Preprints.org

Article

Not peer-reviewed version

Five Myths About Influence Operations: What 25 Million Tweets Across Seven State Campaigns Reveal

Emilio Ferrara *

Posted Date: 29 June 2026

doi: 10.20944/preprints202606.2093.v1

Keywords: influence operations; coordinated inauthentic behavior; disinformation; social media

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC, OpenAlex.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Five Myths About Influence Operations: What 25 Million Tweets Across Seven State Campaigns Reveal

Emilio Ferrara [1,2]

[1] Thomas Lord Department of Computer Science, University of Southern California, Los Angeles, CA, USA; emiliofe@usc.edu
[2] Information Sciences Institute, University of Southern California, Marina del Rey, CA, USA

Abstract

A set of "stylized facts" about state-backed influence operations now circulates across journalism, policy, and the peer-reviewed literature: that they are monolithic troll armies; that they win by weaponizing moral-emotional language; that they manufacture their own virality; that they learn and optimize against feedback; and that they have become indistinguishable from ordinary users. Most rest on single-campaign studies, uncontrolled comparisons, and large-sample significance reported without baselines or multiple-comparison control, and are rarely re-tested. We assemble complete, government-attributed archives of seven state campaigns (25,076,853 tweets from 9,071 accounts) with a matched organic-user baseline for five of them, and re-test all five claims under one protocol: pre-registration, Benjamini–Hochberg false-discovery control, permutation nulls, a future-reception placebo, and a takedown-snapshot decomposition. Scoped to these campaigns, every claim weakens or reverses. The operations are narratively segregated, thinly staffed production desks, not a unified army; an organic moral-contagion law fails to replicate in any of them, and a meaningless placebo predicts engagement as well; internal amplification supplies only 0.10%–5.31% of top-percentile reach, the rest captured from an external audience the operation does not control; behavior is scripted, with rare apparent feedback mean-reverting toward baseline; and while their per-account language has drifted off the 2016 "troll" fingerprint, they still coordinate 7–70× more tightly than matched real users—a regularity a frozen re-test reproduces, with the language-drift and segregation patterns, across twelve further country-groups. The corrected picture is coherent: an industrial content factory *siloed in production but coordinated in execution*, whose reach it does not internally manufacture (its external composition, organic versus undetected coordination, left unresolved). The detectable signature has *migrated from language to coordination*: per-account content fingerprints age out, while cross-account coordination remains the durable, cross-national marker. Because these operations run scripts rather than optimize against feedback, an adversary that genuinely optimized—now feasible with large language models—would look measurably different from the operations studied here: a forward warning, not a present finding. The recurring lesson is methodological: on corpora of confirmed manipulation, baseline-free significance reconstructs the analyst's expectations, and platform and policy decisions rest on those beliefs.

Keywords: influence operations; coordinated inauthentic behavior; disinformation; social media

1. Introduction

Few online phenomena have been studied as intensively as state-backed influence operations (IO). Since the public release of the first platform-attributed archives, a research and policy ecosystem has grown around them, and with it a set of widely repeated claims about *how* these operations achieve influence. Many of these claims have hardened into "stylized facts": statements that circulate across investigative journalism, government and think-tank reporting, and the peer-reviewed literature with the confidence of settled knowledge.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

This paper examines five of them. (1) That an influence operation is a *monolithic troll army*, acting in concert. (2) That it *wins on emotion*, that moral-emotional and outrage-laden language is what drives the spread of its content. (3) That it *manufactures its own virality* through sockpuppet self-amplification. (4) That it is a *sophisticated, optimizing adversary* that learns from reception and adapts what it says to what works. (5) That it has become *indistinguishable from real users*, or, in the mirror-image folk belief, that it is still detectable by the crude linguistic tics of the 2016-era troll. Figure 1 previews each claim, its provenance, the test we apply, and the corrected finding; Table 1 gives the signature statistics.

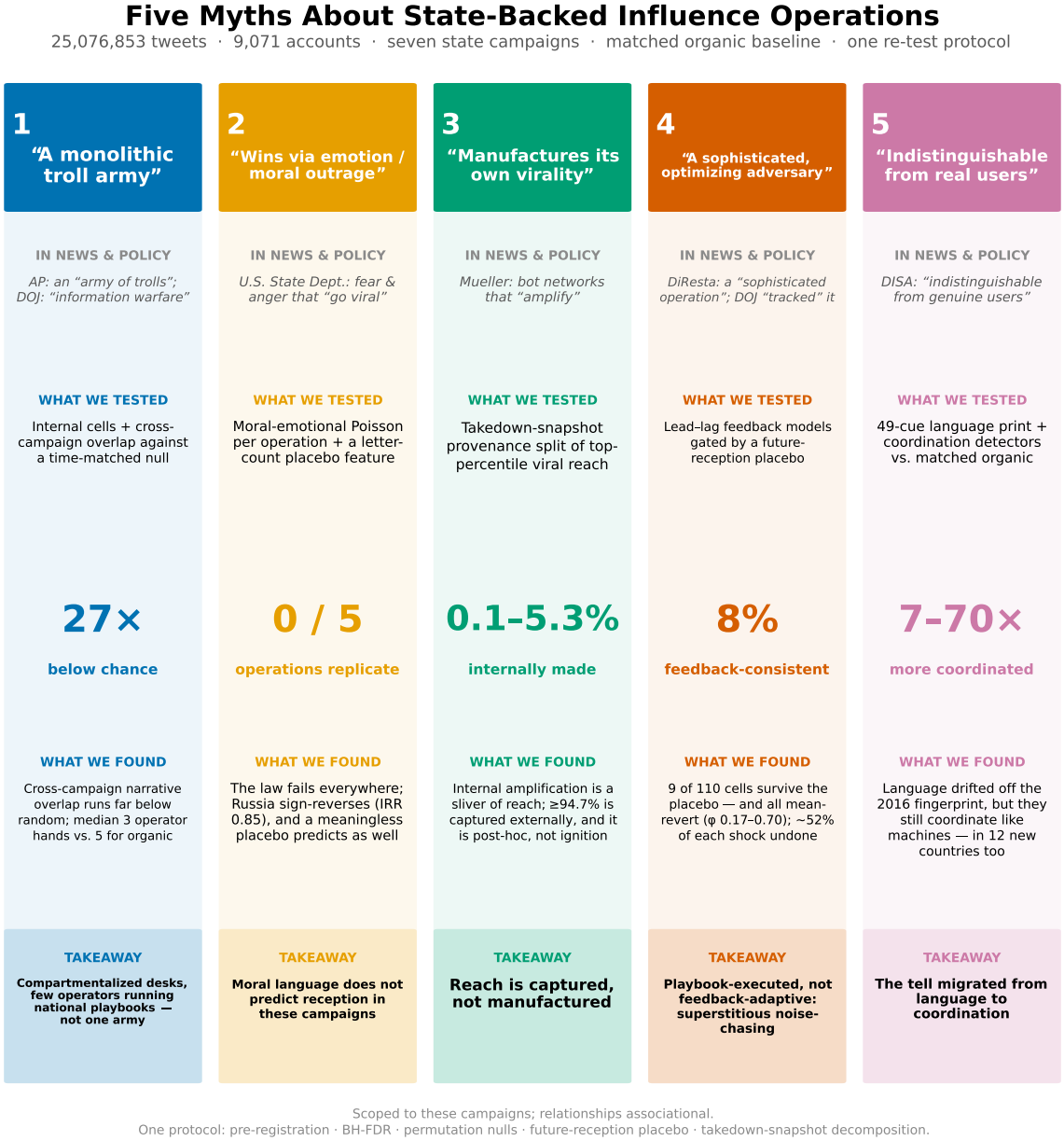

Figure 1. The five myths at a glance. For each: the folk belief and a representative verbatim framing of it in authoritative news or policy reporting (catalogued in Appendix A), what we tested, the headline finding when scoped to these seven campaigns, and the scoped takeaway. All relationships are associational; every statistic is detailed in the corresponding section.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

Table 1. The five myths at a glance: the folk belief, the corrected finding when scoped to these seven campaigns, and one signature statistic. Full evidence and section references follow.

Folk belief	Corrected finding (these campaigns)	Signature statistic
M1: a monolithic troll army	Narratively segregated, thinly staffed desks executing national playbooks	spanning $\sim 27\times$ below null; median 3 hands/desk
M2: wins on moral emotion	The moral-contagion law does not replicate; a meaningless placebo predicts as well	Russia IRR 0.848 vs. placebo 1.537
M3: manufactures its virality	Top-percentile reach is externally captured, not internally generated	internal share 0.10–5.31%
M4: an optimizing adaptive adversary	Scripted; the rare apparent feedback mean-reverts toward baseline	Galton $\phi \in [0.17, 0.70]$, all CIs < 1
M5: indistinguishable / a crude 2016 troll	Language drifted off the 2016 fingerprint, yet coordination stays many-fold above organic	synchrony 7.7–70.3$\times$; replicated in 12 countries

Each of these beliefs has a published provenance, and we attach that provenance to each myth rather than debunking a strawman: every section below opens by attributing its claim to at least one peer-reviewed or authoritative source, and wherever possible to a source that makes the claim *about state influence operations specifically* rather than about social media or human psychology in general. The point is not that these sources are careless. It is that the claims were, for the most part, established on evidence that cannot by itself separate a genuine signal from a measurement artifact: single campaigns rather than portfolios, the manipulation arm with no organic comparison, statistical significance on enormous samples reported without placebos or false-discovery control, and a near-total absence of re-testing. Beyond the scholarly provenance, each belief also circulates as received wisdom in journalism and policy reporting; Appendix A catalogues representative statements of each myth across authoritative news outlets and government, intelligence, and policy-institution documents.

The move.

We assemble *complete* archives (not samples) of seven government-documented influence campaigns released through the Twitter Information Operations program: 25,076,853 tweets posted by 9,071 accounts attributed to operations linked to Russia, Iran, Venezuela, and Bangladesh, spanning 2009–2018. We pair these with a matched organic-user baseline drawn from an independent release of same-country, same-period accounts that were never taken down [19]. Against this evidence base we re-test each of the five claims under a single, pre-registered rigor protocol described in Section 3. Where a claim was originally established on the manipulation arm alone, we re-establish it as a contrast against real users; where it rested on uncorrected significance, we apply false-discovery control and permutation nulls; and where it could be confounded by trend or regression-to-the-mean, we introduce placebo features that a genuine effect must beat.

Scope guard.

Three cautions bound every claim in this paper. First, our evidence concerns these *specific, already-investigated* campaigns; we make no claim about influence operations that platforms never caught, and "these operations do not do X" is never "no influence operation does X." Second, all reported relationships are *associational*: we describe what co-occurs, never what causes what. Third, finding that a popular belief is an artifact of these campaigns does not make the campaigns harmless: several of our corrected findings describe a *more* efficient adversary than the folk picture, not a less consequential one.

Contributions.

(i) A portfolio-scale, baseline-anchored, pre-registered re-test of five recurring IO claims, each weakened or reversed when scoped to these campaigns. (ii) A corrected, internally coherent mechanism picture: the compartmentalized, thinly staffed content factory whose reach is captured rather than manufactured. (iii) A reusable methodological lesson, instantiated repeatedly: baseline-free significance on a corpus of confirmed manipulation reproduces the very patterns analysts expect,

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

and only placebos, matched baselines, and pre-registration distinguish signal from mirage. (iv) An out-of-sample replication on twelve further country-groups establishing that the corrected regularities are cross-national, not idiosyncratic to the original four countries.

2. Related Work

From content to coordination.

Early empirical accounts of state-backed influence operations characterized campaigns through the content and behavior of disclosed troll accounts [1,13], while a parallel literature warned that automated amplification could distort public discourse [54,55]. Two developments have since reshaped the field. First, large-scale studies of organic diffusion indicate that human resharing, not automation, accounts for much of the differential spread of low-credibility content, with bots accelerating true and false material comparably [30]. Second, the disclosed operations proved to be staffed largely by human-operated accounts whose individual behavior resembles that of ordinary users [14,18], leaving individual-account features a weak discriminator. Together these results shifted the analytic target from *what* operation accounts say toward *how* they coordinate.

Network signatures of coordinated inauthentic behavior.

A now-substantial body of work detects coordinated inauthentic behavior through behavioral synchrony rather than content alone, fusing co-retweet, co-URL, text-reuse, and hashtag-sequence similarity into account-similarity networks [23,26], and extending these signatures across platforms [24, 52]. Recent work emphasizes that coordination is temporally unstable and decomposes into behavioral archetypes, so that static snapshots can misrepresent its structure [25]. In parallel, classifiers trained on archived disclosures generalize across campaigns by exploiting shared operational tactics [20,22], including graph-learning detectors that generalize inductively across operations [42], and explainable detectors recover interpretable operator signatures [21]. The recurring lesson is that coordination intensity, rather than any single linguistic feature, is the most durable indicator of an operation.

What actually drives reach and engagement.

