Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

From form to function: an ontology of communicative influence beyond truth and intent

Timothy Paul Westlake

t.p.westlake@tilburguniversity.edu

Tilburg University

Lee Rowland

London School of Economics and Political Science

Article

Keywords: communicative influence, influence operations, ontology, computational communication, persuasion, framing, narrative, polarization, misinformation

Posted Date: November 9th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11295959/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

From form to function: an ontology of communicative influence beyond truth and intent

Anonymised for peer review

Abstract

The widespread use of AI-generated slopaganda and influence campaigns creates a need for computational analysis at scale. Research on disinformation has largely asked whether messages are true or intended to deceive. Yet much communicative influence operates without falsity, through framing, narrative, group positioning, and orientation toward civic or commercial outcomes. Existing frameworks address complementary aspects of such influence but do not represent them as distinct dimensions of the same message; this paper presents an ontology that does. The ontology is truth- and intent-agnostic and operates at the level of message form, resting on the premise that strategic communication is goal-directed and tends to leave observable traces in the message itself. It distinguishes four conceptually orthogonal dimensions: Method (how influence operates), Narrative (how meaning is organised), Polarization (how groups are positioned), and AboutObject (what broader outcome the message bears toward), while TopicDomain independently indexes what the message is about. A single inferential primitive, the cue, grounds base-level labels in message evidence, from which higher-level labels are derived through fixed mappings. Discriminant validity is assessed through irreducibility and content validity through systematic framework mapping; formal verification checks logical coherence and structural and derivation constraints. Empirical annotation reliability and computational annotation performance are not assessed. The resulting structured representation supports both close analysis of individual messages and comparison across large corpora. The codebook, formal verification artefacts, and worked annotation examples are released as open supplemental material.

Keywords: communicative influence; influence operations; ontology; computational communication; persuasion; framing; narrative; polarization; misinformation

1 Introduction

In 2016, UKIP's campaign *Breaking Point* (UKIP, 2016) shared a public X post featuring a documentary photograph of refugees at the Croatian-

Slovenian border (see Figure 1) as a motivation for Brexit. The image contained no false statement, explicit or implicit. Its communicative force came from selection, framing, and political anchoring: a real scene positioned to signal threat, position groups, and indicate a preferred response.



Figure 1: *Breaking Point* campaign image posted publicly by UKIP on X, 16 June 2016 (UKIP, 2016). Reproduced for scholarly criticism and review.

A decade later, generative AI has increased the scale and speed of targeted online messaging used for communicative influence. With volumes now well beyond what any analyst can read, computational approaches are needed to distinguish between different ways messages aim to manipulate cognition and behaviour. AI-enabled disinformation systems can already automate persona generation, persuasive content production, coordinated deployment, and iterative tactical optimisation at scale (Bergmanis-Korāts and Chia Tee Hiang, 2026). The term 'slopaganda' captures the rise of persuasive generative AI-enabled political rhetoric, propaganda and misinformation (Klincewicz et al., 2025). Like the Big Lie broadcast by Nazi-ideologues, it is effective because it can be intermixed with truth or even amplify true information in a way that captures attention of the target audience (Klincewicz et al., 2025).

Here we present a message-internal, truth- and intent-agnostic ontology of communicative influence designed for computational implementation in large-scale message analysis. It separates four conceptually orthogonal dimensions: Method, Narrative, Polarization, and AboutObject, while Top-

icDomain independently indexes subject matter. Each dimension answers a question the others cannot: identical tactics can position groups differently, and messages on the same topic can bear toward different outcomes. A single inferential primitive, the cue, grounds base annotations in observable message evidence; specified higher-level labels are then derived deterministically. The resulting structured representation allows influence-relevant features to be compared and aggregated across messages. The present study specifies and formally validates this representation; annotation reliability and computational annotation performance are not evaluated here.

The rest of the paper reviews existing frameworks and the resulting gap, states the premises and design questions, specifies and validates the ontology, presents the annotation workflow and worked example, and concludes with excluded dimensions, implications, and limitations.

2 Related frameworks

Existing frameworks address this terrain at several units of analysis: veracity and intent typologies, campaign- and incident-level frameworks, message-level technique inventories, and framing and narrative frameworks. Wardle and Derakhshan's information-disorder typology (Wardle and Derakhshan, 2017) separates falseness from intent to harm through misinformation, disinformation, and mal-information. For present purposes, mal-information is especially relevant because it covers accurate content used to harm, but remains defined by inferred intent rather than observable message form. Even if truth status were supplied externally, misinformation and disinformation would remain indistinguishable under intent-agnostic annotation, because the distinction between them turns on deceptive intent. Once truth status is also excluded, the three categories yield no distinction recoverable from message form alone. Disinformation typologies similarly centre false or misleading information coupled to harmful or profit-seeking intent (Kapantai et al., 2021; Plikynas et al., 2025). Knowledge-graph approaches such as Shahi (2024) extend this work computationally by aggregating fact-checked claims to support prescreening against prior verdicts.

These contributions are indispensable for governance and fact-checking, but they answer the question *is this true?*, or in the mal-information case *is this intended to harm?* They do not ask whether a message exhibits the form of influence work. Posts associated with different fact-checking verdicts show differences in textual and affective presentation, consistent with different kinds of claims and distortions appearing in distinct message-level packages that the claim-level verdict itself does not represent (Westlake et al., 2026). As the UKIP example illustrates, influence may operate through framing, emotional amplification, and identity activation without introducing falsity. The present study therefore asks: *Does the message itself bear traces of an*

orientation toward shaping interpretation, judgement, or response through means not reducible to the evidential, argumentative, or substantive merits of its claims?

Campaign- and incident-level frameworks instead model actors, infrastructure, behaviours, and lifecycle dynamics surrounding the artefact. The Disinformation Analysis & Risk Management (DISARM) Framework (DISARM Foundation, n.d.; Newman, 2022) organises operational tactics across Plan, Prepare, Execute, and Assess; the Influence Operation Ontology (Cayuela Tudela et al., 2025) models Threat, Channel, and Social domains; and the joint ENISA–EEAS work on Foreign Information Manipulation and Interference (FIMI) treats manipulative behaviour rather than content falsity as its unit of analysis (European Union Agency for Cybersecurity (ENISA) and European External Action Service (EEAS), 2022). Related approaches organise analysis by actor, behaviour, and content, procedural kill chains, or computational-propaganda and lifecycle models (François, 2019; Pamment, 2020; Nimmo and Hutchins, 2023; Oxford Internet Institute, n.d.; Bradshaw and Howard, 2019; Media Manipulation Casebook, n.d.). These frameworks are complementary because coordination, attribution, targeting, dissemination, and other operation-level properties require evidence beyond the individual artefact.

Message-level technique inventories operate where campaign-level frameworks go quiet. The Institute for Propaganda Analysis's seven classical devices (Institute for Propaganda Analysis, 1939) and the SemEval propaganda and persuasion shared-task lineage, from the 14-class news inventory of SemEval-2020 to multimodal meme and multilingual news tasks in 2021, 2023, and 2024, show that rhetorical mechanisms can be operationalised at span, paragraph, and message level with reusable taxonomies and computational benchmarks (Da San Martino et al., 2020; Dimitrov et al., 2021; Piskorski et al., 2023; Dimitrov et al., 2024).

Framing and narrative have long disciplinary anchors in Entman's account of framing as selection and salience (Entman, 1993) and in Benford and Snow's diagnostic/prognostic/motivational decomposition of collective action frames (Benford and Snow, 2000). Lenk (2026) more recently proposes a four-level hierarchy (semantic frames, strategic narratives, master-narratives, meta-frames) that connects two literatures that have often developed separately.

Taken together, these contributions do not treat method, narrative organisation, group positioning, and influence object as conceptually *orthogonal* dimensions of the same message. Orthogonality here is conceptual rather than statistical: dimensions may co-occur and correlate empirically, but no dimension is definitionally determined by another. Technique inventories operate primarily at the rhetorical-mechanism dimension. Lenk's framework extends the analysis into narrative structure, but it does not treat group-positioning effects or influence-object orientation as separate analytic

axes.

