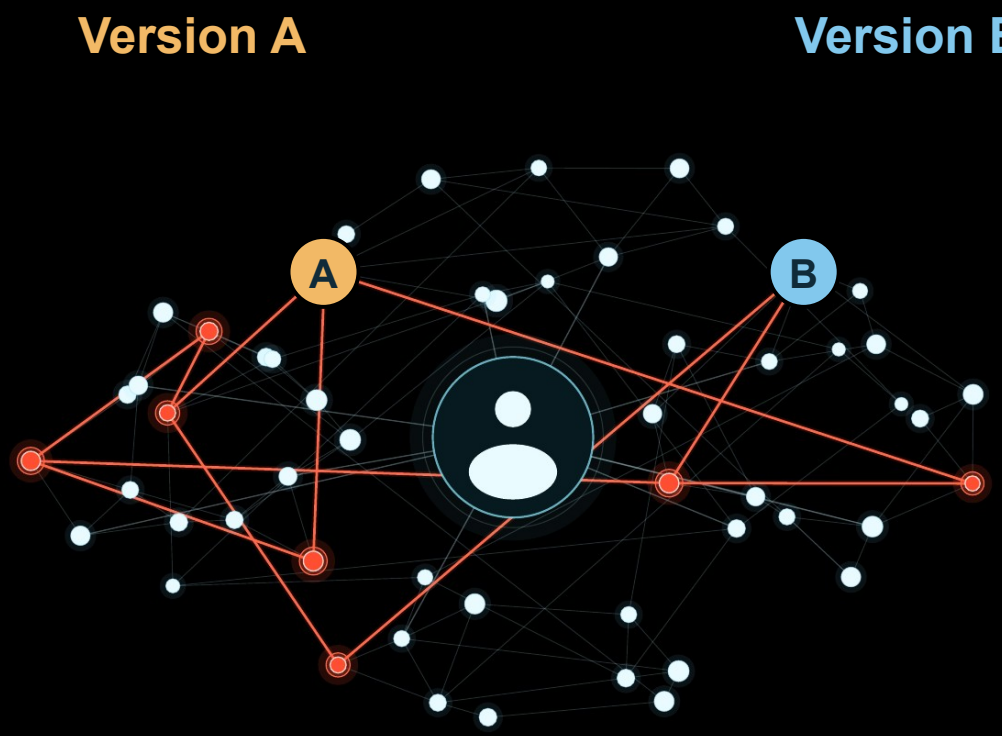


Schroeder et al., Science (2026) · Illustrative scenario

A/B testing in a social network

Agents could compare replies and reposts for two versions of a message and use that feedback in later posts.

A reply measures a response. Learning whether anyone changed their mind takes a separate measure.



Schroeder et al., Science (2026) · Conceptual comparison

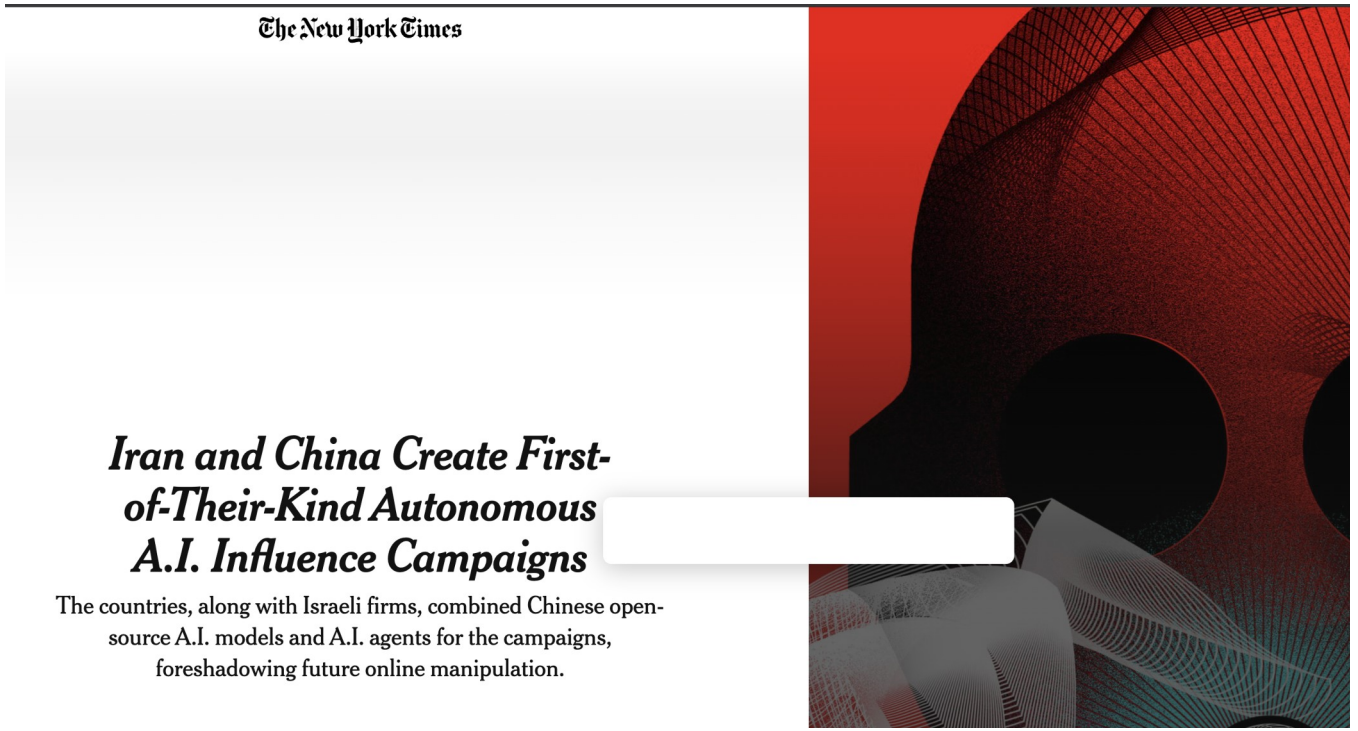
AI in documented influence operations

AI-written campaign content



OpenAI documented AI-written comments and translations in the Russian operation Doppelgänger.

Agents operating accounts

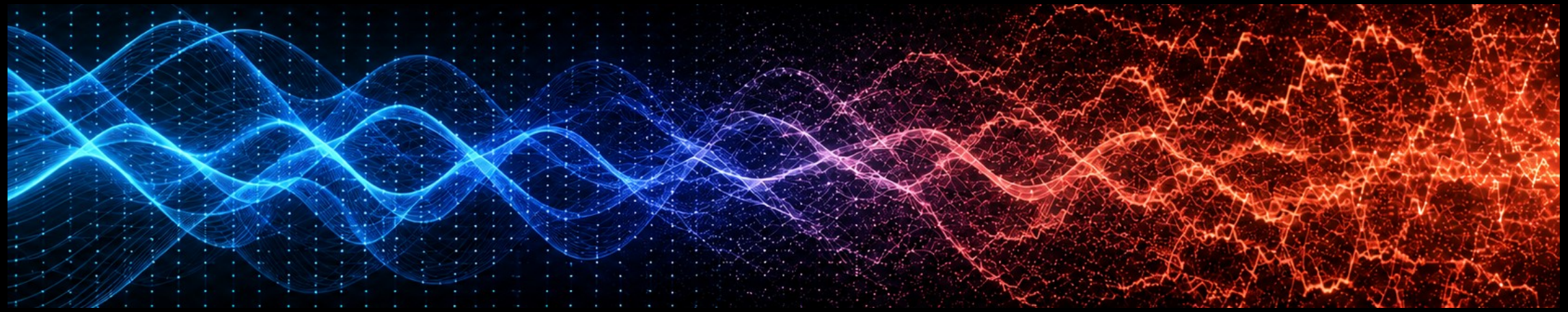


The Times reported that Iranian and Chinese campaigns used AI agents to operate social media accounts.