The mechanisms behind online attention have been progressively disentangled. Engagement appears most strongly associated with out-group animosity [28] (though in-group solidarity can dominate engagement around conflict events [44]) and, in causal settings, with negativity [29], while misinformation may spread in part by evoking moral outrage that prompts resharing without reading [33]. The robustness of moral-emotional "contagion" is itself contested: a pre-registered replication and meta-analysis by the original authors reports a positive but small and heterogeneous effect, including a sign reversal in one corpus [27], a fragility that motivates placebo-controlled re-testing rather than reliance on a single headline effect. Linked exposure–outcome studies of the Russian Internet Research Agency find that exposure was concentrated among a few strongly partisan users, small relative to domestic media exposure, and not measurably associated with shifts in attitudes or voting [31,32].

Generative AI and the limits of "sophistication."

The generative-AI turn has prompted concern that operations will scale and adapt [56,57]. The evidence is mixed: AI-generated propaganda can approach human-authored material in persuasiveness, but typically requires human curation to do so [34], and machine-generated text can be perceived as at least as credible as human-written text [48]; large randomized trials find that LLM-based microtargeting does not reliably outperform untargeted messaging [35], although conversational LLMs supplied with basic personal information can out-persuade humans in dialogue [47]. These results temper the framing of influence operations as sophisticated, autonomously adaptive optimizers.

Reproducibility and the baseline problem.

Computational social science has increasingly confronted measurement-validity and reproducibility concerns, including demonstrations that analytic flexibility can manufacture almost any effect from

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

a single dataset [36,49], and that reporting results across the full set of defensible specifications is a corrective [50]. A recurring finding is that claims about influence operations can change substantially once an appropriate null or organic baseline is introduced: reported inter-state coordination among separate operations [37], for instance, is substantially attenuated under proper baseline comparison [38]. Building on this trajectory, we re-test five recurring claims under a pre-registered protocol that anchors every comparison to a matched organic baseline and treats coordination intensity, not content, as the primary observable.

3. Data and Methods

3.1. Campaigns

The primary corpus comprises seven state-attributed campaigns released through the Twitter Information Operations archive [58]: two operations attributed to the Russian Internet Research Agency and associated Russian activity, two attributed to Iran, two attributed to Venezuela, and one to Bangladesh. In total the corpus contains 25,076,853 tweets from 9,071 released accounts (8,275 of which authored at least one tweet), spanning posting dates from 2009 to 2018. Because the data are takedown snapshots, engagement counters reflect their state at suspension; we treat this explicitly (Section 3.5) rather than ignoring it. Throughout, accounts and internal groupings are referred to only by pseudonymous desk or cluster identifiers; no real handles or numeric account identifiers appear in this paper. Table 2 summarizes the seven campaigns; account and tweet totals are public Twitter-release metadata.

Table 2. The seven state-attributed campaigns in the primary corpus (Twitter Information Operations archive). "Released" counts all disclosed accounts; tweet counts are the released totals. "Organic control" indicates whether a matched, never-suspended same-country/period arm is available for the Myth 1/3/5 contrasts.

Operation (attribution)	Accounts	Tweets	Span	Organic control
Russia (IRA)	3,608	8,768,633	2009–2018	Yes
Russia (other)	416	765,246	2010–2018	Yes (power-flagged)
Iran (Oct-2018)	770	1,122,936	2010–2018	Yes
Iran (Jan-2019)	2,311	4,447,056	2009–2018	Provisional (not pooled)
Venezuela-1	1,196	8,961,788	2010–2018	No (descriptive)
Venezuela-2	755	984,980	2015–2018	Yes
Bangladesh	15	26,214	2009–2018	Below network power floor
Total	**9,071**	**25,076,853**	2009–2018	5 usable + 1 provisional

3.2. Organic Baseline

For five of the seven campaigns we draw a *matched organic-control* arm from an independent release of accounts that were active in the same country and period but were never taken down [19]. "Matched" here means matched on topic/period activity, not on demographics or account age. This baseline is what converts a within-corpus description ("IO accounts do X") into the comparison that the claims actually require ("IO accounts do X more than comparable real users"). One campaign (the largest Venezuelan operation) has no available control and is reported descriptively; one Iranian mapping is provisional and is never pooled into the confirmed-five contrasts. All organic-control results are reported at the aggregate class level only (no per-account dossiers, rankings, identifiers, or text), consistent with the control accounts being ordinary users and with the baseline's license terms.

3.3. Rigor Toolkit

A shared protocol underlies every test:

- **Pre-registration.** Each analysis run committed a timestamped pre-registration file (fixing its hypotheses, estimators, feature definitions, and pass/fail decision rules) to the project repository before any model was fit on the outcome data, and a fixed random seed was used throughout for

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

reproducibility. Deviations forced during analysis (for example, a campaign dropping below the power threshold for a given test) are logged as registered deviations rather than silently absorbed.

- **Multiple-comparison control.** Benjamini–Hochberg false-discovery control is applied *within* each pre-registered test family, and we report q-values rather than raw p-values for those families. A family is the full set of campaign-level (or pair-level) contrasts for one research question, so that, for instance, all cross-campaign sharing tests form one family and all moral-contagion specifications another. Where the design admits only a small number of matched control draws (the Seckin baseline supplies ten, giving a one-sided add-one p-floor of .0909 that BH cannot push below .05), we say so explicitly and carry the conclusion on confidence-interval exclusion of the null ratio rather than on a q-value.
- **Permutation and rewiring nulls.** Reference distributions for the network and segregation statistics are built by resampling rather than assumed: recovered coordination cells are validated against a *degree-preserving rewiring* null [59] that holds each account's connectivity fixed while randomizing partners; cross-campaign narrative, domain, and target overlap is tested against a *time-matched* (and, separately, popularity-matched) null that preserves each campaign's activity volume and timing; nulls use ≥ 1000 draws where the per-permutation cost permits, with the smaller draw counts used for the most expensive tests disclosed in place and framed as exact tests.
- **Effect sizes with uncertainty.** We report effect sizes (incidence-rate ratios, Cliff's δ, abnormality ratios) with bootstrap confidence intervals rather than p-values alone.
- **Three signature placebos/decompositions.** (a) A *future-reception placebo* for the learning test (Section 7): because future reception cannot precede present behavior, a genuine feedback signal must show a large dependence on *past* reception and a near-zero dependence on *future* reception; where the two are equal, the apparent "learning" is trend or autocorrelation. (b) A *takedown-snapshot decomposition* for the virality test (Section 6), isolating the one campaign whose snapshot permits a clean external/internal subtraction. (c) A *letter-count placebo* and *dual moral-foundations dictionaries reported side-by-side* for the emotion test (Section 5).

3.4. Estimands and Test Statistics

Four estimands recur across the five tests. First, reception is modeled per operation i with a no-follower fixed-effects Poisson regression,

$$\mathbb{E}[y_{it}] = \exp\left(\alpha_i + \beta\, x_{it} + \boldsymbol{\gamma}^{\top}\mathbf{z}_{it}\right), \tag{1}$$

where y_{it} is the engagement count of tweet t, x_{it} the standardized feature of interest (moral-emotional load, or the letter-count placebo), $\mathbf{z}_{it}$ persona controls, and the reported effect is the incidence-rate ratio $\mathrm{IRR} = e^{\beta}$ per standard deviation. The future-reception placebo (Myth 4) re-fits a lead–lag form and contrasts the coefficient on *past* reception, β_{past}, with that on *future* reception, β_{fut}: a genuine feedback signal requires $\beta_{\mathrm{past}} \gg \beta_{\mathrm{fut}} \approx 0$, whereas $\beta_{\mathrm{past}} \approx \beta_{\mathrm{fut}}$ indicates trend or autocorrelation.

Second, coordination and overlap statistics are scored against resampled nulls as a lift,

$$\mathrm{lift} = \frac{S_{\mathrm{obs}}}{\mathbb{E}_{\mathrm{null}}[S]}, \tag{2}$$

with S a coordination or cross-campaign-overlap count and the null built by degree-preserving rewiring (internal structure) or time-/popularity-matched resampling (cross-campaign); lift < 1 with the observed value below the null interval indicates segregation rather than convergence.

Third, the matched-baseline contrast (Myth 5) is an abnormality ratio per detector d,

$$A_d = \frac{S_d^{\mathrm{IO}}}{S_d^{\mathrm{organic}}}, \tag{3}$$

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

declared abnormal-high only when the bootstrap 95% confidence interval of A_d excludes 1 *and* $A_d > 1$.

Fourth, whether reception-chasing yielded a durable benefit (Myth 4) is read from a Galton mean-reversion coefficient,

$$r_{t+1} = \phi\, r_t + \varepsilon_t, \tag{4}$$

where r_t is the rank-reception of a reallocated-toward content class in window t: $\phi \to 1$ would indicate a persistent, rational hold-up, while $\phi < 1$ with its bootstrap interval strictly below 1 indicates mean-reversion (a transient response to noise). Within each pre-registered test family, the resulting p-values are converted to Benjamini–Hochberg q-values [51].

3.5. Measurement Honesty

Three limitations are stated explicitly rather than understated. The engagement counters are takedown-final snapshots, a proxy we validate where possible but never treat as live telemetry. The external audience of these operations is structurally unobservable: the archives record only the operations' own accounts, so an outside user who amplified an operation's content leaves no row, a hard ceiling we return to in Sections 6 and 10. And the organic baseline's license forbids redistribution, so only derived aggregate statistics are reported. None of these constraints is a reason to withhold the analysis; each is a reason to scope its claims precisely.

4. Myth 1: "A Monolithic Troll Army"

The claim and its provenance.

In popular and policy discourse the "troll factory" is imagined as a single, unified army acting in concert (Appendix A). The foundational scholarship is in fact more specific: Linvill and Warren, analyzing the IRA's English-language activity, describe an *industrial* operation "mass produced from a system of interchangeable parts, where each class of part fulfilled a specialized function," identifying distinct handle categories [1]. The myth, then, is the popular reading of "troll factory" as one undifferentiated mass; the published account already points toward specialization. Recent network studies reinforce this reading: coordinated communities decompose into behavioral archetypes rather than a single mass [25,39], while reported coordination among separate operations [37] weakens once proper baselines are applied [38]. Our task is to measure that compartmentalization directly and ask what it implies about how the factory is staffed.

The test.

We reconstruct each operation's internal structure from co-activity, shared-infrastructure, and content-overlap networks, validating recovered cells against a degree-preserving rewiring null. We then measure cross-campaign overlap of narratives, domains, and external retweet targets against a time-matched null and, as a production-side capstone, estimate how many distinct authorial "hands" sit behind each desk [40] using a joint stylometric-and-behavioral clustering with the number of operators selected by prediction strength and bracketed by bootstrap.

The finding.

The operations are not one army; they are *narratively segregated desks*, and the segregation is far below chance. Across the portfolio, cross-campaign near-duplicate narrative spanning is observed at 39,113 shared clusters against a time-matched null of roughly 1,064,682, a lift of 0.037, i.e. about 27× *below* chance (Figure 2). Operations share narratives, domains, and external targets *less* than random pairs would: 12 of 21 campaign pairs share fewer domains, and 14 of 21 fewer external retweet targets, than popularity-matched chance. The internal topology is national-playbook-specific: the Iranian operations are genuinely cellular (one resolves into eight language-aligned consensus cells); the IRA is a single dominant fabric concealing a 569-account Russian-language retweet desk that is Cyrillic-dominant (Cliff's $\delta = +0.61$), distinct in client mix (Jensen–Shannon divergence 0.54 against a null maximum of 0.01, $p = .001$), yet keeps the same office-hour rhythm as the English side: one organization running parallel

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

desks. The only reliable cross-campaign link is between the two Iranian operations (shared-domain lift 1.42, $q = .018$; shared external-target lift 1.52; six bespoke clients exclusive to the pair). And campaign identity explains roughly twice as much cell-feature variance as automation level (adjusted R^2 0.155 vs. 0.084): the desks are campaign-shaped first, automation-shaped second.