The cost of this gap is concrete: identical tactics can support different group-relational orientations, while messages on the same topic can bear toward different civic outcomes; those distinctions cannot be recovered from technique, topic, or narrative labels alone. Existing frameworks therefore answer complementary questions, but do not provide a representation of message-internal influence work when only the artefact is available. The present ontology addresses that gap. We develop the irreducibility argument under Construct validity and provide the full framework coverage check in Supplementary Appendix H.

3 Current study

3.1 Premises & assumptions

Two premises anchor the framework and deserve explanation because the analytical choices that follow depend on them.

Premise 1: Communicative influence leaves traces in message form. We assume that strategic communication is goal-directed. Messages do not merely describe the world; they warn, exhort, legitimise, delegitimise, mobilise, and discourage (Austin, 1962; Searle, 1969) because moving audiences toward action typically requires motivation (Ajzen, 1991; Fogg, 2009; Michie et al., 2011). Persuasion research further suggests that influence is pursued not only through extended deliberation but also through cues and devices that heighten affective engagement (Petty and Cacioppo, 1986; Kahneman, 2011; Green and Brock, 2000; Braddock and Dillard, 2016). Together, these traditions support a limited claim: when influence is carried in communicative form rather than solely through targeting, it will often leave observable traces in the message itself. We code these traces without treating them as evidence of authorial intent or achieved influence. Mechanisms that operate through targeting, distribution, or algorithmic amplification fall outside the present scope.

Premise 2: Communicative influence is not defined by falsity. Propaganda and influence operations aim to shape perceptions, identities, trust relations, and behaviour toward an intended end (Jowett and O'Donnell, 2018). Falsehood may serve as an instrument to achieve this, and fact-checking instruments remain valuable for adjudicating specific factual claims (Graves, 2018), but deception is neither necessary nor sufficient. Communicative influence can also operate through framing, association, identity activation, and moral positioning without fabrication; even accurate factual claims can shift interpretation through salience and selection (Entman, 1993).

Falsity is therefore not generally constitutive of the target of influence. A false claim may be used to alter belief, mobilise behaviour, damage trust, or shift policy support, but these influence targets are analytically distinct from the claim's truth status. A limiting case is the "firehose of falsehood" model, in which high-volume, contradictory, and unreliable messaging can undermine audiences' ability to distinguish what is credible or knowable (Paul and Matthews, 2016). Even there, however, the influence target is the audience's epistemic orientation rather than falsity as such. The full firehose model depends on cross-message and distributional properties and therefore falls outside the ontology's message-level scope.

We therefore adopt a truth-agnostic stance. The ontology codes observable cues in the message itself and treats them as evidence of influence-relevant form, not as evidence of intent, factual accuracy, or effect.

3.2 Design questions

The aim of this study is to specify a representation of communicative influence that can be applied when only the message is available and is designed for computational deployment at scale. Three design questions follow.

DQ1. Which dimensions are required to capture message-internal communicative influence, such that each can be operationalised from the artefact alone and none is reducible to the others?

DQ2. Can the ontology be specified so that its fixed derivation relations are deterministic and machine-checkable?

DQ3. Does the resulting representation cover the relevant distinctions made by existing frameworks, without redundancy within the message-internal scope?

The paper's contribution is therefore the specification itself: the dimensions, the literature-grounded canonical vocabularies through which they are operationalised, the criteria by which the dimensions were selected, and the formal properties of the derivation. Reliability and specification coherence are distinct questions: agreement establishes that annotation is reproducible, but cannot by itself establish that the categories are conceptually distinct or substantively warranted. Whether annotators or models can apply the vocabulary reliably is therefore a subsequent empirical question, not addressed here.

DQ1 is addressed by the inclusion criteria under Inclusion criteria, their application to each dimension under Basic components, and the irreducibility demonstrations under Construct validity. DQ2 is addressed by the reasoner and constraint-validation results under Formal verification. DQ3 is addressed by the coverage assessment under Construct validity, with the full mapping in Supplementary Appendix H.

4 Ontology structure

Building on these premises, the framework distinguishes four conceptually orthogonal dimensions of communicative influence, each addressing a different analytical question about the same message. Table 1 illustrates the resulting components using a simplified profile of the UKIP *Breaking Point* artefact; fuller annotation may license additional labels. The rightmost column shows the message-internal evidence supporting the base assignments.

The cue is the ontology's single inferential primitive: observable textual, visual, or other message-internal evidence that licenses a canonical base assignment. Annotators judge whether the available evidence satisfies a base definition. Once a base label, applicable attribute, or configuration membership has been assigned, only relations explicitly fixed by the specification are derived deterministically, as shown in Figure 2.

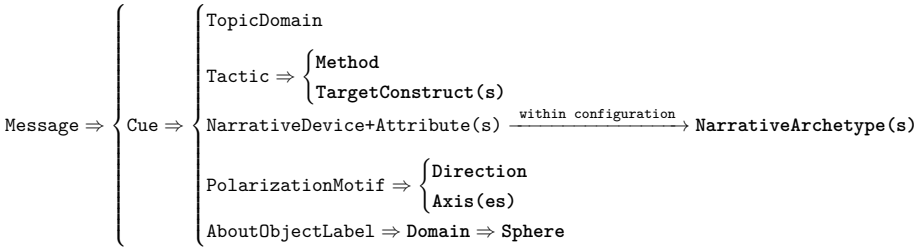


Figure 2: Message-level annotation pathway. Boldface denotes derived outputs.

In this schema, Method, Narrative, Polarization, and AboutObject are the four core analytical dimensions. TopicDomain independently indexes subject matter, while TargetConstruct is inherited deterministically from assigned Tactics. Cue-supported base assignments comprise Tactics, NarrativeDevices and applicable attributes or configuration memberships, PolarizationMotifs, AboutObjectLabels, and TopicDomain labels. Higher-level Method, NarrativeArchetype, Polarization direction and axes, AboutObject ancestors, and TargetConstruct labels are derived only where specified. Full vocabularies, spotting cues, boundary rules, and disambiguation examples appear in the supplemental appendices.

4.1 Inclusion criteria

A dimension qualifies for separate inclusion if it meets three conditions. The first two establish construct differentiation; the third follows from the ontology's message-level unit of analysis.

1. **Discriminant validity.** The dimension captures a distinct type of information not recoverable in full from the others, satisfying the requirement that constructs not simply restate one another (Campbell

Table 1: Ontology components illustrated with the UKIP *Breaking Point* artefact

Component	Question	Illustrative output	Evidence / derivation
Method	How does it influence?	*Loaded language*; *Causal oversimplification*; *Threat framing*; *Appeal to fear*; *Unity (shared identity)* → Framing; Persuasion	"BREAKING POINT"; "The EU has failed us all"; "we", "us all", "our borders"; advocated withdrawal and border control.
Target-Construct	What audience state is structured to move?	Belief; Attitude; Identity; Agency/Choice; Behavior	Deterministically inherited from the assigned Tactics.
Narrative	How is meaning organised?	*Role-casting (actants)*; *Stakes & urgency cues* → *Threat / Enemy at the gates*	External villain/opponent role-casting in relation to the message's "we/us" ingroup + existential/systemic stakes ("BREAKING POINT").
Polarization	How are groups positioned?	*Zero-sum* → Divide / Issue–Ideological; *Scapegoating / concentrated blame* → Divide / Affective–Identity; *Existential threat* → Divide / Affective–Identity	Ingroup ("we/us")–EU relation: control and concentrated blame. Ingroup ("we/us")–depicted migrant/refugee relation: existential threat.
About-Object	What broader outcome does the message bear toward?	`Policy-Support`; `VoteTurnout-Increase` → Politics → CivicOutcome	Leave-EU advocacy; referendum checkbox and "23rd June" voting cue.
Topic-Domain	What is the message about?	Immigration; Elections & Governance	Refugees/border control; Brexit referendum.

and Fiske, 1959; Messick, 1995). The evidence offered here is definitional and structural: operational definitions distinguish the analytical questions answered by the dimensions. Correlational multitrait–multimethod evidence would require annotated data and is not evaluated here.

2. **Independent variation.** The dimension is not uniquely determined by the others: values on the remaining dimensions do not in general fix its value. This is consistent with facet theory's treatment of facets as

conceptually distinct components (Guttman and Greenbaum, 1998).