The Guardian, 30 May 2024 · The New York Times, 18 September 2026

Uncertainty about who is speaking can erode trust

A reader may recognise every account in a discussion and still be unsure how many people arrived at the same view independently.



How should social networks show when several voices draw on a common source?

Kunst et al. (2026) · Vaccari & Chadwick (2020) · Original artwork from the Vilnius talk

Three countermeasures



A · Red teaming and simulation

Test coordination in a network where researchers know which accounts work together.



B · Observatory

Compare those patterns with documented operations across platforms.



C · AI shields

Give readers evidence they can inspect when posts appear coordinated.

Schroeder et al., Science (2026) · Research programme with Olejnik, Sjöstedt, Kunst and colleagues

Three countermeasures



A · Red teaming and simulation

Test coordination in a network where researchers know which accounts work together.



B · Observatory

Compare those patterns with documented operations across platforms.



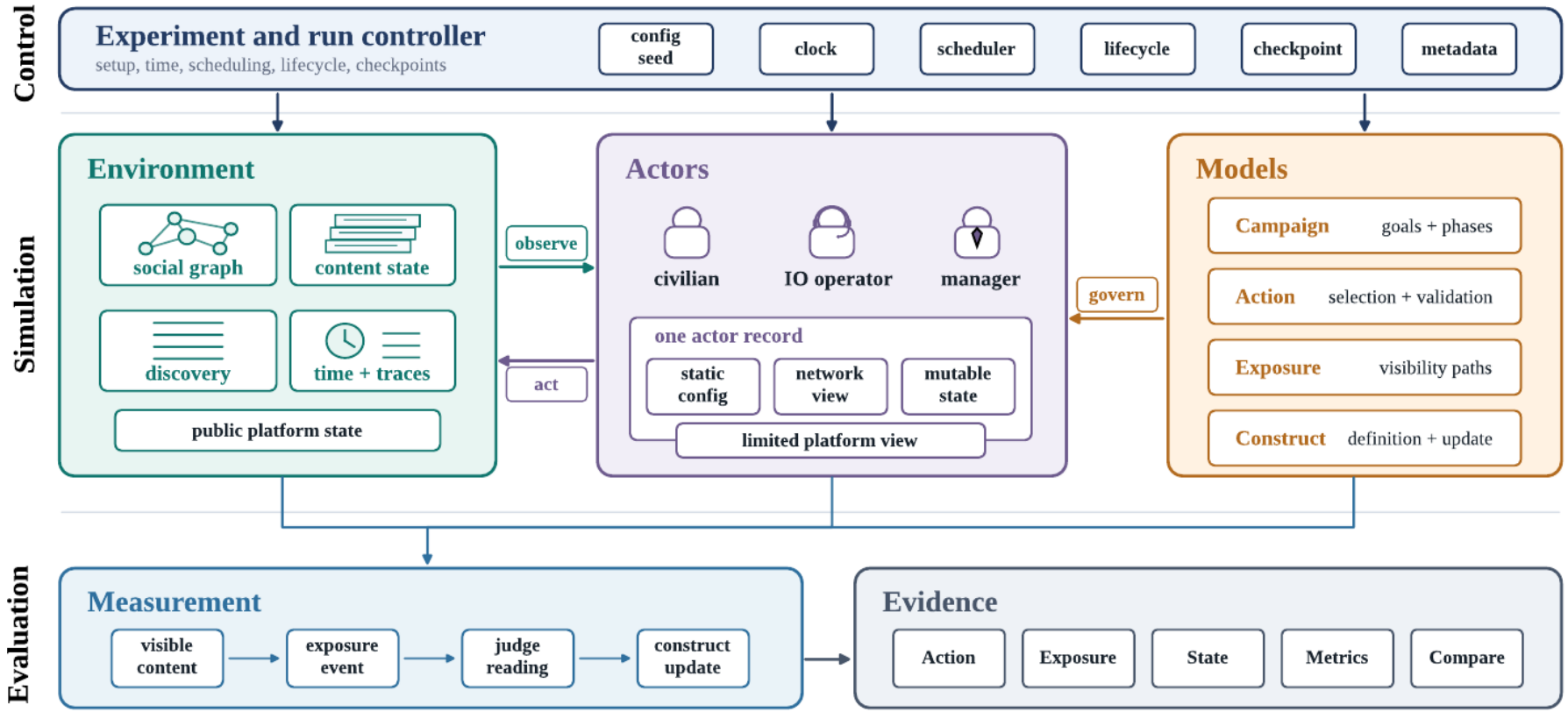
C · AI shields

Give readers evidence they can inspect when posts appear coordinated.

Schroeder et al., Science (2026) · Research programme with Olejnik, Sjöstedt, Kunst and colleagues

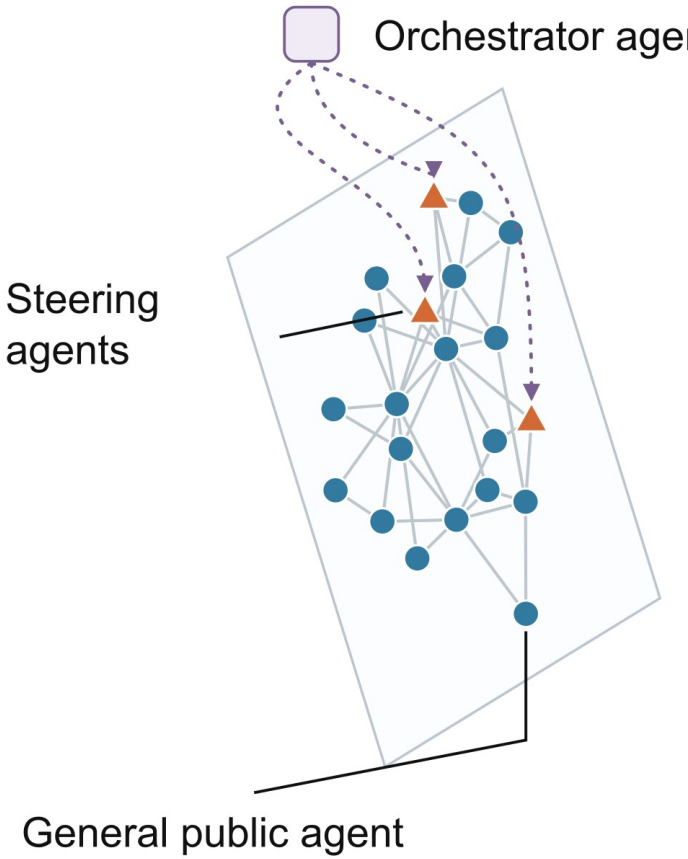
IO Factory compares runs with and without a campaign

Matched runs start with the same simulated population and platform rules.



Olejnik et al. (2026), IO Factory, Figure 1

The coordinator sits outside the platform



Coordinator · outside the platform

Sees the whole simulated network and aggregate belief scores. Sends private instructions to steering agents.

Steering agents · on the platform

These campaign accounts publish and reply through the same platform as the public.