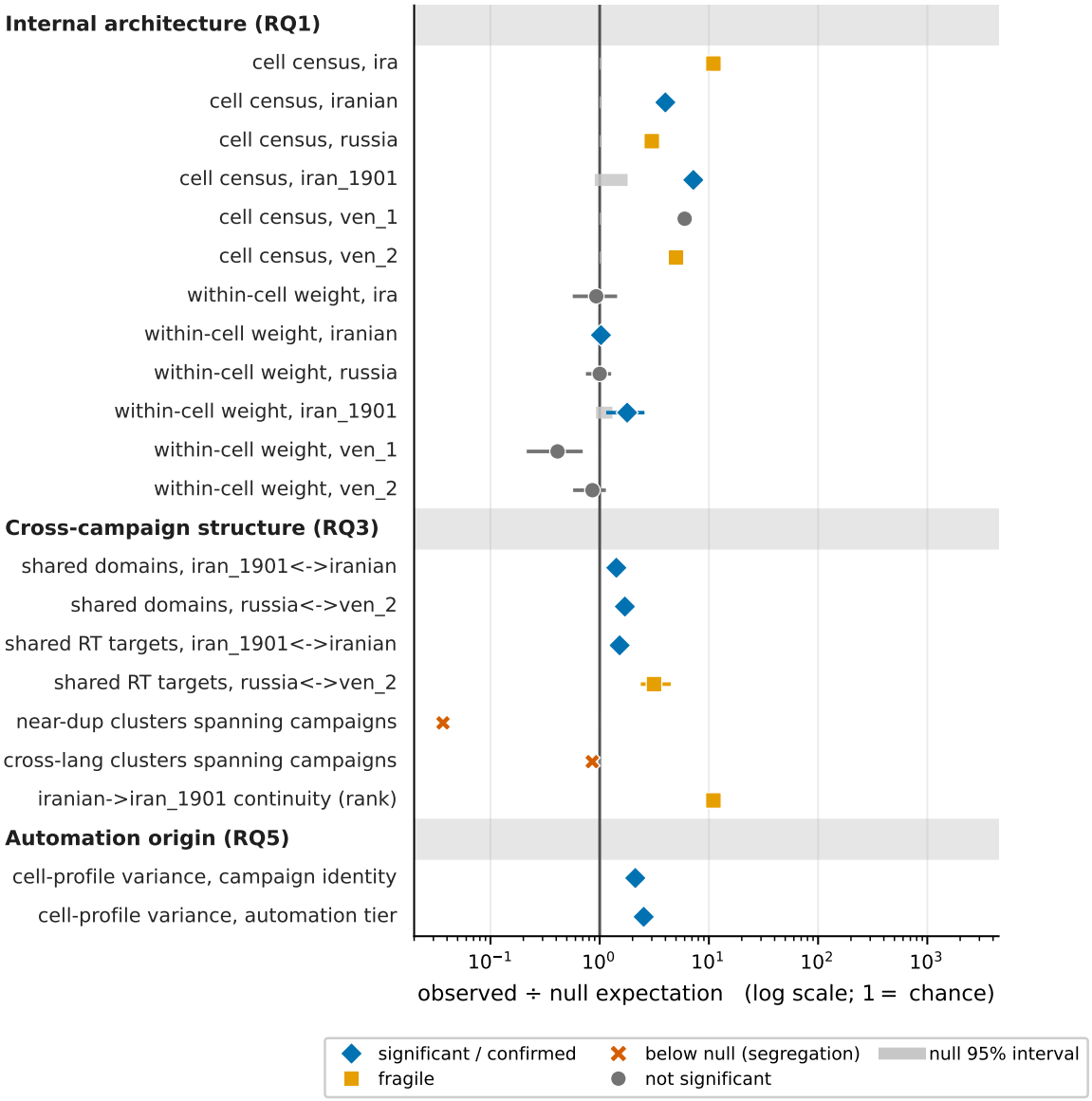

Figure 2. Myth 1. Observed cross-campaign narrative spanning (and related coordination statistics) against time-matched and degree-preserving nulls. Cross-campaign narrative overlap falls about 27× *below* chance: the operations are mutually segregated, not a single coordinated army.

How many hands.

Behind these large multi-account desks sit notably few operators. A joint style-and-behavior clustering returns a single stable signature ($\hat{k} = 1$) for 20 of 33 analyzable desks, with every desk at $\hat{k} \leq 6$; the two largest IRA desks (1,249 and 1,100 active accounts) and the largest Venezuelan desks each collapse to a single hand. Because short, multilingual tweets under-individuate authors (the same pipeline collapses *organic* accounts too), the defensible comparison is the absolute operator count at matched sample size: IO desks carry a median of 3 distinct hands versus 5 for comparable organic crowds (one-sided Mann–Whitney $p = 0.023$). A synthetic positive control confirms the estimator tends to *over*-split, making "few hands" a conservative reading. Staffing density is itself playbook-specific (a $\sim 5\times$ spread): the scripted IRA and Venezuelan fabrics show the fewest hands

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

per account, the cellular Iranian operations the most. A neural authorship-style representation (LUAR content-independent embeddings) [60], computed independently on the same desks, reproduces the collapse and is if anything sharper (24 of 33 desks at a single neural signature; median one hand in both arms); the two representations agree at the aggregate level but not on the precise per-desk count (Spearman $\rho = -0.08$), so we report "hands" as a *range, not a headcount*.

Convergent validity against independent hand labels.

Because Linvill and Warren assigned each IRA account to a functional category *by hand* [1], their labels furnish an external check on desks we recovered from coordination and style with no access to those labels. Matching their public account-category release to the IRA arm by screen name returns a role for 2,604 of our IRA accounts (91.6% of their distinctly labeled handles; the shortfall concentrates in the RightTroll category, consistent with its heavier suspension and handle turnover). The hand labels fall *along* our recovered structure rather than across it (Table 3). The two giant single-hand desks split by language exactly as the manual categories do: the Russian-language desk (giant hand #2) is essentially entirely NonEnglish, while the English-language desk (giant hand #1) absorbs all four English functional categories—RightTroll, LeftTroll, HashtagGamer, and Fearmonger. Two readings follow. First, a desk reconstructed without the hand labels reproduces the first cut an independent team drew manually. Second, and more consequentially, that single English desk is one *stylometric* hand yet co-hosts all four English categories: the much-discussed handle "types" are job functions resident in a single authoring desk, not separate armies—real at the level of content function, collapsed at the level of authorship. The lone partial exception, NewsFeed (automated headline-amplifier accounts), sits mostly in smaller desks, consistent with a low-coordination feed function rather than a manned role. This check is confirmatory and IRA-only—no comparable external hand labeling exists for the other operations—and alters no primary estimate.

Table 3. Myth 1, convergent validity. Share of each Linvill–Warren [1] hand-labeled IRA category whose accounts fall in the two giant single-"hand" desks recovered here (row %; $n = 2{,}604$ IRA accounts carrying an external label). Pseudonymous desk IDs; IRA-only; associational.

Hand label (L&W)	D0 (Eng.)	D1 (Rus.)	Reading
RightTroll	61.5%	0.0%	English partisan → English desk
LeftTroll	53.6%	0.0%	English partisan → English desk
HashtagGamer	95.5%	0.0%	trend amplifier → English desk
Fearmonger	78.4%	0.0%	English alarm → English desk
NewsFeed	27.8%	0.0%	mostly smaller automated desks
NonEnglish	0.3%	74.0%	Russian-language → Russian desk

Scoped takeaway.

These operations are compartmentalized *production* run by *few operators executing national playbooks*, not a unified army. This finding extends rather than contradicts the specialization Linvill and Warren first described, and quantifies just how thinly the factory is staffed.

The architecture, drawn.

Figure 3 renders this compartmentalization directly for the two largest operations. Each node is an account and each edge a within-operation co-retweet; node colour marks the largest coordination desks the clustering recovers, and node size scales with within-operation degree. The IRA (top) is a single dense fabric that nonetheless braids two co-equal desks: a low-retweet, hashtag-driven English-language content desk and a high-retweet, link-amplifying Russian-language desk—the same D0/D1 division that the Linvill–Warren overlay in Table 3 independently labels English versus non-English. Iran (January 2019, bottom) instead resolves into functionally distinct cells—a retweet-amplifier cell, a reply/engagement cell, and a hashtag-campaign cell—rather than one undifferentiated mass. The two operations are wired along visibly different national playbooks, yet each is a compartmentalized production floor of specialized desks rather than a unified army; the contrast between a single braided fabric and a set of separated cells is exactly the national-playbook specificity the cell statistics report.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

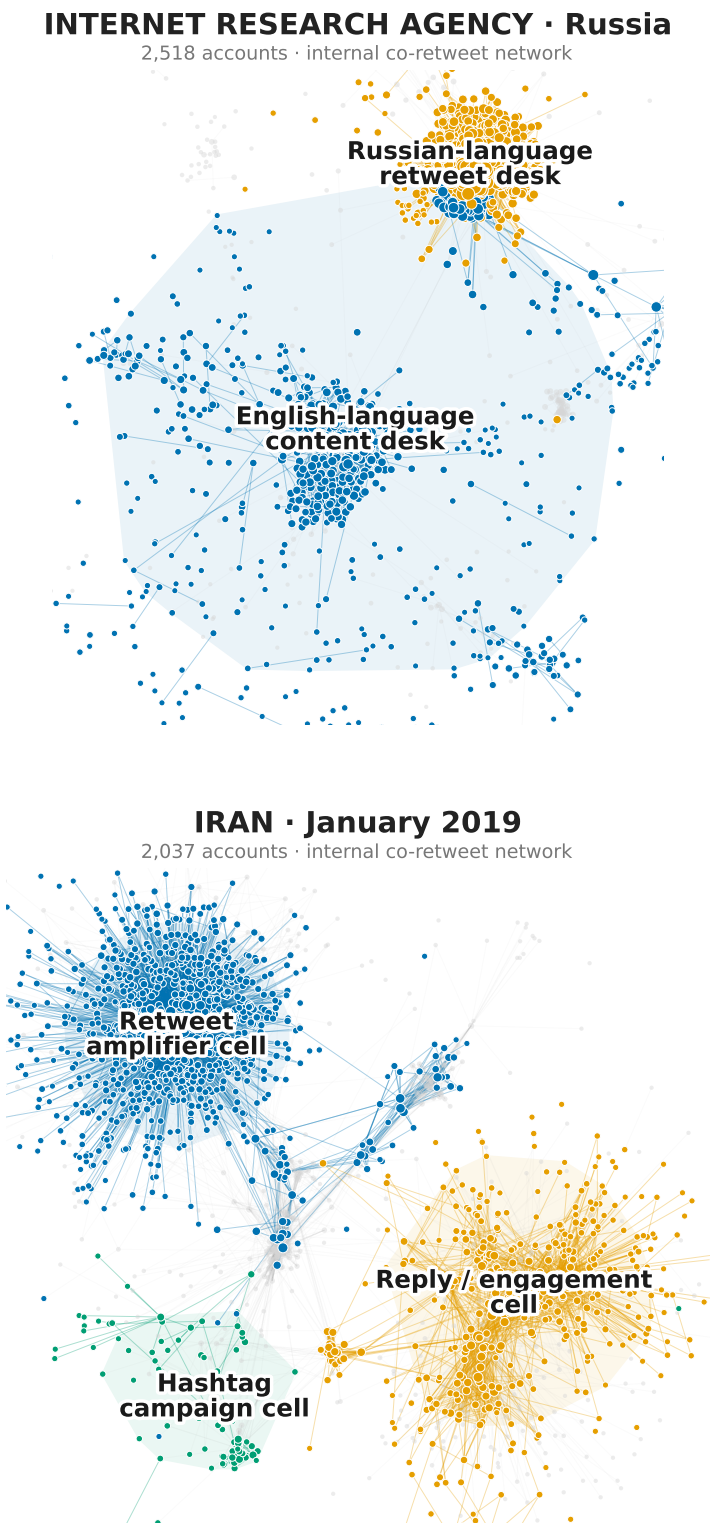

Figure 3. Internal co-retweet coordination networks of the two largest operations (top: the IRA; bottom: Iran, January 2019). Nodes are accounts; edges are within-operation co-retweets; node size scales with within-operation degree; node colour marks the largest coordination desks (with soft "territory" hulls, and labels derived from each desk's retweet/reply/link/hashtag behaviour and, for the IRA, the Linvill–Warren overlay of Table 3), while smaller desks and unclustered accounts are grey. Layout is force-directed (Fruchterman–Reingold), shown for accounts in connected components of at least eight. The IRA forms one dominant fabric that braids two co-equal desks (the D0/D1 split of Table 3); Iran resolves into functionally distinct cells—the same compartmentalization, organized along a different national playbook. Pseudonymous desk identifiers only; all relationships associational.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

5. Myth 2: "Wins via Emotion / Moral Outrage"

The claim and its provenance.

Having established what the factory *is*, we turn to what it produces and whether that content earns attention. The general principle is Brady and colleagues' moral-contagion law: in organic networks, each added moral-emotional word is associated with substantially greater diffusion [3], an effect whose expression is itself amplified by social-feedback learning [43]. The belief that this law drives influence operations has been instantiated directly on real troll data: analyzing the released IRA Twitter corpus, Suk and colleagues report that "negative tweets had 1.206 times the rate of retweets than those without negative sentiment" [2], and the policy literature describes the operations as engineering emotionally resonant memes and human-interest content for spread [5] (Appendix A). There is also independent reason for caution: Burton, Cruz, and Hahn show that the contagion model can perform no better than an implausible alternative [4], precisely the fragility we test for inside state propaganda. A large pre-registered replication and meta-analysis by the original authors likewise finds the effect positive but small and heterogeneous, even reversing sign in one corpus [27], while adjacent work locates the engagement lever in out-group animosity [28] and negativity [29] rather than in moral-emotional content as such.

The test.

On the 16,459,645 original (non-retweet) tweets in the corpus, we fit the moral-emotional reception model per operation, using fixed-effects Poisson with persona controls. We report two moral-foundations dictionaries side-by-side, never averaged because they genuinely diverge (mean Spearman $\rho = 0.25$), and, critically, a *letter-count placebo*: a meaningless feature that should predict nothing if the corpus is well-behaved.

The finding.