3. **Operationalisability.** The dimension can be operationalised from evidence available in the message artefact itself, without requiring auxiliary evidence such as truth adjudication, actor attribution, or inferred intent. This follows the measurement-validity requirement that operationalisations be grounded in observations capable of supporting the concept being measured (Adcock and Collier, 2001).

The third condition is epistemological: private authorial intent, truth adjudication, and actor attribution require evidence beyond the artefact, consistent with the intentional-fallacy argument (Wimsatt and Beardsley, 1946), external verification requirements (Goldman, 1999), and attribution research using metadata or network evidence (Starbird et al., 2019).

The unit of analysis is the artefact submitted for annotation, whether a short post, extended thread, speech, or long-form document. A sequence presented as one artefact is annotated as one message, and relations internal to it count as message-internal evidence. Relations between separately annotated artefacts do not. The same rule governs the sequential Tactics in Supplementary Appendix B: any prior stage must be explicitly recoverable from the available artefact.

The vocabularies are not claimed exhaustive; the four dimensions are claimed the minimal set within the stated scope.

4.2 Basic components

4.2.1 Method: how influence operates

The Method dimension captures the functional mode of influence at the level of message form. Annotators assign a Tactic where observable message-internal evidence satisfies its operational definition and spotting rule, recording the supporting evidence as one or more cues. Each Tactic belongs to exactly one of six Method families: *Persuasion*, *Framing*, *Manipulation*, *Deception*, *Coercion*, and *Distraction* (Figure 3). Detailed tactic definitions, include-exclude rules, and method-boundary cases appear in Supplementary Appendix B.

Truth-agnostic annotation of deceptive form. The Deception family does not adjudicate factual falsity or authorial intent. A Deception Tactic is assigned only where a potentially misleading communicative operation is recoverable from the annotated artefact, including relations across communicative acts within it (Inclusion criteria). Mere falsity, fabrication, omission, impersonation, or insincerity is insufficient where establishing it requires external evidence. Qualifying operations instead depend on recoverable message-internal relations such as a comparison baseline, informational

Figure 3: Method–Tactic matrix with fixed TargetConstruct profiles.

Persuasion (25)				
Altercasting Iden · Behav	Appeal to anger / indignation Bel · Att · Behav	Appeal to authority Bel	Appeal to fear Agen · Behav	Appeal to novelty Bel · Att
Appeal to tradition Bel · Att	Associative transfer / guilt-or-honor by association Bel · Att	Authority-based directive Behav	Commitment / consistency Behav	Door-in-the-face (DITF) Behav
Doubt-seeding Bel	Foot-in-the-door (FITD) Behav	Ingratiation Att	Low-ball Behav	Preemptive inoculation / prebunking Bel · Att · Agen
Presuppositional embedding Bel	Reciprocity Behav	Repetition Bel	Scarcity appeal Att · Agen · Behav	Social proof (bandwagon) Bel · Behav
Source / channel credibility Bel	Statistical evidence Bel	Storytelling / narrative packaging Bel · Att	Two-sided with refutation Bel	Unity (shared identity) Bel · Att · Iden · Behav
Framing (16)				
Attribute framing Perc · Att	Causal oversimplification Bel	Collective nostalgic framing Bel · Att · Iden	Dehumanizing representation Bel · Att	False dilemma Bel · Agen
Gain/loss goal framing Behav	Inevitability framing Bel · Agen	Loaded language Att	Metaphor / analogy Bel	Moral framing Bel · Att
Name-calling / labeling Att	Ridicule / mockery Att	Risky-choice framing Behav	Scapegoating Bel · Att	Slippery-slope / consequential escalation Bel · Att
Threat framing Bel				
Distraction (8)				
Ad hominem / personal-credibility attack Perc · Bel · Att	Agenda shifting / evasive refocusing Perc	Humor as deflection Perc · Att	Information flooding / overload Perc · Agen	Obfuscation (jargon/complexity) Perc · Agen
Red herring Perc	Straw man Perc · Bel	Whataboutism Perc · Bel		
Manipulation (10)				
Affective associative cueing Perc	Appeal to pity / sympathy Bel · Att · Behav	Dangling contingent rewards Behav	Gaslighting / epistemic destabilization Bel · Iden · Agen	Guilt trip Bel · Att · Behav
Induced hypocrisy / self-inconsistency pressure Perc · Behav	Love-bombing Bel · Att · Behav	Manipulative choice architecture Agen · Behav	Pride / ego appeals Bel · Att · Iden · Behav	Shaming / derogation Iden · Behav
Deception (7)				
Exaggeration / overstatement Bel · Att	False balance / distorted evidential parity Bel	False equivalence / distorted comparability Bel · Att	Minimization / understatement Bel · Att	Salience-based concealment Perc · Bel
Selective evidence presentation (cherry-picking) Bel · Att	Strategic vagueness / equivocation Bel			
Coercion (5)				
Blackmail / disclosure leverage Agen · Behav	Coercive ultimatum Agen · Behav	Collective punishment threats Agen · Behav	Explicit threats Agen · Behav	Implied threats / coercive intimidation Agen · Behav

demand, represented evidence set, evidential parity, or material qualification. Full operational boundaries and worked examples are provided in Supplementary Appendix B.

4.2.2 TargetConstruct: what audience state is structured to move

TargetConstruct records the audience-state profile constitutively associated with a Tactic. It is not coded independently; instead, one or more values are inherited deterministically from the assigned Tactic. The ontology distinguishes six constructs: *Perception/Attention*, *Belief*, *Attitude*, *Identity*, *Agency/Choice*, and *Behavior*. Although TargetConstruct is not independently annotated, retaining the derived profile explicitly makes an analytically important distinction visible. How influence operates is not the same as what audience state its constitutive operation bears on. Tactics within the same Method family may therefore yield different TargetConstruct profiles, while Tactics from different Method families may converge on the same profile. The full Tactic-to-construct mapping appears in Supplementary Appendix A.

4.2.3 Narrative: how meaning is organised

Method captures communicative influence operations; Narrative captures how a message organises actors, events, relations, and stakes into represented story structure. Annotators identify NarrativeDevice instances such as role-casting, temporal sequencing, causal emplotment, metaphorical mapping, moral valuation, and stakes/urgency, and populate applicable closed-vocabulary attributes from message-internal evidence. Where several device instances occur, annotators additionally assign each instance to a message-local narrative configuration indicating which devices belong to the same compositional narrative thread. A message may contain several such configurations, including within a single overarching story, and a configuration may span non-contiguous portions of a longer artefact.

NarrativeArchetypes, by contrast, are recurring macro-story templates such as *Threat / Enemy at the gates*, *Enemy within (fifth column)*, or *Corruption / rot at the top*. They are not selected directly. Once the NarrativeDevice instances, their applicable attributes, and their configuration memberships are fixed, explicit signature rules are evaluated separately within each configuration. A configuration may therefore yield zero, one, or multiple NarrativeArchetypes, while unrelated narrative elements are not combined merely because they occur in the same artefact.

This separation captures narrative organisation in both short and long-form messages without collapsing story form into topic or requiring coders to reconstruct a campaign narrative from a single artefact. The ontology

does not assign a message an intrinsic position in a larger narrative arc such as setup, escalation, or resolution; such higher-order structure emerges only through aggregation across messages, sources, and time. Message-internal temporal sequencing or conflict–resolution structure remains codable where recoverable within the artefact. Device schemas, configuration rules, archetype signatures, and derivation rules are given in Supplementary Appendix C.