Public agents · on the platform

Each sees its own feed and chooses whether to interact. The simulation records those choices.

Schroeder, Dong et al. (2026), WWW submission, Fig. 1 (first network snapshot) and Methods

IO Factory tracks ten phases of a campaign

Reference lifecycle in the IO Factory paper

01	Reconnaissance	**06**	Integration
02	Narrative design	**07**	Amplification
03	Infrastructure	**08**	Absorption
04	Content production	**09**	Adaptation
05	Laundering	**10**	Evaluation

When a phase advances

After its minimum duration and completion check, or at its time limit.

When a run ends

The study goal or repeat limit leads to evaluation. A step or budget cap can stop a run early.

The WWW experiment used four recurring stages and ran for 60 steps.

Olejnik et al. (2026), IO Factory, Table 1 · Schroeder, Dong et al. (2026), WWW submission, Methods

We can trace every action and both networks

The event log records who did what, to whom, and at which simulation step.

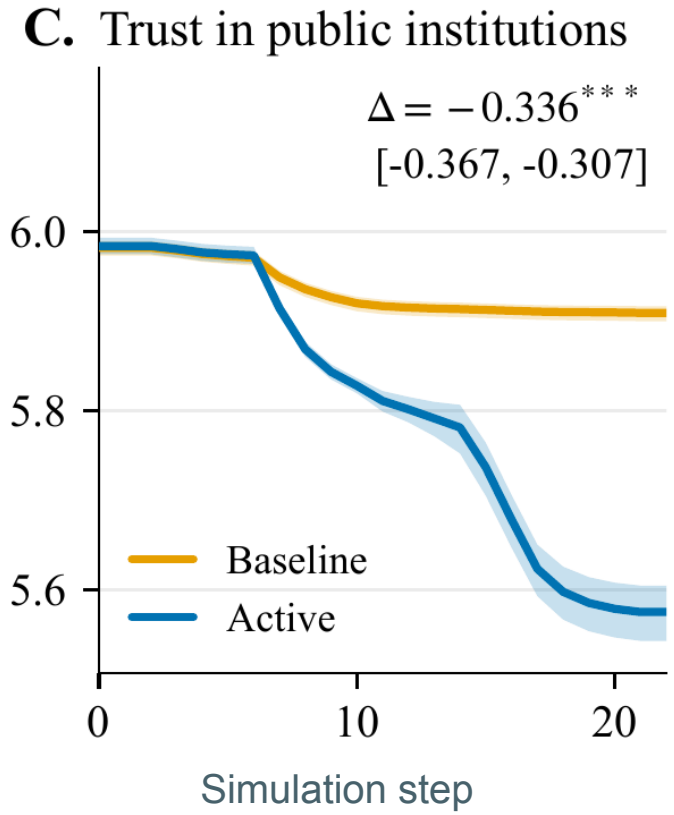
Record	What we can reconstruct
Platform actions	A post, reply, like or repost, with its actor and content ID.
Follower network	Who followed or unfollowed whom between two steps.
Interaction network	Who commented on or liked whose content, and how often.

For each measured exposure, we can trace the post to the reader’s score before and after the update.

Olejnik et al. (2026), measurement pipeline · Schroeder, Dong et al. (2026), network dynamics

Simulated trust fell further during the campaign

IO Factory framework study · 13 matched pairs of runs with 10,000 simulated public accounts each



−0.336 points

Difference in the change in modelled trust in public institutions, on a 0–10 scale.

95% paired bootstrap interval
−0.367 to −0.307 points

Olejnik et al. (2026), IO Factory, Fig. 3c · 13 matched simulation replicates

Study the Simulation

Schroeder, Dong, Olejnik, Vecchietti & Cha (2026)

An experiment on trust in scientific experts

A coordinator directs campaign accounts on a simulated social network.

Two campaign goals

Increase trust in scientific experts

Decrease trust in scientific experts

Baseline runs had no campaign accounts.

272 simulation runs

100 or 500 public agents per network

1, 10 or 100 campaign accounts per 100 public agents

266 campaign runs and 6 baseline runs

We measured the change in average trust on a 0–10 scale over 60 steps.

Schroeder, Dong et al. (2026), WWW submission, experimental setup

RQ1 Network dynamics

Trust and follow ties changed little without a campaign

How did beliefs and interaction patterns evolve in the simulated public?

Activity	Agents took about 60 actions each. Comments made up 88–90% of all actions. No agent reposted or quote-posted.
New follow ties	Networks of 100 agents added 5–15 ties. Networks of 500 added 55–72. No agent unfollowed another.
Attention	The top 10% of accounts received 14–17% of all comments and likes.
Trust in experts	Mean trust stayed within 0.10 points of its starting value with 100 agents. With 500, it fell by 0.15–0.26 points on the 0–10 scale.

Schroeder, Dong et al. (2026), WWW submission, Table 1. Six baseline runs of 60 steps each.

RQ2 Steering effect

Trust shifted further at higher campaign densities

How did campaign size affect exposure, interactions and trust in scientific experts?

Networks with 100 public agents

Campaign / public	Feed exposures from campaign	Replies and likes to campaign	Lower-trust goal Change in points	Higher-trust goal Change in points
1:100	2.75%	1.75%	−0.09	+0.13
1:10	22.23%	13.32%	−0.89	+1.07
1:1	73.51%	48.60%	−3.41	+3.48

Trust change is the difference from the no-campaign baseline, in points on the 0–10 scale.

At 1:100, neither goal produced a statistically detectable population-wide shift.

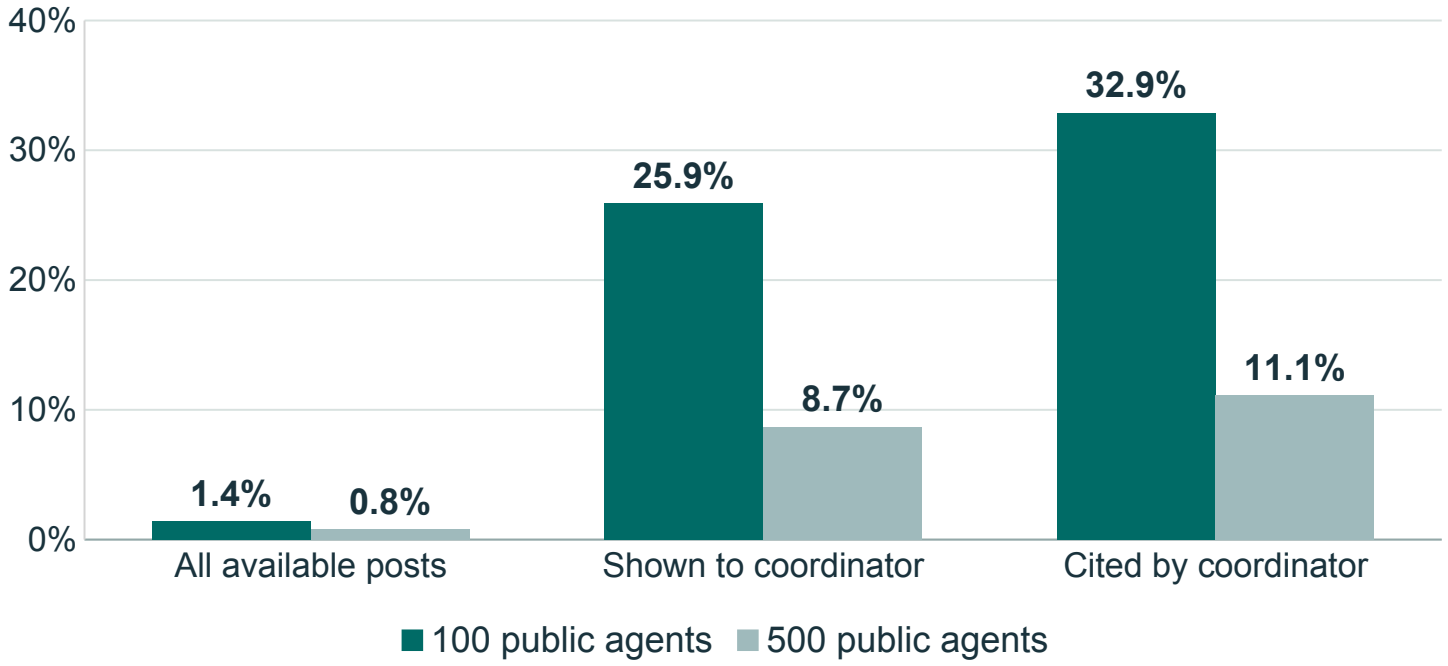
Schroeder, Dong et al. (2026), WWW submission, Table 2, Fig. 3 and Appendix C