The organic moral-contagion law does not replicate in any operation (Figure 4). Against an organic anchor incidence-rate ratio (IRR) of about 1.13, the Russian operation *sign-reverses*: moral-emotional wording is associated with *less* engagement net of persona (IRR 0.848, 95% CI $[0.810, 0.889]$, 0 of 9 specifications positive), while its letter-count placebo is strongly positive (1.537). The corpus manufactures a spurious positive effect for a meaningless feature even where morality is genuinely negative. The Iranian (Oct-2018) operation's moral effect (IRR 1.083, $[1.025, 1.145]$) is statistically *indistinguishable* from its letter-count placebo (1.103, $[1.053, 1.155]$). The remaining operations sit weakly positive but below the organic anchor and outside its confidence interval (Iran 1.037, Venezuela 1.044), or at the reception floor. This is the "large-corpus mirage" Burton and colleagues warned of, demonstrated inside state propaganda: on samples this large, baseline-free significance attaches to noise.

Scoped takeaway.

Moral-emotional language does not predict reception *in these campaigns' data*, and a meaningless placeholder predicts as well. This is a claim about these operations and this measurement, *not* that affect or morality is irrelevant to virality in general. One methodological caveat distinguishes this myth from the others: because the organic baseline lacks engagement counts (Section 10), Myth 2 alone is *placebo-anchored rather than baseline-anchored*. We do not contrast IO against a like-for-like organic reception model; instead the letter-count placebo shows that the apparent moral-emotional effect is not feature-specific: a meaningless placeholder earns the same or larger coefficient on the same samples. That is sufficient to defeat the specific claim (that *morality* drives reception in these operations) without resting on a within-corpus positive, but it is a weaker instrument than the matched-baseline contrasts that carry Myths 1, 3, and 5, and we flag it as such. If anything, the prior literature sharpens the result: even the IO-specific instantiation finds sentiment to be one driver among informational and topical

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

ones, with positive sentiment slightly *suppressing* retweets [2], so the belief we test is narrow, and it does not hold here.

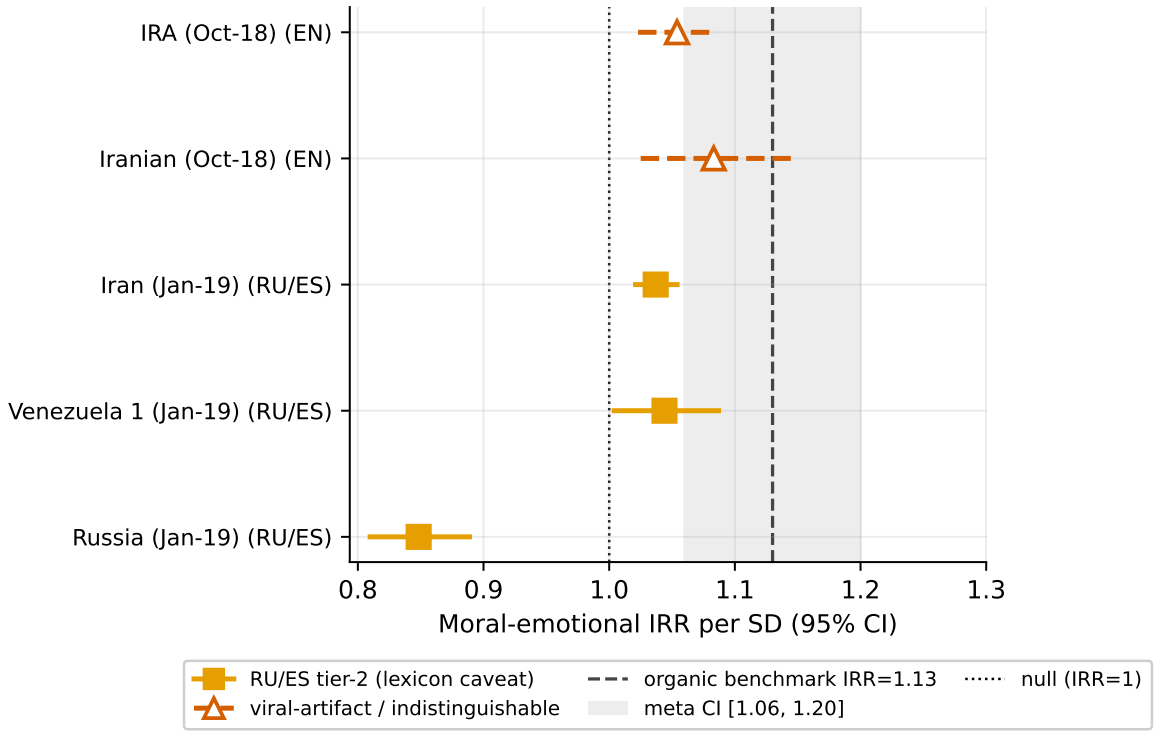

Figure 4. Myth 2. Moral-emotional reception (IRR) by operation against the organic anchor and the letter-count placebo. The contagion law fails to replicate everywhere; Russia sign-reverses, and the placebo predicts engagement as well as morality.

6. Myth 3: "Manufactures Its Own Virality"

The claim and its provenance.

A central image of influence operations is sockpuppet self-amplification: the operation engineers its own viral ignition (Appendix A). The mechanism is well established in the diffusion literature: automated accounts amplify low-credibility content, especially early in a cascade and by targeting influential users [7], and coordinated accounts can boost a cascade's reach up to a saturation threshold [8]. Against this, linked exposure studies find that Russian IRA reach was concentrated among a few partisan users and dwarfed by domestic media, with no measurable attitudinal effect [31], that differential spread is carried more by human resharing than by automation [30], that automated accounts are less central to diffusion than verified or high-profile human accounts [46], and that amplification, where present, often runs through cross-platform laundering of external media [24]. The question is how much of these operations' *actual* viral reach this internal machinery accounts for.

The test.

We decompose the reach of each operation's top-percentile viral originals into internally manufactured amplification (retweets by the operation's own accounts) versus externally captured reach, using a takedown-snapshot decomposition (the IRA snapshot permits a clean external subtraction). We then ask whether the internal amplification that *is* present looks like synchronized manufactured ignition or like diffuse after-the-fact sharing, by relating within-author seeding synchrony to external reach.

The finding.

Internal manufacture is a small fraction of viral reach (Figure 5). Across operations it supplies 0.10% (IRA), 0.57% (Iran, Jan-2019), 0.69% (Iranian, Oct-2018), 0.78% (Venezuela-1), and 5.31% (the Russian hub) of top-percentile reach; the captured share is therefore $\geq$ 99.2% everywhere except the

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

Russian hub (94.7%), and the per-original median manufactured share is 0 in every stratum, consistent with prior findings that the bulk of state-propaganda dissemination is carried by ordinary users rather than the operation's own accounts [9], and that dissemination of low-quality content concentrates in a small set of ordinary "supersharer" users [45]. Internal amplification *is* enriched among the winners (1.7–12.5×), but enrichment is consistent with either manufactured ignition or undetected external coordination, and it does not account for the reach. The enrichment looks like *diffuse post-hoc sharing, not synchronized ignition* (Figure 6): within-author seeding synchrony is associated with *fewer* external accounts, not more, in the IRA (IRR 0.68, $q = .0004$) and Iran (0.93, $q = .003$), null in Venezuela, and positive in only one operation, the confirmed Iranian (Oct-2018) operation (IRR 1.27, $q < 10^{-11}$), the lone ignition-like signature. (This ignition result, and the cellular Iranian structure noted under Myth 1, both derive from the operations' own released archives; the *provisional* Iranian mapping noted in Section 3 concerns only the matched-control contrast of Myth 5, where it is flagged in place.) Within-author seeding synchrony here means same-author co-timed posting; it is distinct from the cross-account same-minute synchrony that distinguishes IO from organic users under Myth 5: the former anti-predicts reach, the latter marks inauthenticity. External reach tracks the *breadth* of amplifiers, not their same-minute synchrony. Producer and amplifier roles are distinctly divided by playbook (Cramér's $V = 0.304$, a moderate association), from a 12-account Russian seeding squad to a 717-account Iranian one.

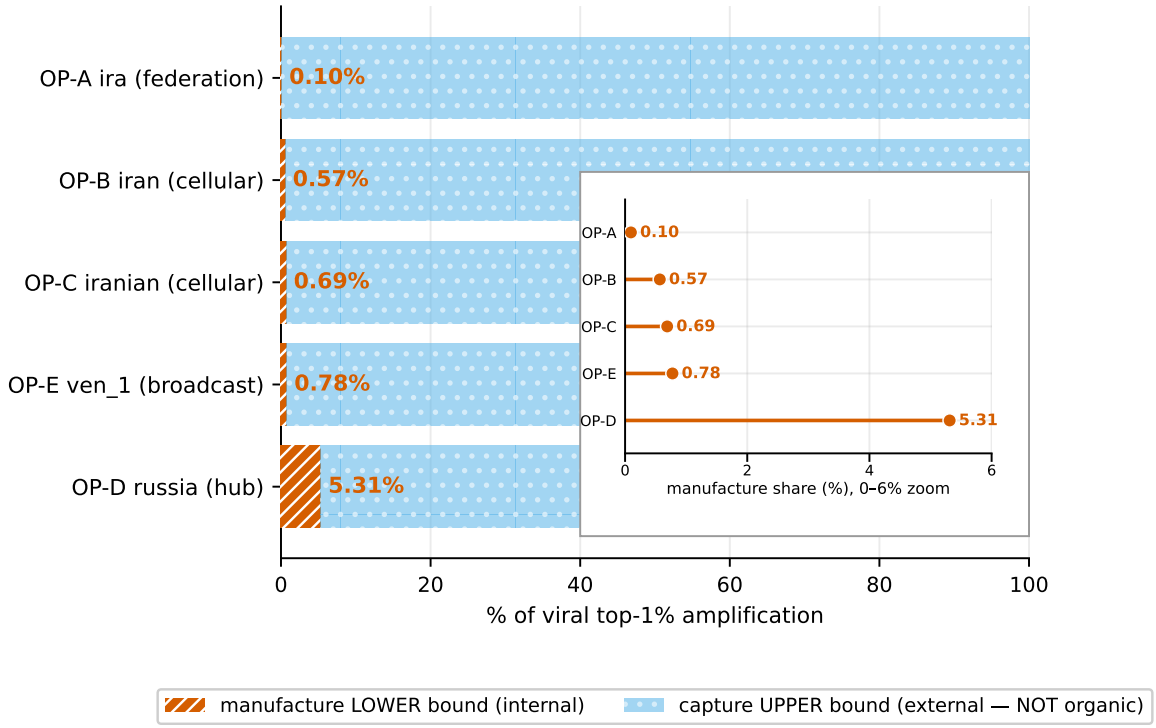

Figure 5. Myth 3, provenance decomposition of top-percentile viral reach by operation. Internally manufactured amplification (manufacture *lower* bound, 0.1–5.3%) versus externally captured reach (capture *upper* bound, ≥ 94.7%); the external bound is not "organic": it mixes genuine users, unrelated bots, and any undetected coordination the archive cannot see. The inset re-plots the manufactured share alone on a 0–6% axis, where the Russian hub (5.3%) stands well above the others (all ≤ 0.8%). Pseudonymous operation labels; associational throughout.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

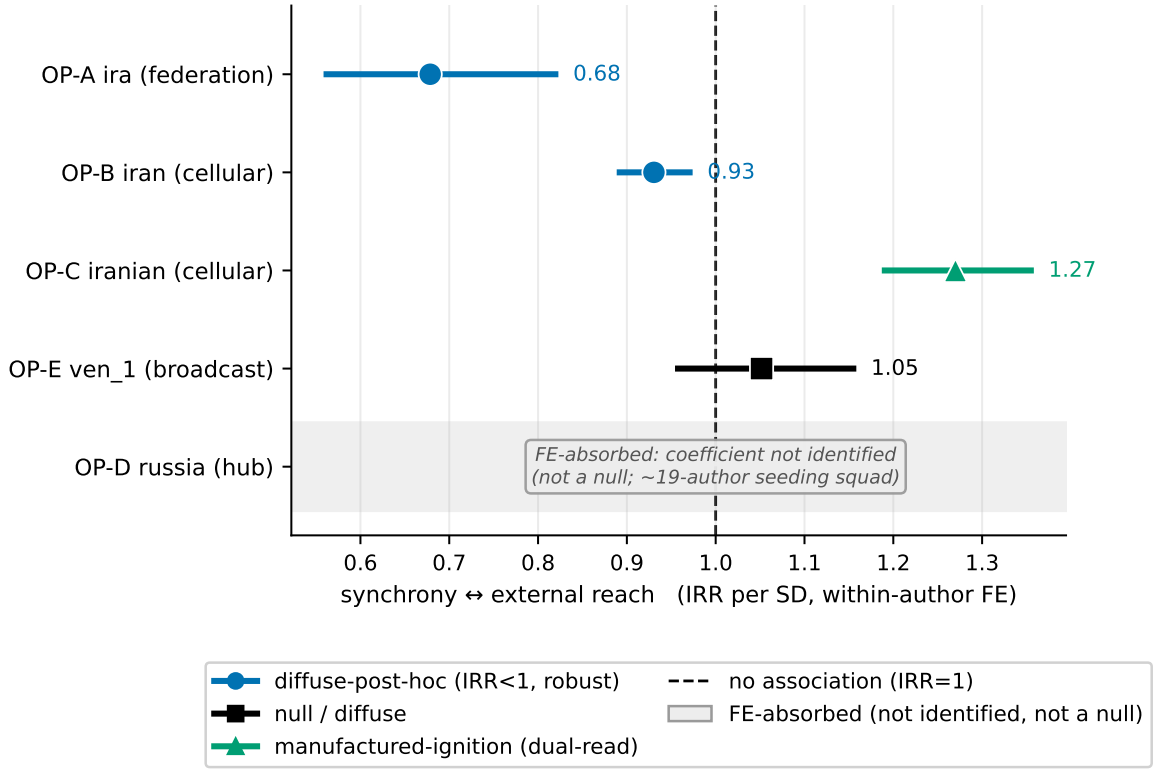

Figure 6. Myth 3, ignition verdict: the association between within-author seeding synchrony and external reach (incidence-rate ratio per SD, within-author fixed effects). IRR < 1 (IRA, Iran) indicates diffuse post-hoc sharing rather than synchronized ignition; the Iranian (Oct-2018) operation is the lone ignition-like positive; the Russian hub is fixed-effect-absorbed (not identified, not a null). Winner-enriched internal amplification is diffuse post-hoc sharing, not synchronized manufactured ignition. Pseudonymous operation labels; associational throughout.