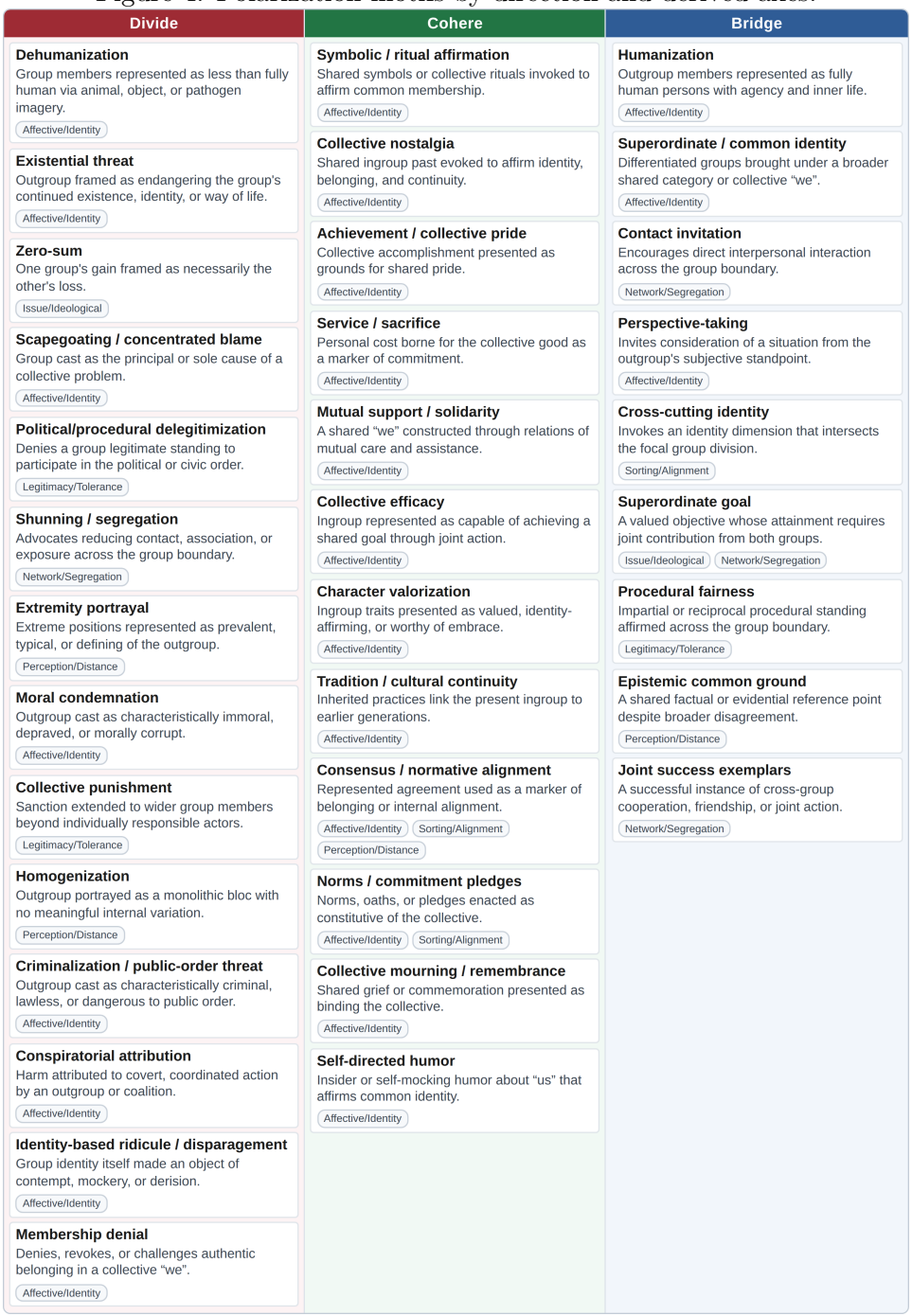
4.2.4 Polarization: how groups are positioned

The Polarization dimension captures group-relational positioning that is not reducible to individual attitude or identity targets. Its operationalisation draws on social identity theory for social categorisation and ingroup–outgroup differentiation (Tajfel and Turner, 1979), affective-polarisation research for group evaluation and social distance (Iyengar et al., 2012), the common ingroup identity model for boundary recategorisation (Gaertner et al., 1993), and relational accounts of politically salient "us"–"them" relations (McCoy et al., 2018).

Annotators identify Polarization motifs from message-internal evidence. Each motif maps deterministically to exactly one direction and a fixed nonempty set of axes. Direction records what the motif does to the group relation: Divide constructs, sharpens, or reinforces a salient boundary; Cohere constructs or reinforces a shared collective "we"; and Bridge crosses, softens, or supersedes a salient boundary. The axes record the dimensions of group relations instantiated by the motif: affective/identity, issue/ideological, sorting/alignment, perception/distance, network/segregation, or legitimacy/tolerance.

Separating direction from axis makes otherwise different motifs commensurable across directions. For example, shunning and contact invitation are opposite directional moves on the network/segregation axis. Where identifiable from the message, annotation may also record the group referents being positioned and their message-relative relational roles. A message may instantiate multiple Polarization motifs when independently supported by message-internal evidence; where no salient group-positioning operation is recoverable, no Polarization motif is assigned. Figure 4 presents the codable Divide/Cohere/Bridge motif board and their derived axes. Full motif definitions and deterministic direction/axis mappings are provided in Supplementary Appendix D.

Figure 4: Polarization motifs by direction and derived axes.



4.2.5 AboutObject: the broader outcome toward which influence is oriented

AboutObject captures the broader outcome toward which a message is oriented, as indicated by message-internal cues; it does not imply subjective authorial intent. It is distinct from both topic and audience state. TopicDomain records what a message is about; TargetConstruct records which audience state a Tactic is structured to move; AboutObject records the broader outcome toward which the message is oriented, such as policy support, election legitimacy, institutional trust, protest participation, purchase decisions, or subscription uptake.

The ontology models AboutObject through a three-level structure: broad sphere tags at the top (civic, commercial, or both), domain-specific categories in the middle, and specific influence-object labels at the bottom. Annotators assign specific L3 AboutObject labels from message-internal evidence using the operational definitions and evidence templates. The corresponding L2 domain and L1 sphere are then derived deterministically. Figure 5 visualises the `AboutObject` hierarchy as an L1–L2–L3 matrix. The full label inventory, evidence templates, and hierarchy are provided in Supplementary Appendix F.



Figure 5: AboutObject vocabulary. L3 labels are grouped by L2 domain and inherit L1 sphere; symbols indicate directional or relational label variants.

4.2.6 TopicDomain: what the message is about

`TopicDomain` independently indexes what the message is about, using macro-domains adapted from established policy and issue taxonomies including EuroVoc, the Comparative Agendas Project, and IPTC Media Topics (Publications Office of the European Union, n.d.; Baumgartner and Jones, 2009; Kuzman and Ljubešić, 2025), with additions for salient contemporary communication issues. Topic does not itself indicate influence: a message about immigration, for example, is not thereby an influence message. Its role is substantive indexing, enabling annotations to be compared and aggregated across issue areas. The full domain list appears in Supplementary Appendix E.

4.3 Validation and verification

Validation and verification proceed along two tracks. Construct-validity evidence, in the sense of Messick (1995), concerns whether the specified constructs are substantively grounded, whether the dimensions are distinct from one another, and whether together they span the domain the ontology claims to represent. Formal verification separately concerns the ontology's specified structural and derivation properties using its OWL and SHACL reference representation.

4.3.1 Construct validity

Substantive grounding The canonical vocabularies are literature-grounded rather than generated solely from the architecture. The operational appendices provide definitions, evidence conditions, boundary rules, and mappings, while Supplementary Appendix J documents the literature basis of the canonical categories and an extensive provenance ledger linking candidate constructs to the final specification, including 11 rejection decisions.

Discriminant validity The four-dimensional architecture did not begin from a fixed list. An earlier version contained narrative, a manipulation layer that later became Method, and a call-to-action layer. Polarization split off once it became clear that group-relational positioning, although often carried by the same tactics, could not be recovered from those tactics alone. AboutObject had a different origin. The call-to-action layer assumed that influence ultimately aims at behaviour, but communicative influence may instead shift evaluation, undermine institutional trust, alter policy support, or bear toward other civic or commercial outcomes without requesting immediate action. Subject-matter experts exposed the limitation by asking how messages using similar influence methods but bearing toward different outcomes should be distinguished. AboutObject was developed to capture

that distinction; calls to action remain recoverable where relevant but no longer define the dimension.

This pressure reflects a well-documented risk in measurement design. Conflating separable constructs undermines both discriminant validity and the interpretive utility of the resulting measurement (Campbell and Fiske, 1959; Messick, 1995). In text-as-data research, annotation schemes that merge separable dimensions can produce classification artefacts that masquerade as substantive findings (Grimmer and Stewart, 2013).