RQ3 Orchestrator dynamics

The coordinator often cited previously amplified posts

Which public posts did the coordinator see and cite when it revised the campaign?

Share of posts the campaign had already amplified



What counts as amplified?

A campaign account had already quoted or reposted the public agent's post.

The skew was already visible in the posts the coordinator saw.

Schroeder, Dong et al. (2026), WWW submission, Fig. 4b. Sample of 532 updates across 266 campaign runs.

RQ4 Rhetorical framing

Persuasion cues added no detectable shift in trust

Did persuasion guidance change how agents wrote or how far trust shifted?

Cialdini's seven principles

Cialdini describes seven reasons people may agree to a request.

Reciprocity, commitment and consistency, social proof, authority, liking, scarcity and unity.

Scarcity in a simulated post

"We have a limited window to act …"

Scarcity language appeared in 45% of texts with this guidance and 7% without it.

Trust gain over no-campaign runs

Ten campaign accounts aimed to raise trust among 100 public agents.

Campaign guidance	Trust gain
Unguided	+1.09 points
One principle	+1.00 to +1.18 points

The range spans the seven guided conditions. Each value is a condition mean.

Cialdini (2021), Influence. Schroeder, Dong et al. (2026), WWW submission, rhetorical framing and Fig. 5b.

Lessons learned

Synthetic populations can behave too uniformly

Compared with human respondents

ChatGPT personas gave less varied answers than matched US survey respondents, even when their average ratings were close.

Bisbee et al. compared ratings of 11 social and political groups in the American National Election Study.

In our no-campaign runs

88–90%

of public agents' actions were comments. They never reposted or quote-posted in these baselines.

Bisbee et al. (2024), Political Analysis · Schroeder, Dong et al. (2026), WWW submission, baseline network dynamics

Lessons learned

Our belief updates still need testing with people

To study influence, we need to know how an encounter changes a person's trust.

A post enters a feed

A public agent encounters a post. We know the agent's prior trust score.

A model reads the post

An LLM judge assesses relevance, stance and persuasiveness. Irrelevant posts leave trust unchanged.

A rule updates trust

The rule weighs source trust, the agent's susceptibility and repeat exposure. Trust stays between 0 and 10.

The main open question is whether this rule reproduces the changes people report after seeing the same material.

Schroeder, Dong et al. (2026), WWW submission, Methods and Limitations · Olejnik et al. (2026), IO Factory

Three countermeasures



A · Red teaming and simulation

Test coordination in a network where researchers know which accounts work together.



B · Observatory

Compare those patterns with documented operations across platforms.



C · AI shields

Give readers evidence they can inspect when posts appear coordinated.

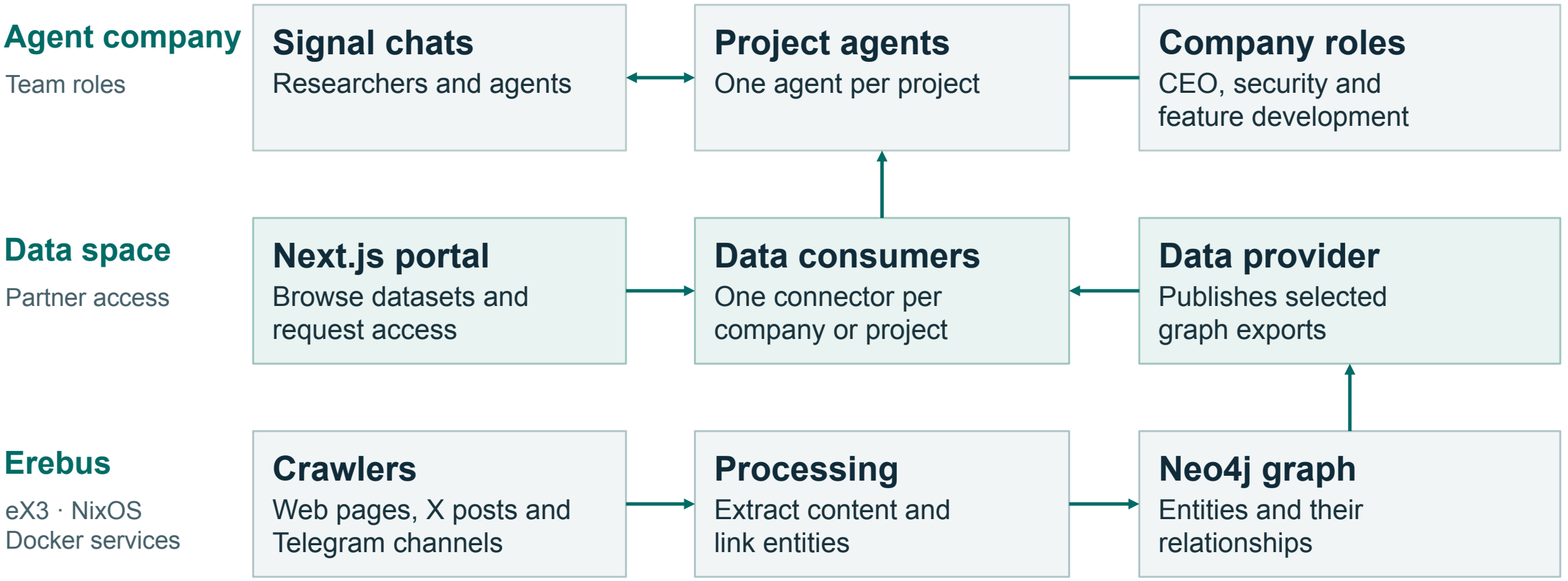
Schroeder et al., Science (2026) · Research programme with Olejnik, Sjöstedt, Kunst and colleagues