Scoped takeaway.

Reach in these operations is *captured, not manufactured*, which is not "no reach." We state the bounding limit plainly and quarantine it in Section 10: because the archives record only the operations' own accounts, the external pool cannot be split into genuine organic uptake versus undetected coordination the corpus sees only the internal tip of. The captured/manufactured contrast holds; the internal composition of the captured share is not resolvable here.

7. Myth 4: "A Sophisticated, Optimizing, Adaptive Adversary"

The claim and its provenance.

Having found that reception (Myth 2) and viral reach (Myth 3) lie largely outside these operations' control, we now ask whether they nonetheless reallocate toward them. The operations are widely described as learning machines that optimize against feedback (Appendix A). The authoritative policy account portrays the IRA as "run like a sophisticated marketing agency" that "developed their content using digital marketing best practices" [5], and the doctrine literature calls Russian propaganda "remarkably responsive and nimble" [6]. The academic literature makes a weaker, temporal claim (that tactics, language, and identities change over time [12,13], which we do not dispute), while our own prior work finds trolls act "regardless of the feedback" they receive from genuine users [14] and that automated accounts lack the time-varying behavioral dynamics that characterize genuine human activity [11]. Consistent with a scripted reading, a campaign-agnostic classifier across nineteen state operations attributes their detectability to shared, templated tactics [22], field experiments find no measurable persuasive effect from IRA contact [32], and even large-language-model microtargeting does not reliably outperform untargeted messaging [34,35]. We test the strong belief: do operators reallocate content toward what earns reception?

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

The test.

Exploiting the temporal axis, we fit lead–lag reception-to-behavior models at the operation, desk, and account levels, and gate every apparent feedback signal through the future-reception placebo. For the cells that survive, we ask whether reception-chasing yielded a durable benefit: whether a reception shock persisted or mean-reverted.

The finding.

The operations are scripted, not feedback-adaptive (Figures 7 and 8). Of 110 feasible operation-level cells, only 9 (8%) are genuinely feedback-consistent once the placebo is applied; 18 (16%) are the corpus's *largest* apparent-adaptation coefficients, unmasked by the placebo as time-symmetric trend; and 83 (76%) are flat. Eight of the nine surviving cells are surface-presentation axes (whether to include a URL, hashtag, or mention), never content; the cleanest case (IRA URL-formatting) shows the canonical signature (past-reception coefficient $\approx +1.3$, placebo coefficient slightly negative). And every feedback-consistent cell *mean-reverts*: Galton coefficients of 0.17–0.70 with every bootstrap interval strictly below 1, a median $\sim$ 52% of each reception shock reverting within one window: transient responses to noise rather than durable optimization, with no persistent hold-up. The IRA shows content drift *negatively* associated with its own reception gradient: it moves away from, not toward, the agendas that earned the most engagement. Feedback, where present, is centralized at the operation or desk level and essentially absent at the account level (the feedback-consistent count collapses $9 \to 6 \to 2$ from operation to desk to account). Without the placebo, the 9 genuine cells and the 18 trend artifacts together would have read as some 27 "learning" cells.

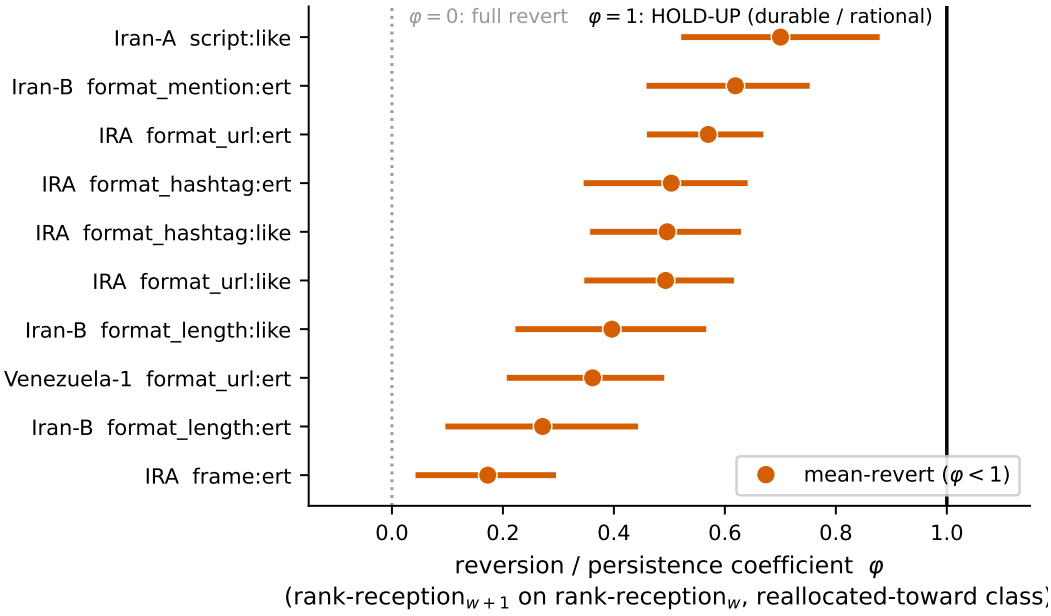

Figure 7. Myth 4, pay-off of reception-chasing (reversion). Reversion/persistence coefficient φ for each feedback-consistent cell (regression of next-window on current-window rank-reception, reallocated-toward class), with a $\geq$2000-iteration block-bootstrap CI; $\varphi = 1$ marks durable hold-up, $\varphi = 0$ full reversion. Every feedback-consistent cell mean-reverts (φ in $[0.17, 0.70]$, all 95% CIs below 1): apparent "learning" reflects transient responses to noise, not durable optimization. Associational; pseudonymous/aggregate cell labels; reception is rank of takedown-final counts (snapshot-timing proxy).

Scoped takeaway.

These operations are *playbook-executed, not feedback-adaptive*. We use "learn" descriptively, never as a claim of cognition or operator intent, and all relationships are associational. Scripted is not the same as harmless, and it sets a baseline: an adversary that genuinely optimized against reception, as large

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

language models now make feasible [53], would look measurably different from this. That contrast is a forward warning, not a present finding.

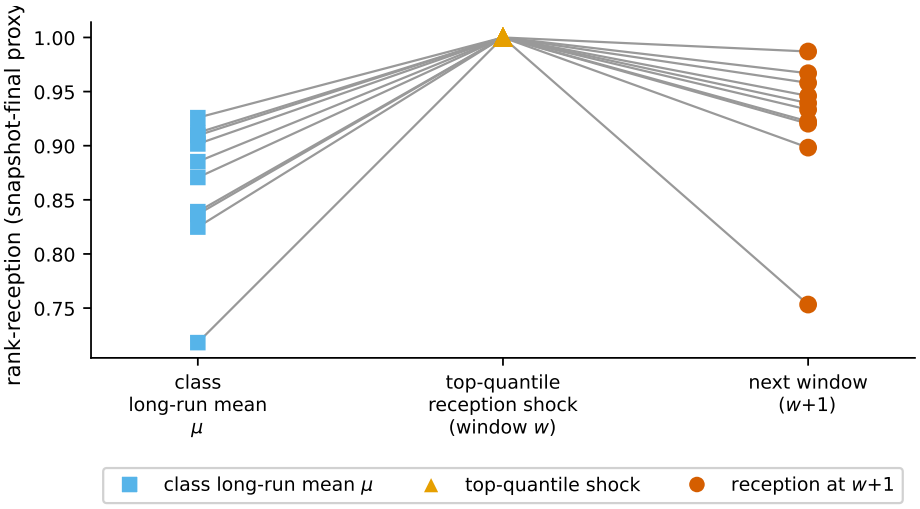

Figure 8. Myth 4, pay-off of reception-chasing (trajectory). A top-quantile reception shock and its partial reversion one window later (median ∼52% of the shock reverts within one window). The reception gain does not persist, consistent with transient responses to noise rather than durable optimization. Associational; pseudonymous/aggregate cell labels; reception is rank of takedown-final counts (snapshot-timing proxy).

8. Myth 5: "Indistinguishable from Real Users" (and Its Mirror, "Still a Crude 2016 Troll")

The claim and its provenance.

Two opposing folk beliefs coexist, both widely circulated in news and policy reporting (Appendix A). One holds that modern IO accounts are individually indistinguishable from real users: the human-operated troll accounts in coordinated networks "present traits more similar to regular users" and lack the synchronization signatures of bots, so that loose coordination, not individual features, is what detection must target [18]; our own clustering work similarly finds trolls "appear indistinguishable" on behavior and intermingle with genuine accounts [14,15]. The other holds that operations remain catchable by 2016-era linguistic fingerprints: behavioral and linguistic signatures separate trolls from users at high accuracy [16], content-based features generalize across campaigns and platforms [41], and dozens of deception-linked language markers achieve strong classification [17]. Recent detectors bear out both halves of this tension: cross-campaign linguistic drift forces continual model adaptation [20,21], while fused coordination signatures remain the durable tell [23,26]. We test both against the matched organic baseline.

The test.

On the five campaigns with a matched control arm, we re-compute the 49-cue deception fingerprint and a battery of coordination-network detectors in the family established for uncovering coordinated activity [10] (same-minute synchrony, copypasta text similarity, co-retweet synchrony, co-hashtag) for IO versus matched real users. Because the baseline draws are budget-limited (ten matched draws give a one-sided p-floor of .0909), a contrast is declared abnormal only when the 95% confidence interval excludes 1 *and* the ratio exceeds 1.

The finding.

The language has evolved off the 2016 fingerprint, but the coordination has not (Figure 9). The deception fingerprint only partially replicates (mean sign-agreement 0.54 across the confirmed five, ranging 34%–66%): the "non-immediate, low-affective stance" markers persist, but the conversational markers the 2016 literature relied on (questions, punctuation, hashtags) *reverse*, with IO using fewer in

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

every campaign. Modern IO in this corpus is more moralized and more negatively emotional than organic users while being *less* conversational, the opposite of the 2016 high-hashtag, high-engagement conversational profile. Yet on coordination the operations are abnormal relative to matched real users: same-minute synchrony runs 7.7–70.3× organic, copypasta similarity 2.2–16.4×, and co-retweet synchrony 15.7–35× in the powered campaigns, while co-hashtag, as pre-registered, realizes the predicted honest null (0.73–1.0×, never above 1) in every campaign, confirming that the matched baseline did not manufacture the contrasts.