These limitations run in both directions. A tactic-only scheme cannot recover group-relational positioning, while Polarization does not recover Method in full; topic cannot recover influence object; and narrative cannot determine how claims are supported, while Method cannot reconstruct plot. Removing a dimension therefore loses information the others cannot recover. Table 2 demonstrates this by holding one dimension constant across paired hypothetical messages while varying another.

Table 2 shows independent variation rather than mutual exclusivity. A shared Method does not determine Polarization; a shared TopicDomain does not determine AboutObject; and a shared Narrative does not determine Method. These contrasts are reinforced by explicit cross-layer boundary rules in Supplementary Appendices C–D, including for similarly named constructs across Method and Polarization; cross-layer annotations require independent satisfaction of each branch's criteria. Applied adjudications in Supplementary Appendix I illustrate the same boundary logic: in Case 3, Method-level *Threat framing* does not license Polarization *Existential threat* because the stronger existential-threat criterion is not met. The worked example under Example adds a Narrative–Polarization case, deriving *Enemy within (fifth column)* without DIVIDE, since represented internal antagonism does not by itself satisfy the group-relational criteria of a Polarization motif.

Table 2: Irreducibility contrasts across dimensions. Each row holds one dimension constant across two hypothetical messages while varying another. If the varying distinction cannot be recovered from the held-constant dimension, the two dimensions are not reducible to one another.

Held constant across messages	**What varies**	**Illustrative contrast**
Method: *Appeal to fear*	Polarization	Both messages use the same threat–protective-response structure: "This threat puts us at risk—protect our own" (DIVIDE) versus "This threat puts us at risk—stand together" (COHERE). The Method is held constant, but the represented group relation differs.
TopicDomain: Immigration	AboutObject	Both messages concern immigration, but one bears toward electoral participation (e.g., "vote them out"), whereas the other bears toward institutional trust (e.g., "the system cannot be trusted"). The topic is held constant, but the broader influence object differs.
Narrative: *Corruption / rot at the top*	Method	Both messages represent the same elite-corruption narrative, but one supports that representation through specific sourced evidence, whereas the other relies on insinuation, unnamed sourcing, or agentless construction. The narrative structure is held constant, but the communicative operation differs.

Content validity Coverage was assessed systematically against the frameworks reviewed under Related frameworks, asking whether each introduced either an unrepresented message-internal dimension or a substantively distinct category requiring extension; the full mapping appears in Supplementary Appendix H. No framework in the bounded comparison set required an additional dimension. Apparent gaps resolved into four outcomes: existing coverage, within-dimension vocabulary refinement, cross-dimensional decomposition, or an external-evidence scope boundary. Veracity- and intent-based frameworks remain complementary rather than unrepresented: their distinctions require evidence outside the artefact and therefore mark a scope boundary rather than a coverage gap. Manufactured

uncertainty, for example, exposed an underspecified within-dimension operation and was canonicalised as *Doubt-seeding* within Persuasion with *Belief* as its fixed TargetConstruct (Michaels, 2006; Stocking and Holstein, 2009; Biddle and Leuschner, 2015; Gustafson and Rice, 2020). This supports content validity within the stated scope, but does not establish exhaustive coverage of every future conceptualisation of communicative influence.

4.3.2 Formal verification

The ontology's OWL 2 and SHACL representation was machine-checked for logical coherence and conformance to the specified structural and derivation constraints (Knublauch and Kontokostas, 2017). OWL 2 DL classification with HermiT identified no named unsatisfiable classes; the conforming graph passed SHACL, the deliberately non-conforming graph returned the expected violations, and NarrativeArchetype and NarrativeConfiguration-scope tests behaved as specified (Musen and Protégé Team, 2015; Glimm et al., 2014). These checks verify the reference representation against the stipulated formal structure; they do not establish empirical construct validity, annotation reliability, or downstream predictive performance. Supplementary Appendix G provides the verification artefacts, software versions, procedure, and detailed results.

5 Annotation workflow

5.1 Workflow

Annotation separates evidence-supported base assignment from deterministic derivation. Coders assign a canonical base category when message-internal evidence meets its definition and minimum-evidence conditions, recording the supporting textual span, visual feature, or other element as the cue. Cue identification is therefore part of, not prior to, base-label assignment.

Base assignment applies to Tactics, NarrativeDevices and applicable attributes or configuration memberships, Polarization motifs, TopicDomains, and specific AboutObject labels. A message element may support multiple annotations when each independently meets its criteria; overlap does not license cross-dimensional carry-over.

For Narrative, annotators additionally assign each NarrativeDevice instance to a local narrative configuration recording which device instances are eligible to compose together. Archetype derivation then evaluates the fixed signatures separately within each configuration.

Deterministic processing begins only after the relevant base annotation has been assigned, and only for relations explicitly fixed by the specification. An assigned Tactic deterministically yields its Method family and

predefined TargetConstruct set. Within each narrative configuration, assigned NarrativeDevice instances and their applicable attributes yield a NarrativeArchetype whenever they satisfy a specified minimal signature. Each assigned Polarization motif yields exactly one direction and a fixed nonempty set of axes. A specific L3 AboutObject label deterministically inherits its L2 domain and L1 sphere. Where the evidence supports only the residual `Unclear` label, no L2 domain or L1 sphere is derived. TopicDomain is independently assigned and does not derive from, or determine, the influence annotations.

This division localises interpretive uncertainty to base assignment while making downstream derivation reproducible and machine-checkable. This accommodates genuine annotation ambiguity (Plank, 2022) and the contextual interpretation required in persuasion-technique detection (Piskorski et al., 2023). Once assignments, applicable attributes, and configuration memberships are fixed, downstream derivation follows the specification: *evidence-supported base assignment → deterministic derivation.*

Computational treatment of localised message evidence and fine-grained communicative categories has precedent in established NLP tasks: CoNLL-2010 required systems to identify hedge cues and resolve their scope (Farkas et al., 2010), while NLP4IF-2019 required fragment-level identification of specific propaganda techniques (Da San Martino et al., 2019). These tasks establish precedents for computational localisation and classification, not for the ontology's deterministic derivation structure. Computational systems may therefore assist with or learn the interpretive judgement that message evidence satisfies an annotation definition; specified rule application begins only after the relevant annotation has been accepted.

5.2 Example

To illustrate message-level application, we use a synthetic election-related message written to resemble contemporary civic discourse without referring to a real case:

> "They're calling it a routine verification.
>
> But why are ballot batches being moved at 2:17 a.m., after observers were sent home?
>
> Local officials insist everything is secure. The same officials who quietly changed the counting procedures last week.
>
> This isn't about party. It's about whether your vote still means something.
>
> If procedures can be rewritten in the dark, what happens next time?
>
> We deserve transparency. We deserve oversight.
>
> Share this. Ask questions. Demand answers before it's too late."

At the level of **Method**, several parts of the message license canonical Tactics. The contrast between "routine verification" and the subsequent challenge, and between the officials' assurance and their represented prior conduct, supports *Two-sided with refutation*; the questions and juxtaposed procedural details support *Doubt-seeding* and *Presuppositional embedding*. The same represented prior conduct also supports *Ad hominem / personal-credibility attack* because it is invoked to discount the officials' assurance, illustrating that the same evidence may satisfy more than one Tactic definition where the criteria for each are met. The threatened continuation of procedural manipulation supports *Threat framing*; where that threatened harm functions as a reason to seek transparency and oversight, it additionally supports *Appeal to fear*. The diminishing opportunity conveyed by "before it's too late" separately supports *Scarcity appeal*. The remaining Method assignments are summarised in Table 3.

The timestamp "2:17 a.m." is a specific detail but does not by itself constitute *Statistical evidence*; nor does the message license a Deception Tactic merely because represented relations are implied, agentless, or unverified externally. The ontology codes communicative operations recoverable from the message rather than adjudicating whether the represented events occurred.

Taken together, the Tactic assignments in Table 3 derive three **Method** families, Persuasion, Framing, and Distraction, and all six **TargetConstruct** values. This saturation results from aggregation across the assigned Tactics and is not a requirement of the ontology.