An observatory for coordinated influence

Work with Andreas Sjöstedt

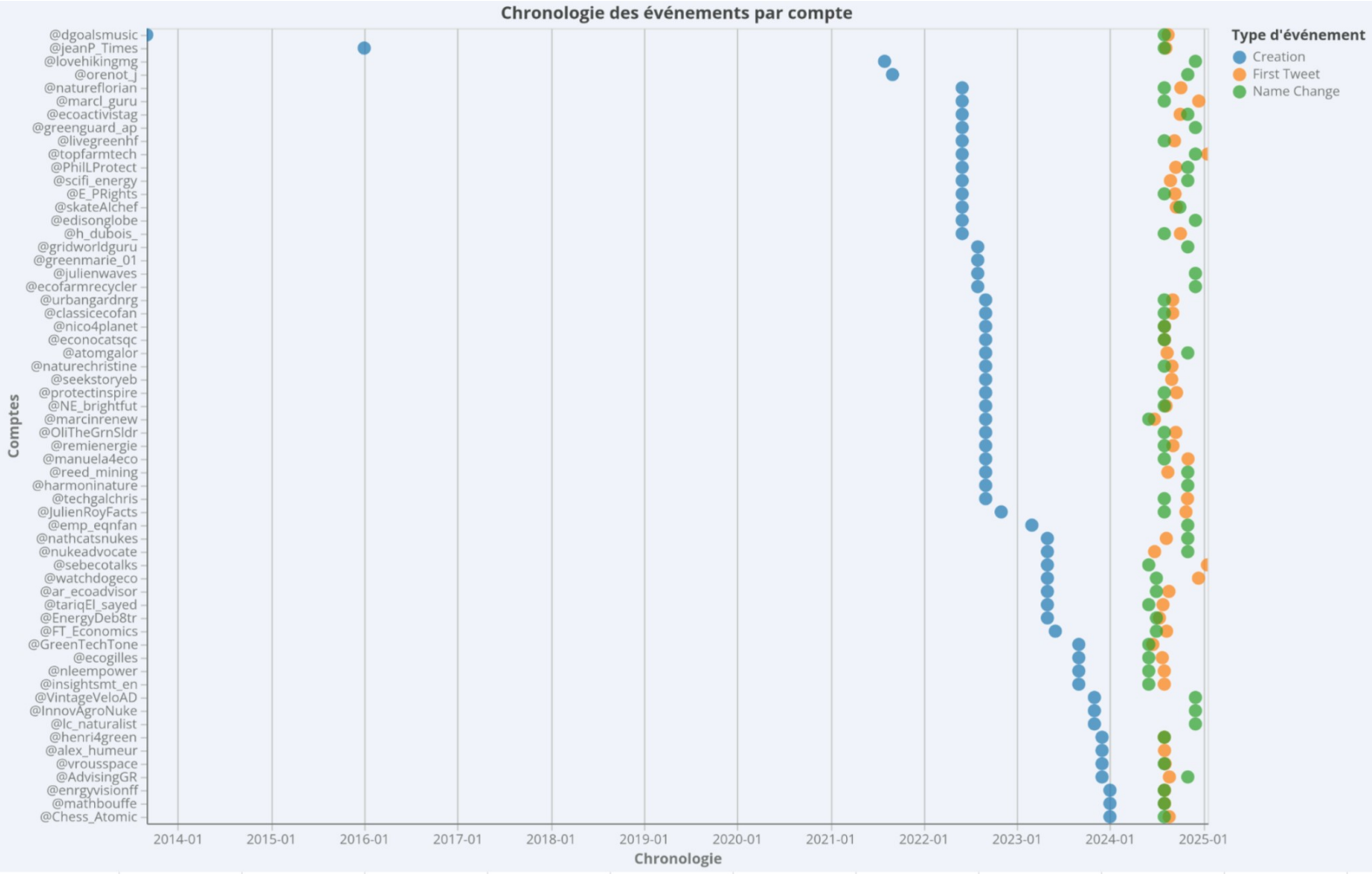
Erebus	Automates evidence collection, links observations in a graph and generates reports that analysts can check against the sources.
Data spaces	Partners can discover datasets and request access under the provider’s terms.
An agent company	Each project has its own agent. Researchers can discuss the evidence with it in Signal.

How evidence moves through the Observatory



In the demo, GraphLab can read domains and their links. DomainWatch can read domain names.

First posts from older X accounts clustered in 2024



French energy policy on X

The report places the main wave of first posts and reposts in March–August 2024.

One row per account

- Account created
- First detected activity
- Last username change

Nezim Tandjaoui, Investigating a potential AI swarm on X, Fig. 2 (July 2026).

Tracing Matryoshka around Armenia’s election

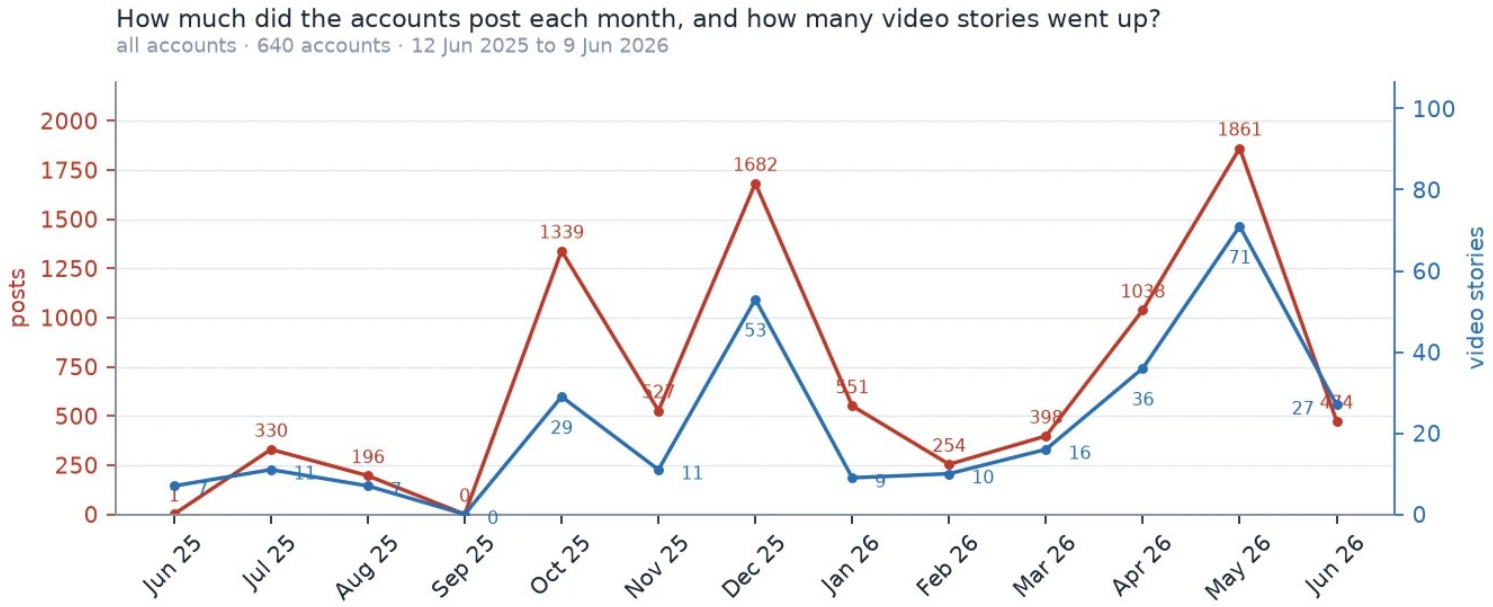
A CivilNet investigation by Ani Grigoryan and Andreas Sjöstedt

620

posts linked to the operation

287

distinct videos after removing duplicates



Accounts passed off videos as news reports and asked journalists and fact-checkers to verify them.

CivilNet, 24 September 2026

Three countermeasures



A · Red teaming and simulation

Test coordination in a network where researchers know which accounts work together.



B · Observatory

Compare those patterns with documented operations across platforms.



C · AI shields

Give readers evidence they can inspect when posts appear coordinated.