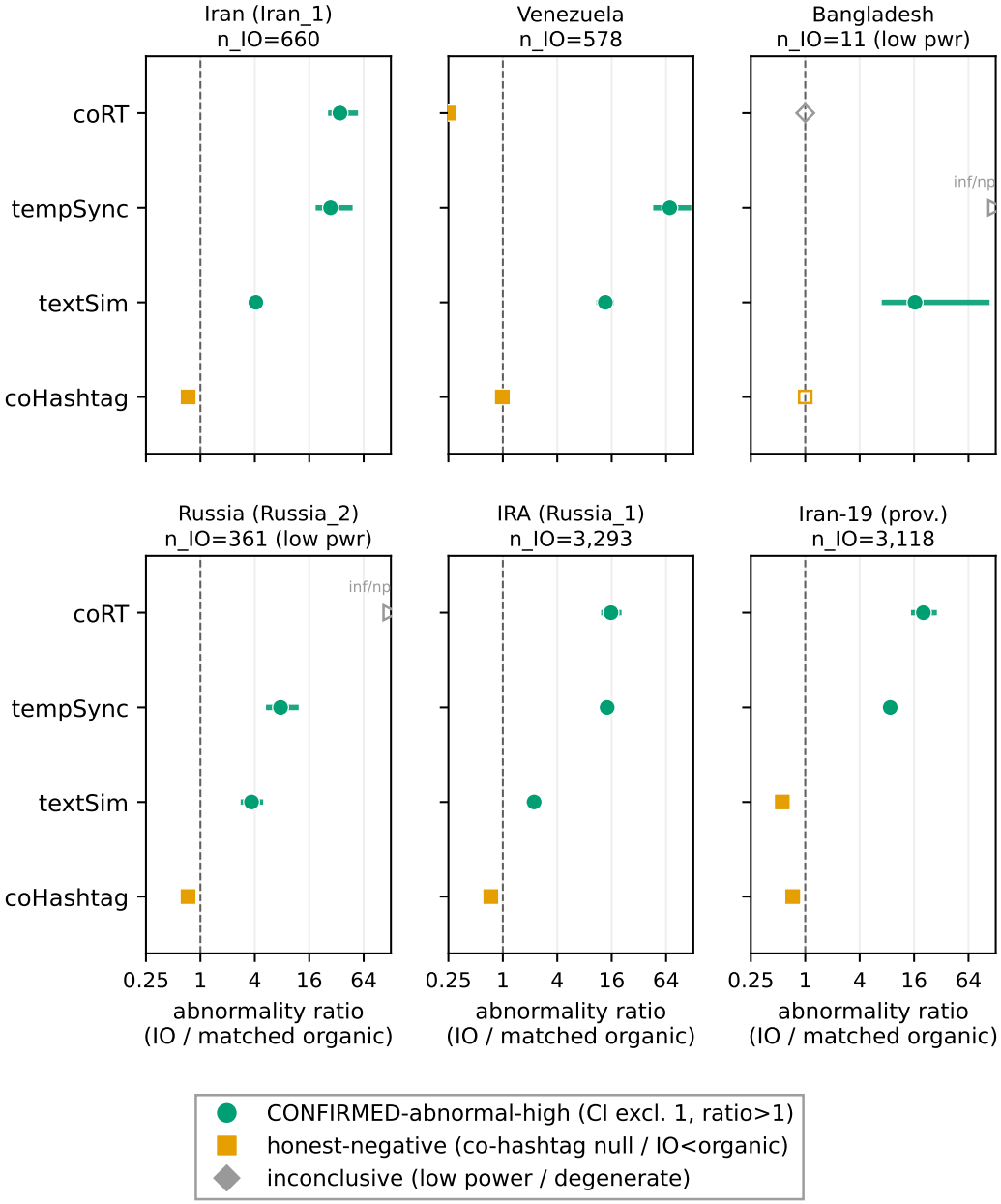

Figure 9. Myth 5 (confirmed five). Coordination abnormality of IO versus matched organic users by detector and campaign, on a $\log_2$ abnormality-ratio axis (IO statistic / matched-size organic draw, 95% bootstrap CI, ratio= 1 reference). Detectors: coRT (co-retweet synchrony), tempSync (same-minute co-activity), textSim (copypasta/near-duplicate), coHashtag (pre-registered honest-negative). Synchrony, copypasta, and co-retweet run many-fold above organic; co-hashtag realizes the pre-registered honest null. Aggregate class-level only. Associational, not causal.

Cross-national replication.

To test whether these are regularities or artifacts of the original four countries, we re-ran the frozen pipeline (no re-tuning) on twelve *new*, non-corpus country-groups against their matched controls (Figure 10, Table 4). All three regularities replicate out-of-sample and the null of indistinguishability is

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

rejected in every country: copypasta similarity is abnormally high in 11 of 12 (pooled random-effects ratio 4.20×, CI [2.56, 6.89]), same-minute synchrony in 11 of 12 (20.7×, [6.2, 69.2]), co-retweet synchrony in 10 of 12 (376× where finite), and the co-hashtag honest-null reproduces in 9 of 12. The language-drift split reproduces (mean English sign-agreement 0.594), and cross-campaign content segregation persists (16× below null, $p = 0.001$), indicating no global template. Four small-arm country-groups (Armenia 31, Qatar 29, Ghana 60, Catalonia 76 IO accounts) lean on degeneracy-driven verdicts, and three co-hashtag matches fail on the smallest arms; we flag these rather than down-weighting them.

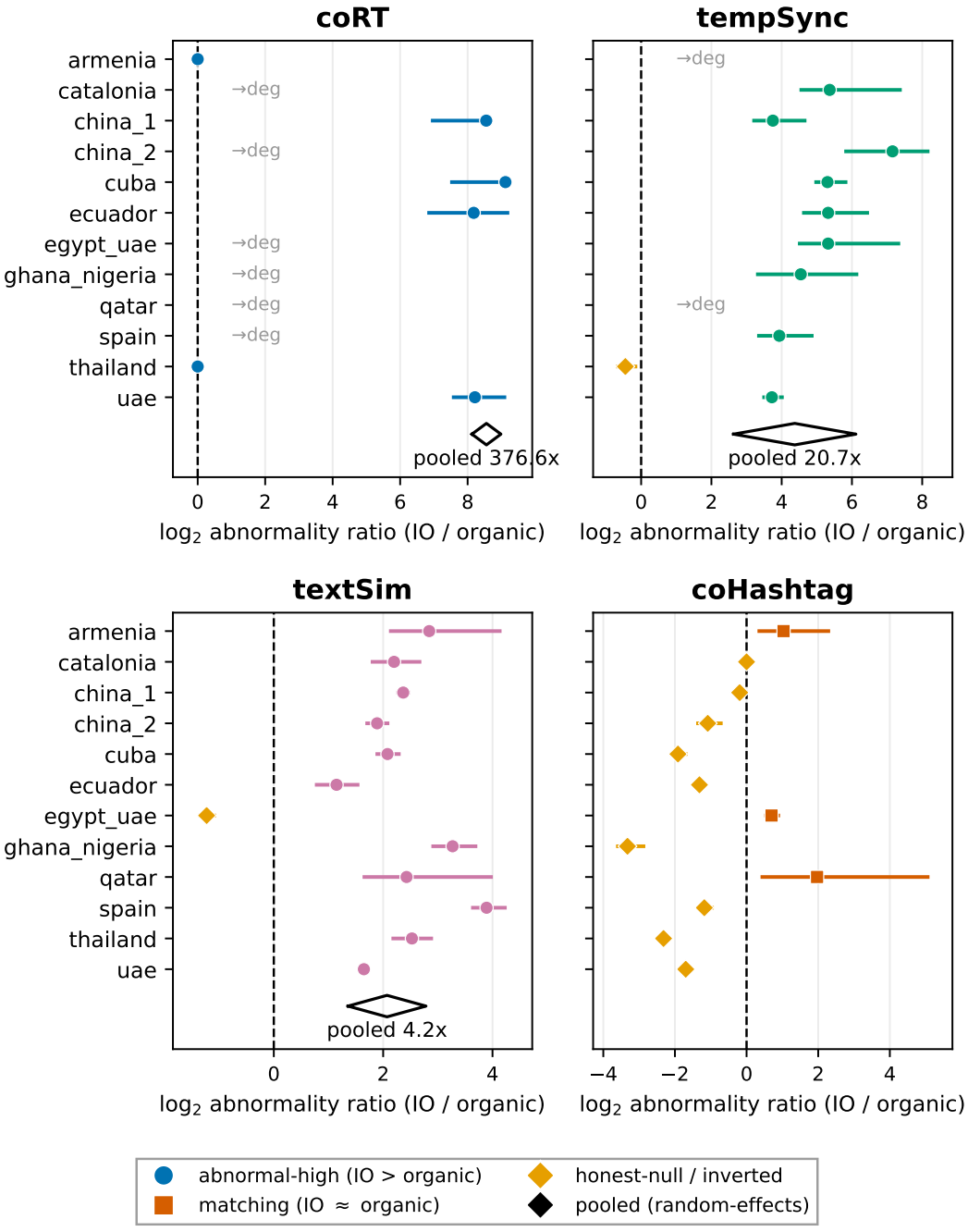

Figure 10. Myth 5 (cross-national replication). Frozen-pipeline re-test on twelve new non-corpus country-groups, by detector ($\log_2$ abnormality ratio, IO versus matched organic; right of the dashed line = IO more coordinated). Black diamonds are random-effects pooled estimates per detector; "→deg" marks degenerate cells (organic statistic 0, IO > 0). The coordination abnormality replicates out-of-sample; the null of indistinguishability from real users is rejected in every country.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

Table 4. Cross-national replication of the coordination abnormality (frozen pipeline, twelve new non-corpus country-groups). Each detector contrasts IO against matched organic users; "abnormal-high" counts country-groups whose ratio exceeds 1 with the confidence interval excluding 1. Pooled ratios are random-effects (DerSimonian–Laird) over the countries with finite estimates. Co-hashtag is the pre-registered *honest null*, which should *not* fire, and largely does not.

Detector	Abnormal-high	Pooled Ratio (95% CI)	Reading
Copypasta text similarity	11 / 12	4.20× [2.56, 6.89]	replicates
Same-minute synchrony	11 / 12	20.7× [6.2, 69.2]	replicates
Co-retweet synchrony	10 / 12	376× (finite $k = 4$)	replicates
Co-hashtag (honest null)	9 / 12 honest-null	0.10–0.88×	null reproduces

Scoped takeaway.

Neither folk belief holds: these operations no longer talk like 2016 trolls, but they still coordinate like machines. The detectable signature has *migrated from language to coordination*, and that migration is cross-national.

9. Discussion

What *is* true: the corrected picture.

The five corrected findings are not five disconnected negatives; they compose a single, coherent account. These operations are *industrial content factories*: compartmentalized into nationally fingerprinted desks (Myth 1), thinly staffed by few operators executing calendars rather than crowds of improvisers (Myth 1), producing content on a script rather than optimizing it against reception (Myth 4). What they cannot do internally is make that content travel: the moral-emotional lever that the folk model treats as their primary mechanism does not move reception in their own data (Myth 2), and their viral reach is predominantly *captured* from an external audience they do not control rather than manufactured by their own retweets (Myth 3). What durably distinguishes them from real users is not the content of any single account, whose language has converged toward the ordinary (Myth 5), but the machine-like coordination across accounts that persists regardless of individual-account linguistic convergence (Myth 5). "Siloed in production, coordinated in execution" is not a contradiction; it is the corrected mechanism.

Why the reach limit and the coordination evidence fit together.

The hardest limit in this paper, that we cannot decompose external reach into organic uptake and undetected coordination (Myth 3), is partly backstopped by the coordination evidence (Myth 5). The two myths use the word "synchrony" for two different objects, and separating them is what makes the reconciliation work. Under Myth 5, the discriminating signal is *cross-account* same-minute synchrony: many accounts acting in lockstep, which is what separates the operations from matched organic users at 7–70×. Under Myth 3, the question is whether *within-author* seeding synchrony, one account's own co-timed posting of a viral original, predicts how far that original then travels externally; it does not, and in the IRA and Iran it weakly anti-predicts reach. These are consistent: cross-account coordination is a reliable *indicator* of inauthenticity, but the same synchrony does not *drive* external reach. Thus, where reach *is* observable, on the operations' own accounts, the synchrony that marks the operation as coordinated is precisely *not* the mechanism that propagates its content externally; the diffuse external sharing that does carry the bulk of reach shows no such lockstep signature in the operations where we can look. The unresolved part of the reach question, whether the external pool is organic or partly undetected coordination, is therefore bounded by what coordination looks like where we *can* measure it: if undetected coordination were carrying the external reach, it would have to do so without the same-minute, copypasta, and co-retweet signatures that coordination otherwise leaves in every measurable context. This is an inference from the internal accounts to an unobservable external pool, not a measurement of that pool; it narrows the question rather than closing it.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

The methodological lesson.

The common thread across all five myths is that a corpus of platform-confirmed manipulation will, analyzed without a baseline, a placebo, or false-discovery control, reproduce the patterns analysts expect to find. A meaningless letter-count predicts engagement on these samples (Myth 2); the corpus's largest "adaptation" coefficients are time-symmetric trend (Myth 4); within-corpus significance on the manipulation arm says nothing about how these accounts compare to real users until an organic arm is added (Myths 2 and 5). Pre-registration, matched baselines, placebo features, and permutation nulls are not procedural formalities here; they are what separates a genuine regularity from a spurious artifact. Several of our positive controls (detectors that *do* fire on coordination, a co-hashtag test that reproduces as a clean null, a fingerprint classifier that separates campaigns at high accuracy) demonstrate that the null findings are not attributable merely to insufficient statistical power.

10. Limitations

These results concern seven specific, already-investigated campaigns and twelve further country-groups; they do not speak to operations that platforms never detected, and no claim here generalizes to "all influence operations." The engagement counters are takedown-final snapshots, a small acknowledged limitation addressed by the snapshot decomposition. The genuinely bounding limitation is the external-reach decomposition of Myth 3: the archives record only the operations' own accounts, so the external audience is structurally invisible and the organic-versus-undetected-coordination split of the captured reach is not a computable quantity in this corpus, a limit partly, not wholly, addressed by the coordination evidence of Myth 5. The matched-baseline contrasts rest on budget-limited draws (a p-floor of .0909), so confidence-interval verdicts, not false-discovery-corrected q-values, carry several of the Myth 5 conclusions; small-arm cross-national countries lean on degeneracy-driven verdicts and are flagged individually. The organic baseline lacks engagement counts, which makes an organic moral-contagion replication impossible and leaves Myth 2's debunk anchored to the within-corpus placebo rather than a like-for-like organic comparison. Stylometric "hands" estimates are ranges, not headcounts, and the classic and neural representations agree only in aggregate. The Linvill–Warren convergent-validity overlay (Myth 1) is confirmatory and IRA-only: it joins by screen name (incomplete coverage, concentrated in the heavily churned RightTroll category), no comparable external hand labeling exists for the other operations, and the NewsFeed category is the one label that does not fall in the two giant desks. All relationships reported are associational. Control-subject content is reported only in aggregate, consistent with its license and with the controls being ordinary users.