At the level of **Narrative**, the assigned NarrativeDevice instances participate in a single represented electoral-procedure thread and are therefore assigned to one message-local narrative configuration, `C1`. Role-casting represents the officials as an internal elite opponent (`ActorGroup=elite_in_group`, `ActorLocation=internal`, `Function=villain/opponent`) in relation to `institutions/rule_of_law`; Causal emplotment represents internal institutional undermining and licenses `subversion/collusion`. Temporal sequencing & contrast is coded with `Direction=decline`, Moral valuation with `Axis=fairness/cheating`, and Stakes & urgency with `Level=existential/systemic` and `Domain=institutional/order`.

Because these device instances share `C1`, the fixed signatures are evaluated within that configuration and derive *Enemy within (fifth column)* and *Crisis & collapse*. *Enemy within (fifth column)* is witnessed by Role-casting together with the `subversion/collusion` Causal emplotment instance; *Crisis & collapse* is witnessed by Temporal sequencing & contrast together with Stakes & urgency at `Level = existential/systemic` and `Domain = institutional/order`. These annotations concern relations represented by the message and do not imply that the officials acted subversively, that the described events occurred, or that the electoral system was in fact collapsing.

At the level of **Polarization**, one motif is assigned. "This isn't about party" makes a partisan distinction salient while rejecting it as the relevant

basis for the appeal; "your vote" and "we deserve" then shift the message toward a broader civic identity centred on voters. Read across the sequence, this recasts the salient partisan boundary within a more inclusive collective "we", supporting *Superordinate / common identity* and deriving BRIDGE and Affective/Identity. The appeal to transparency, oversight, and fair electoral procedure does not additionally license *Procedural fairness*, which requires procedural standing to be affirmed across salient differentiated groups.

The message does not satisfy a DIVIDE motif. Although "quietly", "in the dark", and the juxtaposition of observers being sent home with subsequent ballot movement invite suspicion, they do not establish the covert, coordinated causal plot by a salient outgroup or coalition required for *Conspiratorial attribution*. The derived *Enemy within (fifth column)* archetype therefore does not itself imply DIVIDE: Narrative captures represented internal antagonism, whereas Polarization requires a recoverable group-relational positioning operation.

TopicDomain independently indexes the message as *Elections & Governance*. Four specific **AboutObject** labels are recoverable: `ElectionLegitimacy-Decrease`, `InstitutionalTrust-Decrease`, `InformationSeeking-Increase`, and `PublicExpression-Increase`. Their L2 domains and common L1 `CivicOutcome` sphere are inherited deterministically, as summarised in Table 3.

The example illustrates the ontology's central separation between evidence-supported assignment and deterministic derivation. A single message may support several independently licensed annotations across peer dimensions, while fixed mappings operate only after the relevant assignments—including, for Narrative, configuration membership—have been established. None of these annotations requires a judgement about the factual truth of the represented events or the author's private intent. Additional worked examples and interface details are provided in Supplementary Appendix I; annotation effort and its consequences for deployment are discussed under Annotation cost and the deployment path.

6 Discussion & conclusion

6.1 Excluded candidate dimensions

Candidate dimensions were assessed against the same conditions of discriminant validity, independent variation, and message-internal operationalisability.

Affective tone and emotion are pervasive in communicative influence, but much of this variation is already constitutive of represented Tactics and Polarization motifs. A fear appeal, for example, already incorporates a fear-relevant threat appraisal as part of its defining structure (Lazarus, 1991). A separate emotion dimension would therefore largely redescribe information

Table 3: Compact derivation summary

Branch	**Output**
Tactics (assigned)	*Two-sided with refutation*; *Doubt-seeding*; *Presuppositional embedding*; *Ad hominem / personal-credibility attack*; *Loaded language*; *Moral framing*; *Appeal to anger / indignation*; *Threat framing*; *Appeal to fear*; *Scarcity appeal*; *Slippery-slope / consequential escalation*; *Unity (shared identity)*; *Storytelling / narrative packaging*
Method (derived)	Persuasion; Framing; Distraction
TargetConstruct (derived)	Perception/Attention; Belief; Attitude; Identity; Agency/Choice; Behavior
Narrative (assigned)	`C1`: Role-casting (`elite_in_group`, `internal`, `villain/opponent`, `institutions/rule_of_law`); Causal emplotment (`subversion/collusion`); Temporal sequencing & contrast (`decline`); Moral valuation (`fairness/cheating`); Stakes & urgency (`existential/systemic`, `institutional/order`);
NarrativeArchetypes (derived)	*Enemy within (fifth column)*; *Crisis & collapse*
Polarization	*Superordinate / common identity* → BRIDGE → Affective/Identity
TopicDomain	Elections & Governance
AboutObject	`ElectionLegitimacy-Decrease` → Politics → CivicOutcome; `InstitutionalTrust-Decrease` → Epistemic Trust, Legitimacy & Attribution → CivicOutcome; `InformationSeeking-Increase` → Information & Influence Resilience → CivicOutcome; `PublicExpression-Increase` → Personal Security, Privacy & Public Expression → CivicOutcome

already captured by Method and Polarization. Residual affective variation, such as intensity differences within an identical tactic set, may exist but does not meet the threshold of irrecoverable information loss required for a separate dimension.

Epistemic register, the mode in which a message grounds its knowledge claims, presents a subtler case. Two messages may share identical Tactic profiles yet differ in how they warrant their claims. Its observable form, where a message performs a scientific, testimonial, or revelatory register, is recoverable through message-internal patterns already represented in Method; conspiratorial register, for example, surfaces through devices such as insinuation, unnamed sourcing, and agentless construction (Barkun, 2013). The genuinely independent construct, actual epistemic grounding, requires verification of sources, credentials, and evidence chains outside the artefact (Goldman, 1999). Epistemic register therefore fails either independence or

message-internal operationalisability, depending on how it is defined.

Removing any of the four dimensions loses information the others cannot recover; the candidate dimensions considered above either fail the inclusion threshold or require unavailable evidence. Within the stated scope, the four dimensions therefore form a minimal complete set under the inclusion conditions. TargetConstruct is not an additional peer dimension but a profile inherited deterministically from Tactic assignments, preserving the distinction between how influence operates and what audience state its constitutive operation bears on.

6.2 Implications

The ontology contributes measurement infrastructure rather than an explanatory theory. Each message yields a structured feature vector over Method, Narrative, Polarization, and AboutObject, enabling comparative corpus analysis and downstream modelling within a theoretically motivated, formally constrained feature space (Garla and Brandt, 2012).

In practice, the ontology is best understood as a human-authored, computationally implementable measurement layer for large-scale message streams. Within an augmented-intelligence workflow, it formalises distinctions grounded in expert judgement while computational systems extend those distinctions to volumes of data no unaided analyst could process (Aras and Westlake, 2025). Scale matters substantively, not just operationally: recent work documents AI-enabled disinformation systems that already generate personalised content, synthetic personas, and adaptive campaign behaviour at an operational tempo beyond plausible manual analysis (Bergmanis-Korāts and Chia Tee Hiang, 2026).

6.3 Annotation cost and the deployment path

The representation is deliberately fine-grained: Method distinguishes 71 canonical Tactics across six families; Narrative requires device instances, applicable closed-vocabulary attributes, and message-local configuration membership; Polarization distinguishes 35 motifs; and AboutObject is assigned at the most specific level of a three-level hierarchy. This granularity preserves substantively distinct operations and outcomes within each dimension; the inclusion criteria under Inclusion criteria establish the need for the four dimensions, while their finer distinctions provide the codebook's analytical resolution.

The development-stage exercise in Supplementary Appendix I provides a descriptive annotation-cost benchmark: one analyst with prior content-annotation experience but no prior exposure to the ontology averaged approximately thirteen minutes per message across 34 short social-media messages. Because the exercise used a pre-freeze specification and did not assess

agreement, it is not an estimate of final-specification reliability or trained-annotator performance.