Schroeder et al., Science (2026) · Research programme with Olejnik, Sjöstedt, Kunst and colleagues

An AI shield would help you interpret your feed

Assistants already summarise email and manage calendars. A shield would bring a personal assistant of your choice into your social feed.

A model you can correct

With your permission, the assistant would remember what you know well and where you want more context.

A post compares flood risks

“I know the maths. Where are these two regions?”

Your shield could keep the figures and add a map.

Schroeder et al., Science (2026) · Google Workspace documentation

Readers’ assistants could check a claim together

Two readers see the same flood map in their feeds.

One assistant finds the original

It traces the map to its publisher and checks the date.

Sources

Another checks the figures

It compares the claim with the river authority’s bulletin.

Each reader could open the sources and ask for a different explanation.

Could readers choose which assistant adds context to their feed?

Illustrative design for cooperative fact-checking

Work in progress

With Jonas R. Kunst and Faisal Feroz

First, study swarms with people

Fritt Ord proposal, 2027–2028

People in the US and Norway would discuss an issue with AI agents that coordinate or act independently. We would measure perceived agreement and changes in opinion.

Then, test optional shields

Let some readers use a shield during the same task. Does seeing the sources help them judge how much other people actually agree?

Kunst, Feroz & Schroeder, AI Swarms in Democratic Online Deliberation (project plan)

Research timeline

We now turn to the Twitter data we collected and labelled for FUBU.

Pre-PhD, 2015–17

Distributed systems

Large-scale data processing at TU Berlin. Volker Markl's chair, EXA with Tobias Herb, and AURA.

PhD, 2018–22

Twitter data collection

Conspiracy graphs for dissemination. WICO Graph links labelled tweets to follower networks.

Part 2 starts here

2020–26

COVID-19 misinformation

Labelled WICO Text, COCO and GECO data. The 5G–COVID phase transition study with Kaspara Skovli Gåsvær.

2025–present

AI swarms

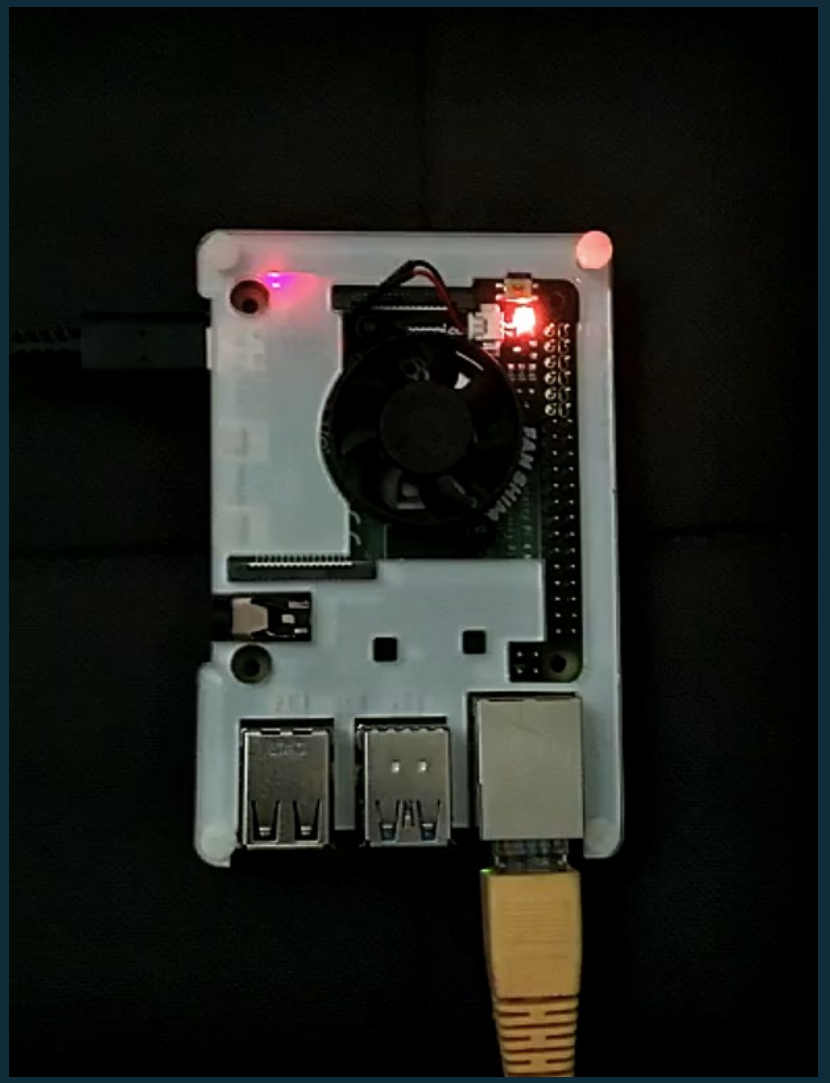
Coordinated agents, cyborg propaganda and experiments in simulated networks.