11. Conclusions

Scoped to these seven state campaigns, five widely circulated beliefs about how influence operations work either weaken or reverse under baseline-anchored, pre-registered testing. The operations are not monolithic armies but compartmentalized, thinly staffed production desks; they do not win on moral-emotional language, which fails to predict reception in their own data; they do not manufacture their own virality, which is predominantly captured rather than internally generated; they do not optimize against feedback, behaving instead on scripts whose rare apparent adaptation is statistically indistinguishable from noise; and they are neither indistinguishable from real users nor still caught by 2016-era language, having drifted linguistically toward the ordinary while continuing to coordinate 7–70$\times$ more tightly than matched organic users. The corrected picture, an industrial content factory whose reach is predominantly captured from an external audience it does not control rather than internally manufactured (with the organic-versus-undetected-coordination composition of that external reach left unresolved), accounts for the evidence more consistently than the folk model it replaces. The recurring lesson for the field is methodological: on corpora of confirmed manipulation, significance without baselines and placebos reconstructs the analyst's expectations, and only pre-registered, baseline-anchored, placebo-gated testing tells us which "stylized facts" are facts.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

Funding: This research received no external funding.

Data Availability Statement: The primary corpus derives from the public Twitter Information Operations archive; the organic baseline derives from an independent release [19] and is reported in aggregate only, without redistribution. Pre-registration files, analysis code, and derived aggregate tables are maintained in a project repository and will be made available on publication; because the organic baseline's license forbids redistribution, the released artifacts contain derived aggregate statistics only, never control-account records or text.

Conflicts of Interest: The author declares no conflict of interest.

Appendix A. Provenance of the Five Myths in Public and Policy Discourse

Each myth is attributed in the body to peer-reviewed or authoritative scholarship. To document that these beliefs are not academic strawmen but circulate as received wisdom, Table A1 catalogues representative statements of each myth in authoritative news reporting and in government, intelligence, and policy-institution documents. This catalogue is illustrative, not exhaustive, and is kept deliberately separate from the scholarly References: its purpose is to evidence the *popular and policy framing* the paper tests, not to endorse these sources. All entries are verbatim excerpts (abridged with ellipses where noted); each was retrieved from the cited outlet, and the full catalogue with URLs and retrieval notes accompanies the project's supplementary materials. Quotation here documents how a claim was publicly framed and implies no judgment of the cited authors.

Table A1. Representative framings of each myth in authoritative news and policy discourse. "Type" is News, Policy/Govt (government, intelligence, or legal documents), or Think-tank / NGO (including reports commissioned by the U.S. Senate Select Committee on Intelligence). Quotes are verbatim, abridged with "…" where indicated.

Myth	Source (outlet, year)	Type	Representative framing (verbatim)	What it portrays / distorts
M1	Associated Press, 2015	News	"Russia's army of trolls lead Putin's online propaganda campaign" (headline)	A single unified "army" of trolls acting in concert.
M1	IRA Indictment (DOJ/Special Counsel), 2018	Policy/Govt	"…conduct what it called 'information warfare against the United States of America' …"	One organization waging unified information warfare.
M1	Reuters, 2018 (Iran)	News	"a sprawling network of anonymous websites and social media accounts in 11 different languages"	The Iran operation as one vast, sprawling, unified network.
M1	The Conversation, 2024	News	"…hired hundreds of online commentators (trolls) …responsible for promoting and amplifying the narratives that suited the Kremlin's agenda"	"Troll factory" of hundreds serving one centralized agenda.
M1	New Knowledge (SSCI report), 2018	Think-tank / NGO	"a sweeping and sustained social influence operation …aimed directly at US citizens"	The flagship Senate-commissioned report frames it as one unified operation.
M2	U.S. State Dept. GEC, *Weapons of Mass Distraction*, 2019	Policy/Govt	"…fear, and anger—are the very characteristics that increase the likelihood a message will go viral."	A government report asserting emotion is what drives spread.
M2	IRA Indictment (DOJ/Special Counsel), 2018	Policy/Govt	"…a strategic goal to sow discord in the U.S. political system …"	Anchors the divisive-emotional framing of the operation.
M2	Oxford Internet Institute / Graphika (SSCI report), 2018	Think-tank / NGO	"spreading sensationalist, conspiratorial, and other forms of junk political news and misinformation"	Attributes reach to sensational, inflammatory content.
M2	PBS NewsHour, 2018	News	"elicit outrage with posts about liberal appeasement of 'others' at the expense of US citizens"	The core tactic framed as eliciting outrage.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

Table A1. *Cont.*

Myth	Source (outlet, year)	Type	Representative framing (verbatim)	What it portrays / distorts
M2	Univ. of Colorado Boulder (Vargo study), 2020	News	"…fear and anger appeals work really well in getting people to engage."	Attributes engagement success to fear/anger appeals.
M3	Mueller Report, via Brookings, 2019	Policy/Govt	"…individual accounts that would post content, which bot networks would then amplify."	A two-tier internal machine where bots amplify own content.
M3	Symantec (threat intelligence), 2019	Think-tank / NGO	"Auxiliary accounts were configured to retweet content pushed out by the main accounts."	A designed self-amplification architecture.
M3	Atlantic Council DFRLab, 2020 (Venezuela)	Think-tank / NGO	"…artificially amplified …to make the hashtags seem more popular and organic than they were."	Manufactured appearance of organic popularity.
M3	CBS News, 2018	News	"Russian bots retweeted Mr. Trump almost 470,000 times …"	Bot-driven retweeting cast as the amplification engine.
M3	Iran International, 2022 (Iran)	News	"…made software for the tweets and retweets and had created at least 256 accounts …"	An automated self-retweet network manufacturing reach.
M4	IRA Indictment (DOJ/Special Counsel), 2018	Policy/Govt	"…tracked the performance of content they posted over social media."	Basis for the data-driven, metrics-optimizing framing.
M4	RAND (Paul & Matthews, *Firehose of Falsehood*), 2016	Think-tank / NGO	"It is rapid, continuous, and repetitive."	Casts the operation as a fast, agile, responsive machine.
M4	Oxford Internet Institute / Graphika (SSCI report), 2018	Think-tank / NGO	"adapted existing techniques from digital advertising to spread disinformation …"	Likens the IRA to a professional digital-marketing operation.
M4	CBS News (quoting DiResta), 2018	News	"a sophisticated operation that does not hesitate to reach out to individual Americans"	The "sophisticated operation" framing.
M5	IRA Indictment (DOJ/Special Counsel), 2018	Policy/Govt	"…posing as U.S. persons and creating false U.S. personas …"	(a) IO accounts blend in as ordinary Americans.
M5	DISA, 2025	Think-tank / NGO	"…often indistinguishable from genuine users …"	(a) States flatly that accounts are indistinguishable.
M5	The Intercept, 2024	News	AI personas "undetectable by social media algorithms"	(a) Undetectable, human-seeming personas.
M5	DOJ press release (RT "Meliorator" bot farm), 2024	Policy/Govt	"…create fake but authentic-appearing social media personas en masse."	(a) A 2024 U.S. government case: AI personas authentic at scale.
M5	Atlantic Council DFRLab (Nimmo), 2018	Think-tank / NGO	"…inability to use the grammatical articles—'a' and 'the'—appropriately."	(b) The crude linguistic-tell stereotype.
M5	Propastop (Estonian Defence League), 2024	Think-tank / NGO	"…they slip and start writing in native Russian."	(b) Trolls betrayed by crude operational slip-ups.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

References

1. Linvill, D.L.; Warren, P.L. Troll Factories: Manufacturing Specialized Disinformation on Twitter. *Polit. Commun.* **2020**, *37*, 447–467. https://doi.org/10.1080/10584609.2020.1718257.
2. Suk, J.; Lukito, J.; Su, M.-H.; Kim, S.J.; Tong, C.; Sun, Z.; Sarma, P. Do I Sound American? How Message Attributes of Internet Research Agency (IRA) Disinformation Relate to Twitter Engagement. *Comput. Commun. Res.* **2022**, *4*, 590–628. https://doi.org/10.5117/CCR2022.2.008.SUK.
3. Brady, W.J.; Wills, J.A.; Jost, J.T.; Tucker, J.A.; Van Bavel, J.J. Emotion Shapes the Diffusion of Moralized Content in Social Networks. *Proc. Natl. Acad. Sci. USA* **2017**, *114*, 7313–7318. https://doi.org/10.1073/pnas.1618923114.
4. Burton, J.W.; Cruz, N.; Hahn, U. Reconsidering Evidence of Moral Contagion in Online Social Networks. *Nat. Hum. Behav.* **2021**, *5*, 1629–1635. https://doi.org/10.1038/s41562-021-01133-5.
5. DiResta, R.; Shaffer, K.; Ruppel, B.; Sullivan, D.; Matney, R.; Fox, R.; Albright, J.; Johnson, B. *The Tactics & Tropes of the Internet Research Agency*; New Knowledge, report prepared for the U.S. Senate Select Committee on Intelligence: Austin, TX, USA, 2018.
6. Paul, C.; Matthews, M. *The Russian "Firehose of Falsehood" Propaganda Model: Why It Might Work and Options to Counter It*; RAND Corporation, Perspective PE-198: Santa Monica, CA, USA, 2016. https://doi.org/10.7249/PE198.
7. Shao, C.; Ciampaglia, G.L.; Varol, O.; Yang, K.-C.; Flammini, A.; Menczer, F. The Spread of Low-Credibility Content by Social Bots. *Nat. Commun.* **2018**, *9*, 4787. https://doi.org/10.1038/s41467-018-06930-7.
8. Cinelli, M.; Cresci, S.; Quattrociocchi, W.; Tesconi, M.; Zola, P. Coordinated Inauthentic Behavior and Information Spreading on Twitter. *Decis. Support Syst.* **2022**, *160*, 113819. https://doi.org/10.1016/j.dss.2022.113819.
9. Golovchenko, Y.; Hartmann, M.; Adler-Nissen, R. State, Media and Civil Society in the Information Warfare over Ukraine: Citizen Curators of Digital Disinformation. *Int. Aff.* **2018**, *94*, 975–994. https://doi.org/10.1093/ia/iiy148.
10. Pacheco, D.; Hui, P.-M.; Torres-Lugo, C.; Truong, B.T.; Flammini, A.; Menczer, F. Uncovering Coordinated Networks on Social Media: Methods and Case Studies. *Proc. Int. AAAI Conf. Web Soc. Media* **2021**, *15*, 455–466. https://doi.org/10.1609/icwsm.v15i1.18075.
11. Pozzana, I.; Ferrara, E. Measuring Bot and Human Behavioral Dynamics. *Front. Phys.* **2020**, *8*, 125. https://doi.org/10.3389/fphy.2020.00125.
12. Zannettou, S.; Caulfield, T.; De Cristofaro, E.; Sirivianos, M.; Stringhini, G.; Blackburn, J. Disinformation Warfare: Understanding State-Sponsored Trolls on Twitter and Their Influence on the Web. In *Companion Proceedings of The 2019 World Wide Web Conference*; San Francisco, CA, USA, 2019; pp. 218–226. https://doi.org/10.1145/3308560.3316495.
13. Zannettou, S.; Caulfield, T.; Setzer, W.; Sirivianos, M.; Stringhini, G.; Blackburn, J. Who Let the Trolls Out? Towards Understanding State-Sponsored Trolls. In *Proceedings of the 10th ACM Conference on Web Science (WebSci '19)*; Boston, MA, USA, 2019; pp. 353–362. https://doi.org/10.1145/3292522.3326016.
14. Luceri, L.; Giordano, S.; Ferrara, E. Detecting Troll Behavior via Inverse Reinforcement Learning: A Case Study of Russian Trolls in the 2016 US Election. *Proc. Int. AAAI Conf. Web Soc. Media* **2020**, *14*, 417–427. https://doi.org/10.1609/icwsm.v14i1.7311.
15. Ezzeddine, F.; Ayoub, O.; Giordano, S.; Nogara, G.; Sbeity, I.; Ferrara, E.; Luceri, L. Exposing Influence Campaigns in the Age of LLMs: A Behavioral-Based AI Approach to Detecting State-Sponsored Trolls. *EPJ Data Sci.* **2023**, *12*, 46. https://doi.org/10.1140/epjds/s13688-023-00423-4.
16. Im, J.; Chandrasekharan, E.; Sargent, J.; Lighthammer, P.; Denby, T.; Bhargava, A.; Hemphill, L.; Jurgens, D.; Gilbert, E. Still Out There: Modeling and Identifying Russian Troll Accounts on Twitter. In *Proceedings of the 12th ACM Conference on Web Science (WebSci '20)*; Southampton, UK, 2020; pp. 1–10. https://doi.org/10.1145/3394231.3397889.
17. Addawood, A.; Badawy, A.; Lerman, K.; Ferrara, E. Linguistic Cues to Deception: Identifying Political Trolls on Social Media. *Proc. Int. AAAI Conf. Web Soc. Media* **2019**, *13*, 15–25. https://doi.org/10.1609/icwsm.v13i01.3205.
18. Saeed, M.H.; Ali, S.; Blackburn, J.; De Cristofaro, E.; Zannettou, S.; Stringhini, G. TrollMagnifier: Detecting State-Sponsored Troll Accounts on Reddit. In *2022 IEEE Symposium on Security and Privacy (SP)*; San Francisco, CA, USA, 2022; pp. 2161–2175. https://doi.org/10.1109/SP46214.2022.9833706.
19. Seckin, O.C.; Pote, M.; Nwala, A.; Yin, L.; Luceri, L.; Flammini, A.; Menczer, F. Labeled Datasets for Research on Information Operations. *Proc. Int. AAAI Conf. Web Soc. Media* **2025**, *19*, 2567–2574. https://doi.org/10.1609/icwsm.v19i1.35958. Data: Zenodo, https://doi.org/10.5281/zenodo.14141549.
20. Tian, L.; Zhang, X.; Lau, J.H. MetaTroll: Few-Shot Detection of State-Sponsored Trolls with Transformer Adapters. In *Proceedings of the ACM Web Conference 2023 (WWW '23)*; Austin, TX, USA, 2023; pp. 1743–1753. https://doi.org/10.1145/3543507.3583417.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