Unaided human annotation at this granularity is implausible for corpus-scale coverage. Coarsening would reduce cost but sacrifice distinctions the specification was designed to preserve. Human annotation remains feasible for bounded subsets, but corpus-scale retrieval and stratification require computational or model-assisted annotation, with humans retained for specification, calibration, adjudication, audit, and close annotation. Hybrid LLM pre-annotation with human verification has made fine-grained labelling tractable on short-form social-media data (Sahitaj et al., 2025); whether the present vocabulary can be applied computationally at acceptable accuracy remains an open empirical question.

6.4 Limitations

Certain areas are deliberately out of scope. The ontology does not by itself detect influence operations, establish authorial intent, measure real-world impact, or capture influence that operates only through targeting, placement, algorithmic amplification, or cross-message sequencing.

Cue exemplars are drawn primarily from English-language political communication, and cross-cultural adoption requires more than translation of the codebook. The four core dimensions together with the derived TargetConstruct profile may provide an analytic scaffold, but transferability is an empirical question. Research on the ethnography of communication and contextualisation cues shows that communicative meaning depends on locally learned norms and inferential conventions (Hymes, 1974; Gumperz, 1982), while cultural psychology documents variation in self-construal, motivation, and social categories (Markus and Kitayama, 1991). Adaptation therefore requires reconstructing the cue layer through domain literature, native-language expertise, area specialists, and pilot annotation rather than merely translating existing cues. What counts as a cue for threat, solidarity, or authority depends on locally available carriers such as honorific systems, historical or religious references, and group categories.

A further design tension concerns the inclusion of *Deception* in a truth-agnostic ontology. Deception Tactics are coded only where a potentially misleading communicative operation is recoverable from the message without adjudicating external truth or authorial intent; mere falsity, unsupported assertion, ambiguity, or omission is insufficient by itself. Full operational boundaries are specified in Supplementary Appendix B.

Dual-use considerations. Any representational framework for influence can be inverted from analysis to synthesis, and this ontology is no exception. By defining a closed, machine-readable feature space for message-internal

influence, it lowers the cost of producing messages that instantiate specified combinations of Method, Narrative, Polarization, and AboutObject. The same formalisation also makes message-level influence more measurable, auditable, and amenable to detection, benchmarking, and corpus-level comparison. Hostile actors iterate on influence content without needing to represent what they do; defenders, whose work is comparison, aggregation, and audit, must. We believe the asymmetry favours publication. Operations that fall outside message-internal analysis require complementary network-, platform-, attribution-, or campaign-level evidence; the ontology is not a protective system on its own.

No policy layer by design. An earlier version included a policy layer intended to register whether a message had actually influenced its audience. That is a claim about effect, and effect cannot be recovered from the artefact alone. As the project shifted from effect to message form, the layer was removed: the four analytical dimensions describe what a message represents and how it is built to do influence work, not whether the attempt succeeded. The truth-agnostic stance, by contrast, was retained throughout: the framework does not assign truth values to represented claims.

The ontology is therefore an analytical and measurement instrument, not an enforcement or risk-scoring framework. Questions about its use in moderation, platform governance, or counter-influence operations remain institutional and normative questions outside the specification (see Pijpers, 2023).

6.5 Conclusion

This article has presented a formally specified, multi-dimensional ontology for analysing communicative influence at the level of message form. The framework is message-internal, truth-agnostic, and deterministically derived above the base annotation layer. It separates four dimensions: Method, Narrative, Polarization, and AboutObject. Together, these distinguish communicative operation, represented story structure, group-relational positioning, and broader influence object, so that variation on one need not be treated as variation on another. For example, messages employing the same tactics can differ in narrative organisation or group positioning, while messages addressing the same topic can bear toward different influence objects.

The ontology is a specification rather than a theory of influence. It defines what can be recorded about a message, on what evidence, and how those records compose, providing a closed vocabulary for annotation, comparison across artefacts, and aggregation across corpora. The specification makes two empirical questions testable: whether these distinctions are consequential, and whether annotators (human or machine) can apply them reliably.

Data availability

The codebook, framework mappings, formal verification results, and worked annotation examples are provided in the Supplementary Information. The machine-readable OWL ontology, SHACL shapes, and associated verification artefacts are available at this OSF view only link.

References

Adcock R, Collier D (2001) Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review* 95(3):529–546. https://doi.org/10.1017/S0003055401003100.

Ajzen I (1991) The theory of planned behavior. *Organizational Behavior and Human Decision Processes* 50(2):179–211. https://doi.org/10.1016/0749-5978(91)90020-T.

Aras D, Westlake T (2025) Advancing early warning mechanisms: an augmented intelligence-driven model for predicting multidimensional vulnerability levels in Afghanistan. *Data & Policy* 7:e16. https://doi.org/10.1017/dap.2024.91.

Austin JL (1962) *How to Do Things with Words*. Harvard University Press, Cambridge, MA.

Barkun M (2013) *A Culture of Conspiracy: Apocalyptic Visions in Contemporary America*. 2nd edn. University of California Press, Berkeley. https://doi.org/10.1525/9780520956520.

Baumgartner FR, Jones BD (2009) *Agendas and Instability in American Politics*. 2nd edn. University of Chicago Press, Chicago. https://doi.org/10.7208/chicago/9780226039534.001.0001.

Benford RD, Snow DA (2000) Framing processes and social movements: An overview and assessment. *Annual Review of Sociology* 26:611–639. https://doi.org/10.1146/annurev.soc.26.1.611.

Bergmanis-Korāts G, Chia Tee Hiang J (2026) Beyond spam bots: The rise of AI-powered disinformation machines and the imperative for strategic response. Technical report, NATO Strategic Communications Centre of Excellence. https://stratcomcoe.org/publications/beyond-spam-bots-the-rise-of-ai-powered-disinformation-machines-and-the-imperative-for-strategic-response/342.

Biddle JB, Leuschner A (2015) Climate skepticism and the manufacture of doubt: Can dissent in science be epistemically detrimental? *European*

Journal for Philosophy of Science 5(3):261–278. https://doi.org/10.1007/s13194-014-0101-x.

Braddock K, Dillard JP (2016) Meta-analytic evidence for the persuasive effect of narratives on beliefs, attitudes, intentions, and behaviors. *Communication Monographs* 83(4):446–467. https://doi.org/10.1080/03637751.2015.1128555.

Bradshaw S, Howard PN (2019) The global disinformation order: 2019 global inventory of organised social media manipulation. Working Paper 2019.2, Project on Computational Propaganda, Oxford Internet Institute, University of Oxford. https://www.oii.ox.ac.uk/news-events/reports/the-global-disinformation-order-2019-global-inventory-of-organised-social-media-manipulation/.

Campbell DT, Fiske DW (1959) Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin* 56(2):81–105. https://doi.org/10.1037/h0046016.

Cayuela Tudela AD, Pastor-Galindo J, Nespoli P et al (2025) The influence operation ontology (IOO). https://arxiv.org/abs/2503.07304.

Da San Martino G, Barrón-Cedeño A, Nakov P (2019) Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection. In: Feldman A, Da San Martino G, Barrón-Cedeño A et al (eds) *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pp 162–170, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-5024.

Da San Martino G, Barrón-Cedeño A, Wachsmuth H et al (2020) SemEval-2020 task 11: Detection of propaganda techniques in news articles. In: Herbelot A, Zhu X, Palmer A et al (eds) *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pp 1377–1414, Barcelona (online). International Committee for Computational Linguistics. https://doi.org/10.18653/v1/2020.semeval-1.186.

Dimitrov D, Bin Ali B, Shaar S et al (2021) SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In: Palmer A, Schneider N, Schluter N et al (eds) *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pp 70–98, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.semeval-1.7.

Dimitrov D, Alam F, Hasanain M et al (2024) SemEval-2024 task 4: Multilingual detection of persuasion techniques in memes. In: Ojha AK,

Doğruöz AS, Madabushi HT et al (eds) *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pp 2009–2026, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.semeval-1.275.

DISARM Foundation (n.d.) DISARM red framework. https://www.disarm.foundation/framework. Accessed 28 Aug 2026.

Entman RM (1993) Framing: Toward clarification of a fractured paradigm. *Journal of Communication* 43(4):51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x.

European Union Agency for Cybersecurity (ENISA), European External Action Service (EEAS) (2022) Foreign information manipulation and interference (FIMI) and cybersecurity: Threat landscape. Technical report, European Union Agency for Cybersecurity. https://www.enisa.europa.eu/publications/foreign-information-manipulation-interference-fimi-and-cybersecurity-threat-landscape.