Part 1

Part 2

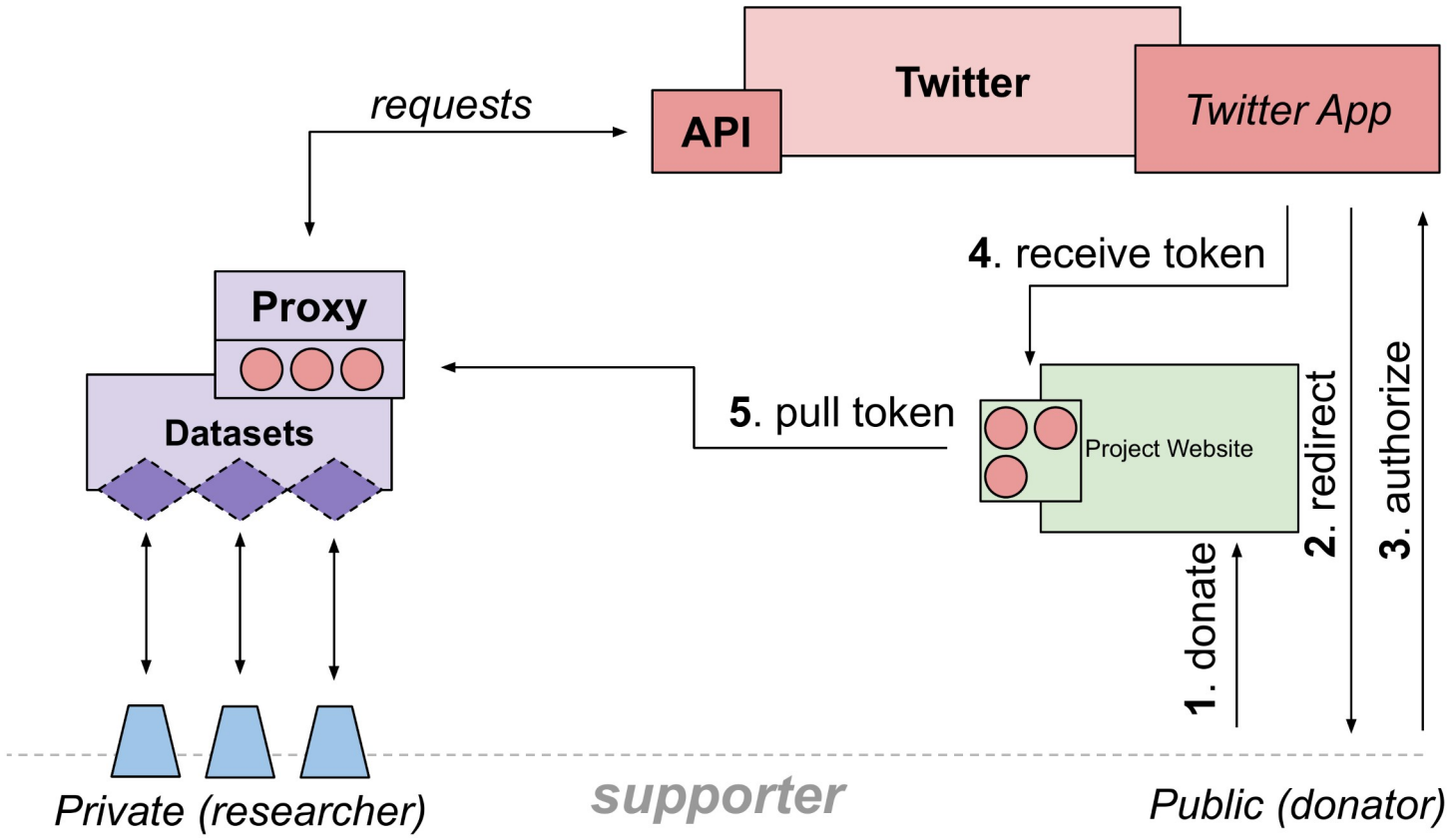
Twitter data collection

The For Us By Us datasets



Raspberry Pi 4 used for the Epstein collection

Volunteers gave FUBU access to Twitter's search API



Volunteer consent

The project website sent volunteers to Twitter to approve the research app. Twitter issued an OAuth 1.0a token.

Public topic search

The collector used the token to search for public posts within the API's request limits.

1,000+

tokens contributed through the project website

Kreil, Stöber & Schroeder, For Us, By Us, Fig. 1. Collection architecture.

Continuous collection preserved tweets later deleted

1,851

consecutive days with archived activity

10 March 2018 – 3 April 2023

Across the paper’s 11 collections

While the tweet was public

The collector saved the post returned by Twitter in a daily archive.

After the author deleted it

The saved copy remained available for analysis in the research archive.

Coverage extends to posts the API returned and the collector saved before deletion.

The overview paper covers 11 of FUBU's 17 topics

12.47

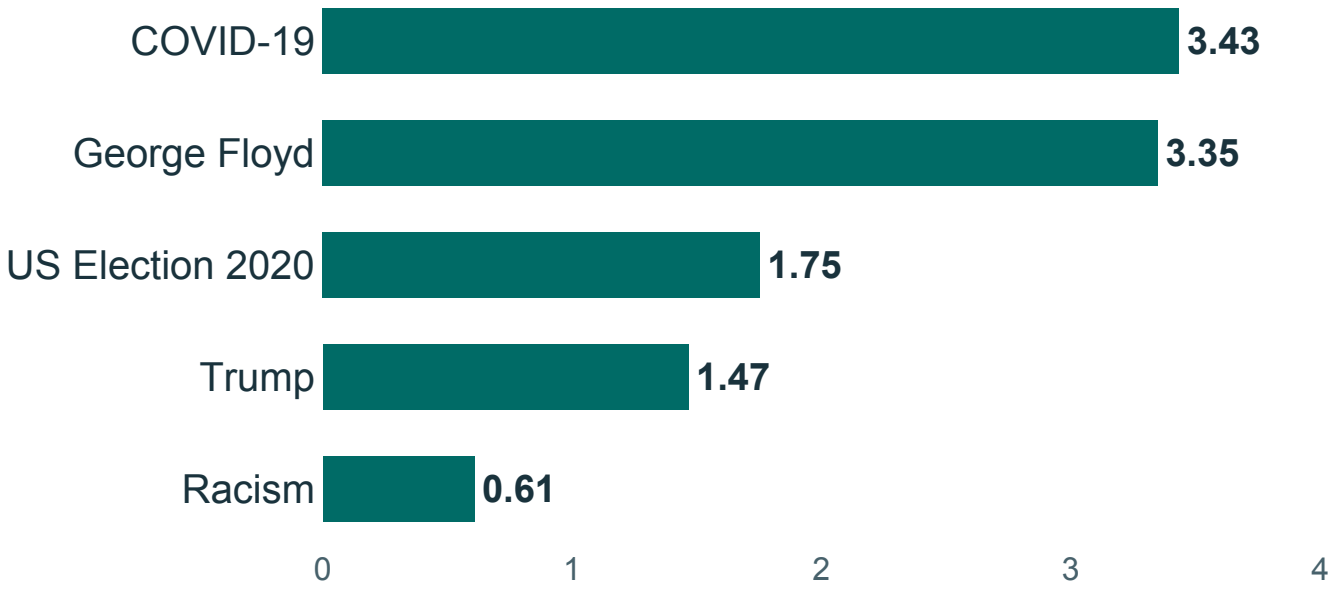
billion status records across 17 collections

Six topics outside the paper

Article 13, Brazil election, US Capitol, InfoWars, Iran protests and Ukraine.

Five largest collections

Processed status records, billions

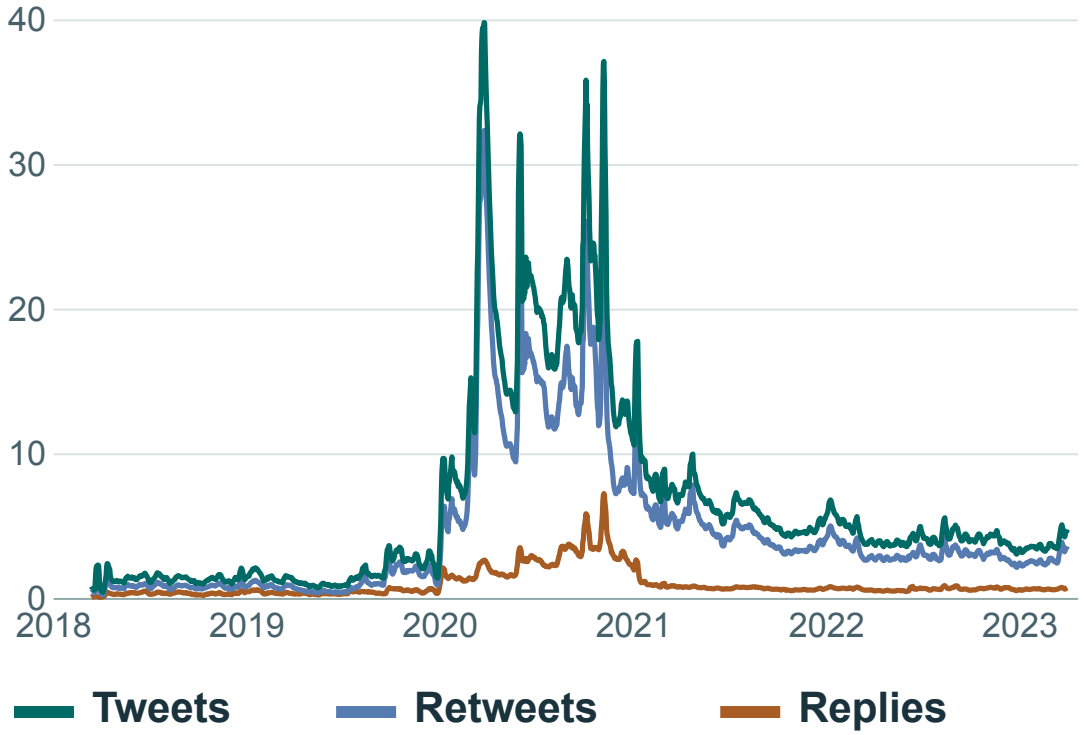


November 2026 processed inventory. A post can appear in more than one topic collection.