21. Tian, L.; Zhang, X.; Kim, M.M.-H.; Biggs, J.; Rizoiu, M.-A. X-Troll: eXplainable Detection of State-Sponsored Information Operations Agents. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM '25)*; Seoul, Republic of Korea, 2025; pp. 2874–2884. https://doi.org/10.1145/3746252.3761028.
22. Saeed, M.H.; Ali, S.; Paudel, P.; Blackburn, J.; Stringhini, G. Unraveling the Web of Disinformation: Exploring the Larger Context of State-Sponsored Influence Campaigns on Twitter. In *Proceedings of the 27th International Symposium on Research in Attacks, Intrusions and Defenses (RAID '24)*; Padua, Italy, 2024; pp. 353–367. https://doi.org/10.1145/3678890.3678911.
23. Luceri, L.; Pantè, V.; Burghardt, K.; Ferrara, E. Unmasking the Web of Deceit: Uncovering Coordinated Activity to Expose Information Operations on Twitter. In *Proceedings of the ACM Web Conference 2024 (WWW '24)*; Singapore, 2024; pp. 2530–2541. https://doi.org/10.1145/3589334.3645529.
24. Cinus, F.; Minici, M.; Luceri, L.; Ferrara, E. Exposing Cross-Platform Coordinated Inauthentic Activity in the Run-Up to the 2024 U.S. Election. In *Proceedings of the ACM Web Conference 2025 (WWW '25)*; Sydney, Australia, 2025; pp. 541–559. https://doi.org/10.1145/3696410.3714698.
25. Tardelli, S.; Nizzoli, L.; Tesconi, M.; Conti, M.; Nakov, P.; Da San Martino, G.; Cresci, S. Temporal Dynamics of Coordinated Online Behavior: Stability, Archetypes, and Influence. *Proc. Natl. Acad. Sci. USA* **2024**, *121*, e2307038121. https://doi.org/10.1073/pnas.2307038121.
26. Tardelli, S.; Nizzoli, L.; Avvenuti, M.; Cresci, S.; Tesconi, M. Multifaceted Online Coordinated Behavior in the 2020 US Presidential Election. *EPJ Data Sci.* **2024**, *13*, 33. https://doi.org/10.1140/epjds/s13688-024-00467-0.
27. Brady, W.J.; Rathje, S.; Globig, L.K.; Van Bavel, J.J. Estimating the Effect Size of Moral Contagion in Online Networks: A Pre-Registered Replication and Meta-Analysis. *PNAS Nexus* **2025**, *4*, pgaf327. https://doi.org/10.1093/pnasnexus/pgaf327.
28. Rathje, S.; Van Bavel, J.J.; van der Linden, S. Out-Group Animosity Drives Engagement on Social Media. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2024292118. https://doi.org/10.1073/pnas.2024292118.
29. Robertson, C.E.; Pröllochs, N.; Schwarz, K.; Pärnamets, P.; Van Bavel, J.J.; Feuerriegel, S. Negativity Drives Online News Consumption. *Nat. Hum. Behav.* **2023**, *7*, 812–822. https://doi.org/10.1038/s41562-023-01538-4.
30. Vosoughi, S.; Roy, D.; Aral, S. The Spread of True and False News Online. *Science* **2018**, *359*, 1146–1151. https://doi.org/10.1126/science.aap9559.
31. Eady, G.; Paskhalis, T.; Zilinsky, J.; Bonneau, R.; Nagler, J.; Tucker, J.A. Exposure to the Russian Internet Research Agency Foreign Influence Campaign on Twitter in the 2016 US Election and Its Relationship to Attitudes and Voting Behavior. *Nat. Commun.* **2023**, *14*, 62. https://doi.org/10.1038/s41467-022-35576-9.
32. Bail, C.A.; Guay, B.; Maloney, E.; Combs, A.; Hillygus, D.S.; Merhout, F.; Freelon, D.; Volfovsky, A. Assessing the Russian Internet Research Agency's Impact on the Political Attitudes and Behaviors of American Twitter Users in Late 2017. *Proc. Natl. Acad. Sci. USA* **2020**, *117*, 243–250. https://doi.org/10.1073/pnas.1906420116.
33. McLoughlin, K.L.; Brady, W.J.; Goolsbee, A.; Kaiser, B.; Klonick, K.; Crockett, M.J. Misinformation Exploits Outrage to Spread Online. *Science* **2024**, *386*, 991–996. https://doi.org/10.1126/science.adl2829.
34. Goldstein, J.A.; Chao, J.; Grossman, S.; Stamos, A.; Tomz, M. How Persuasive Is AI-Generated Propaganda? *PNAS Nexus* **2024**, *3*, pgae034. https://doi.org/10.1093/pnasnexus/pgae034.
35. Hackenburg, K.; Margetts, H. Evaluating the Persuasive Influence of Political Microtargeting with Large Language Models. *Proc. Natl. Acad. Sci. USA* **2024**, *121*, e2403116121. https://doi.org/10.1073/pnas.2403116121.
36. Orben, A.; Przybylski, A.K. The Association between Adolescent Well-Being and Digital Technology Use. *Nat. Hum. Behav.* **2019**, *3*, 173–182. https://doi.org/10.1038/s41562-018-0506-1.
37. Wang, X.; Li, J.; Srivatsavaya, E.; Rajtmajer, S. Evidence of Inter-State Coordination amongst State-Backed Information Operations. *Sci. Rep.* **2023**, *13*, 7716. https://doi.org/10.1038/s41598-023-34245-1.
38. Pantè, V.; Axelrod, D.; Flammini, A.; Menczer, F.; Ferrara, E.; Luceri, L. Beyond Interaction Patterns: Assessing Claims of Coordinated Inter-State Information Operations on Twitter/X. In *Companion Proceedings of the ACM Web Conference 2025 (WWW '25 Companion)*; Sydney, Australia, 2025; pp. 1234–1238. https://doi.org/10.1145/3701716.3715575.
39. Uyheng, J.; Cruickshank, I.J.; Carley, K.M. Mapping State-Sponsored Information Operations with Multi-View Modularity Clustering. *EPJ Data Sci.* **2022**, *11*, 25. https://doi.org/10.1140/epjds/s13688-022-00338-6.
40. Kumar, S.; Cheng, J.; Leskovec, J.; Subrahmanian, V.S. An Army of Me: Sockpuppets in Online Discussion Communities. In *Proceedings of the 26th International Conference on World Wide Web (WWW '17)*; Perth, Australia, 2017; pp. 857–866. https://doi.org/10.1145/3038912.3052677.
41. Alizadeh, M.; Shapiro, J.N.; Buntain, C.; Tucker, J.A. Content-Based Features Predict Social Media Influence Operations. *Sci. Adv.* **2020**, *6*, eabb5824. https://doi.org/10.1126/sciadv.abb5824.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 June 2026 doi:10.20944/preprints202606.2093.v1

42. Gabriel, N.A.; Broniatowski, D.A.; Johnson, N.F. Inductive Detection of Influence Operations via Graph Learning. *Sci. Rep.* **2023**, *13*, 22571. https://doi.org/10.1038/s41598-023-49676-z.
43. Brady, W.J.; McLoughlin, K.; Doan, T.N.; Crockett, M.J. How Social Learning Amplifies Moral Outrage Expression in Online Social Networks. *Sci. Adv.* **2021**, *7*, eabe5641. https://doi.org/10.1126/sciadv.abe5641.
44. Kyrychenko, Y.; Brik, T.; van der Linden, S.; Roozenbeek, J. Social Identity Correlates of Social Media Engagement before and after the 2022 Russian Invasion of Ukraine. *Nat. Commun.* **2024**, *15*, 8127. https://doi.org/10.1038/s41467-024-52179-8.
45. Baribi-Bartov, S.; Swire-Thompson, B.; Grinberg, N. Supersharers of Fake News on Twitter. *Science* **2024**, *384*, 979–982. https://doi.org/10.1126/science.adl4435.
46. González-Bailón, S.; De Domenico, M. Bots Are Less Central than Verified Accounts during Contentious Political Events. *Proc. Natl. Acad. Sci. USA* **2021**, *118*, e2013443118. https://doi.org/10.1073/pnas.2013443118.
47. Salvi, F.; Horta Ribeiro, M.; Gallotti, R.; West, R. On the Conversational Persuasiveness of GPT-4. *Nat. Hum. Behav.* **2025**, *9*, 1645–1653. https://doi.org/10.1038/s41562-025-02194-6.
48. Spitale, G.; Biller-Andorno, N.; Germani, F. AI Model GPT-3 (Dis)informs Us Better than Humans. *Sci. Adv.* **2023**, *9*, eadh1850. https://doi.org/10.1126/sciadv.adh1850.
49. Gelman, A.; Loken, E. The Statistical Crisis in Science. *Am. Sci.* **2014**, *102*, 460. https://doi.org/10.1511/2014.111.460.
50. Simonsohn, U.; Simmons, J.P.; Nelson, L.D. Specification Curve Analysis. *Nat. Hum. Behav.* **2020**, *4*, 1208–1214. https://doi.org/10.1038/s41562-020-0912-z.
51. Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **1995**, *57*, 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
52. Minici, M.; Cinus, F.; Luceri, L.; Ferrara, E. Uncovering Coordinated Cross-Platform Information Operations Threatening the Integrity of the 2024 U.S. Presidential Election Online Discussion. *First Monday* **2024**, *29*. https://doi.org/10.5210/fm.v29i11.13831.
53. Orlando, G.M.; Ye, J.; La Gatta, V.; Saeedi, M.; Moscato, V.; Ferrara, E.; Luceri, L. Emergent Coordinated Behaviors in Networked LLM Agents: Modeling the Strategic Dynamics of Information Operations. In *Proceedings of the ACM Web Conference 2026 (WWW '26)*; 2026; pp. 4805–4816. https://doi.org/10.1145/3774904.3792580.
54. Bessi, A.; Ferrara, E. Social Bots Distort the 2016 U.S. Presidential Election Online Discussion. *First Monday* **2016**, *21*. https://doi.org/10.5210/fm.v21i11.7090.
55. Ferrara, E.; Varol, O.; Davis, C.; Menczer, F.; Flammini, A. The Rise of Social Bots. *Commun. ACM* **2016**, *59*, 96–104. https://doi.org/10.1145/2818717.
56. Goldstein, J.A.; Sastry, G.; Musser, M.; DiResta, R.; Gentzel, M.; Sedova, K. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. *arXiv* **2023**, arXiv:2301.04246. https://doi.org/10.48550/arXiv.2301.04246.
57. Ferrara, E. GenAI against Humanity: Nefarious Applications of Generative Artificial Intelligence and Large Language Models. *J. Comput. Soc. Sci.* **2024**, *7*, 549–569. https://doi.org/10.1007/s42001-024-00250-1.
58. Twitter, Inc. Information Operations. Twitter Transparency Center, 2018–2023. Available online: https://web.archive.org/web/20200822015858/https://transparency.twitter.com/en/reports/information-operations.html (accessed on 22 August 2020).
59. Maslov, S.; Sneppen, K. Specificity and Stability in Topology of Protein Networks. *Science* **2002**, *296*, 910–913. https://doi.org/10.1126/science.1065103.
60. Rivera-Soto, R.A.; Miano, O.E.; Ordonez, J.; Chen, B.Y.; Khan, A.; Bishop, M.; Andrews, N. Learning Universal Authorship Representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; 2021; pp. 913–919. https://doi.org/10.18653/v1/2021.emnlp-main.70.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.