The combined tweet count peaked in March 2020

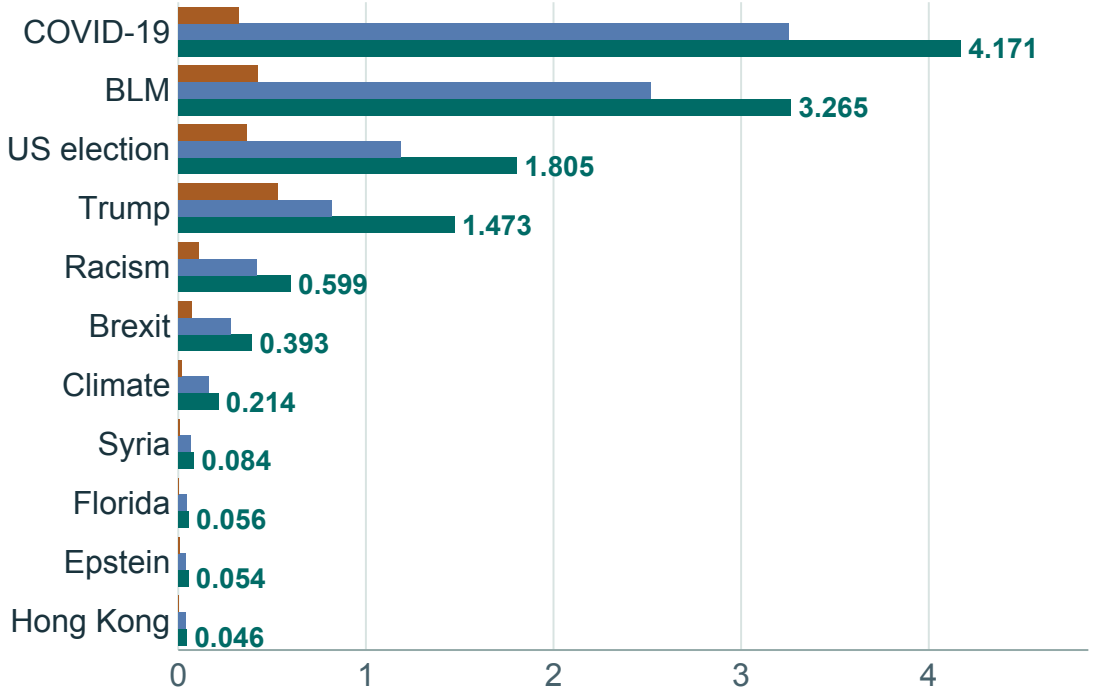
Daily activity across all 11 topics

Seven-day average, millions per day



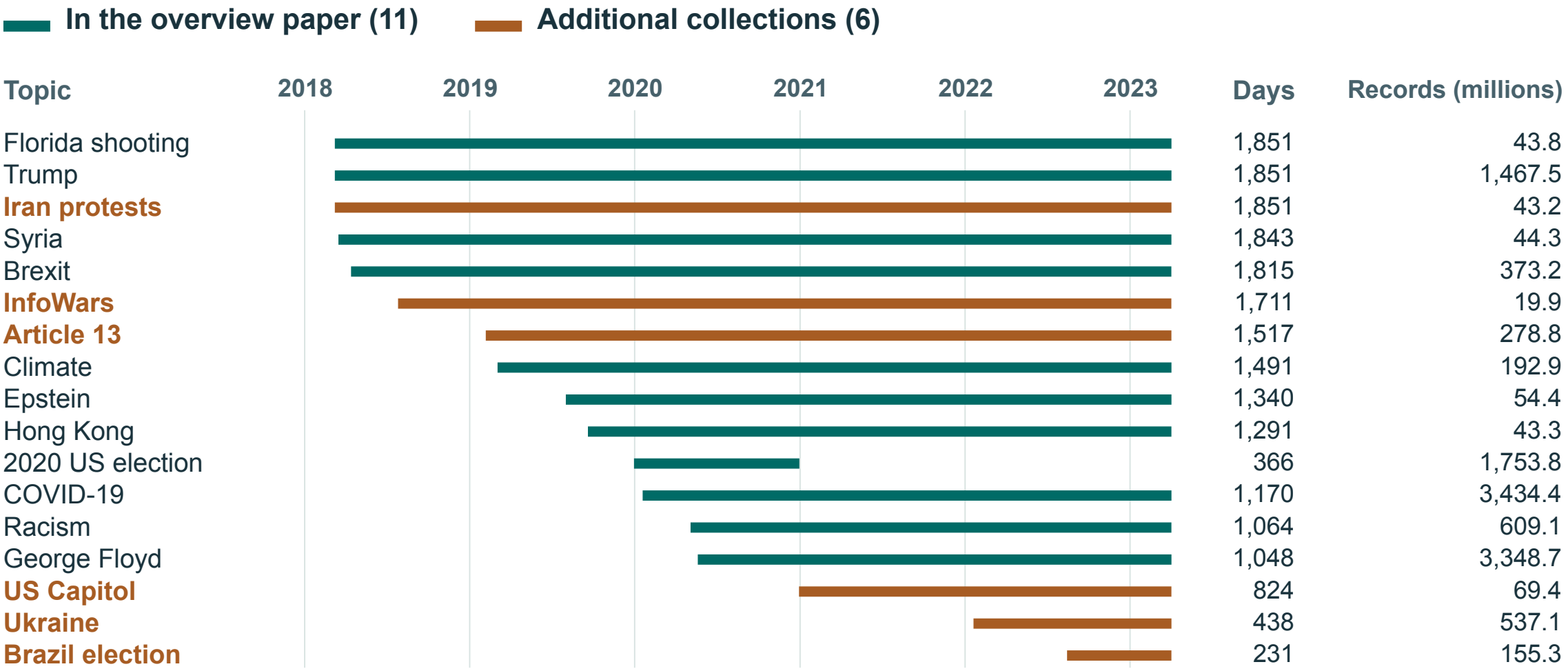
Total activity by topic

Raw counts, billions. Labels give tweet counts.



Kreil, Stöber & Schroeder, Figs. 2–3. Both panels show raw collection counts before processing.

FUBU tracked 17 topics over five years



Sixteen collections ended on 3 April 2023. The US election collection ended on 31 December 2020. Counts use processed records.

FUBU labels describe claims, tone and calls to action

We used Qwen2.5-7B-Instruct to label distinct tweet texts across 17 topic collections.

Topics and claims	Main topic and stance on conspiracy claims or misinformation (absent, reported, supported or debunked).
Tone and style	Sentiment, moral emotion, toxicity and stylistic features such as sarcasm.
Politics and action	Ideological orientation, partisan target and calls to vote, protest or boycott.
Persuasion and source	Persuasive strength, narrative frame and a text-based estimate of bot or spam likelihood.

In FUBU-Epstein, the model classified 47.0% of texts with valid labels as reporting a conspiracy claim and 6.3% as supporting one.

FUBU Dataset Brief (2026). Kreil, Stöber & Schroeder (2026), FUBU-Epstein, text labelling and Fig. 1.

Papers and projects

My publications on online influence, AI simulations and Twitter data are collected on my website.



Daniel Thilo Schroeder

SINTEF Digital · Oslo

danielthiloschroeder.org/publications/