Farkas R, Vincze V, Móra G et al (2010) The CoNLL-2010 shared task: Learning to detect hedges and their scope in natural language text. In: *Proceedings of the Fourteenth Conference on Computational Natural Language Learning – Shared Task*, pp 1–12, Uppsala, Sweden. Association for Computational Linguistics. https://aclanthology.org/W10-3001/.

Fogg BJ (2009) A behavior model for persuasive design. In: *Proceedings of the 4th International Conference on Persuasive Technology*, Persuasive '09, pp 1–7, New York, NY, USA. Association for Computing Machinery. https://doi.org/10.1145/1541948.1541999.

François C (2019) Actors, behaviors, content: A disinformation ABC: Highlighting three vectors of viral deception to guide industry and regulatory responses. Technical report, Transatlantic Working Group on Content Moderation Online and Freedom of Expression. https://docs.house.gov/meetings/SY/SY21/20190926/109980/HHRG-116-SY21-Wstate-FrancoisC-20190926-SD001.pdf.

Gaertner SL, Dovidio JF, Anastasio PA et al (1993) The common ingroup identity model: Recategorization and the reduction of intergroup bias. *European Review of Social Psychology* 4(1):1–26. https://doi.org/10.1080/14792779343000004.

Garla VN, Brandt C (2012) Ontology-guided feature engineering for clinical text classification. *Journal of Biomedical Informatics* 45(5):992–998. https://doi.org/10.1016/j.jbi.2012.04.010.

Glimm B, Horrocks I, Motik B et al (2014) HermiT: An OWL 2 reasoner. *Journal of Automated Reasoning* 53(3):245–269. https://doi.org/10.1007/s10817-014-9305-1.

Goldman AI (1999) *Knowledge in a Social World*. Oxford University Press, Oxford. https://doi.org/10.1093/0198238207.001.0001.

Graves L (2018) Understanding the promise and limits of automated fact-checking. Factsheet, Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/our-research/understanding-promise-and-limits-automated-fact-checking.

Green MC, Brock TC (2000) The role of transportation in the persuasiveness of public narratives. *Journal of Personality and Social Psychology* 79(5):701–721. https://doi.org/10.1037/0022-3514.79.5.701.

Grimmer J, Stewart BM (2013) Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis* 21(3):267–297. https://doi.org/10.1093/pan/mps028.

Gumperz JJ (1982) *Discourse Strategies*. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511611834.

Gustafson A, Rice RE (2020) A review of the effects of uncertainty in public science communication. *Public Understanding of Science* 29(6):614–633. https://doi.org/10.1177/0963662520942122.

Guttman R, Greenbaum CW (1998) Facet theory: Its development and current status. *European Psychologist* 3(1):13–36. https://doi.org/10.1027/1016-9040.3.1.13.

Hymes D (1974) *Foundations in Sociolinguistics: An Ethnographic Approach*. University of Pennsylvania Press, Philadelphia, PA.

Institute for Propaganda Analysis (1939) *The Fine Art of Propaganda: A Study of Father Coughlin's Speeches*. Harcourt, Brace and Company, New York.

Iyengar S, Sood G, Lelkes Y (2012) Affect, not ideology: A social identity perspective on polarization. *Public Opinion Quarterly* 76(3):405–431. https://doi.org/10.1093/poq/nfs038.

Jowett GS, O'Donnell V (2018) *Propaganda & Persuasion*. 7th edn. SAGE Publications, Thousand Oaks, CA.

Kahneman D (2011) *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.

Kapantai E, Christopoulou A, Berberidis C et al (2021) A systematic literature review on disinformation: Toward a unified taxonomical framework. *New Media & Society* 23(5):1301–1326. https://doi.org/10.1177/1461444820959296.

Klincewicz M, Alfano M, Ebrahimi Fard A (2025) Slopaganda: The interaction between propaganda and generative AI. *Filosofiska Notiser* 12(1):135–162. https://www.filosofiskanotiser.com/KlincewiczAlfanoFard.pdf.

Knublauch H, Kontokostas D (2017) Shapes constraint language (SHACL). W3C recommendation, World Wide Web Consortium (W3C). https://www.w3.org/TR/shacl/.

Kuzman T, Ljubešić N (2025) LLM teacher-student framework for text classification with no manually annotated data: A case study in IPTC news topic classification. *IEEE Access* 13:35621–35633. https://doi.org/10.1109/ACCESS.2025.3544814.

Lazarus RS (1991) *Emotion and Adaptation*. Oxford University Press, New York. https://doi.org/10.1093/oso/9780195069945.001.0001.

Lenk T (2026) Towards a deeper understanding of information manipulation: Proposing a multilevel framework for the analysis of manipulative narratives. *Humanities and Social Sciences Communications* 13:343. https://doi.org/10.1057/s41599-026-06656-8.

Markus HR, Kitayama S (1991) Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review* 98(2):224–253. https://doi.org/10.1037/0033-295X.98.2.224.

McCoy J, Rahman T, Somer M (2018) Polarization and the global crisis of democracy: Common patterns, dynamics, and pernicious consequences for democratic polities. *American Behavioral Scientist* 62(1):16–42. https://doi.org/10.1177/0002764218759576.

Media Manipulation Casebook (n.d.) Methods: The media manipulation life cycle. https://mediamanipulation.org/methods/. Accessed 28 Aug 2026.

Messick S (1995) Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist* 50(9):741–749. https://doi.org/10.1037/0003-066X.50.9.741.

Michaels D (2006) Manufactured uncertainty: Protecting public health in the age of contested science and product defense. *Annals of the New York*

Academy of Sciences 1076(1):149–162. https://doi.org/10.1196/annals.1371.058.

Michie S, van Stralen MM, West R (2011) The behaviour change wheel: A new method for characterising and designing behaviour change interventions. *Implementation Science* 6:42. https://doi.org/10.1186/1748-5908-6-42.

Musen MA, Protégé Team (2015) The protégé project: A look back and a look forward. *AI Matters* 1(4):4–12. https://doi.org/10.1145/2757001.2757003.

Newman H (2022) Foreign information manipulation and interference defence standards: Test for rapid adoption of the common language and framework DISARM. Hybrid CoE Research Report 7, European Centre of Excellence for Countering Hybrid Threats and NATO Strategic Communications Centre of Excellence. https://stratcomcoe.org/publications/foreign-information-manipulation-and-interference-defence-standards-test-for-rapid-adoption-of-the-common-language-and-framework-disarm-prepared-in-cooperation-with-hybrid-coe/253.

Nimmo B, Hutchins E (2023) Phase-based tactical analysis of online operations. Technical report, Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2023/03/phase-based-tactical-analysis-of-online-operations.

Oxford Internet Institute (n.d.) Computational propaganda. https://www.oii.ox.ac.uk/research/projects/computational-propaganda/. Accessed 28 Aug 2026.

Pamment J (2020) The EU's role in fighting disinformation: Crafting a disinformation framework. Technical report, Carnegie Endowment for International Peace. https://carnegieendowment.org/research/2020/09/the-eus-role-in-fighting-disinformation-crafting-a-disinformation-framework/.

Paul C, Matthews M (2016) The Russian "firehose of falsehood" propaganda model: Why it might work and options to counter it. Perspective PE-198-OSD, RAND Corporation. https://www.rand.org/pubs/perspectives/PE198.html.

Petty RE, Cacioppo JT (1986) The elaboration likelihood model of persuasion. In: Berkowitz L (ed) *Advances in Experimental Social Psychology*, vol 19, pp 123–205. Academic Press, New York. https://doi.org/10.1016/S0065-2601(08)60214-2.

Pijpers PBMJ (2023) *Influence Operations in Cyberspace and the Applicability of International Law*. Elgar International Law and Technology. Edward Elgar Publishing, Cheltenham.

Piskorski J, Stefanovitch N, Da San Martino G et al (2023) SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In: Ojha AK, Doğruöz AS, Madabushi HT et al (eds) *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pp 2343–2361, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.semeval-1.317.

Plank B (2022) The "problem" of human label variation: On ground truth in data, modeling and evaluation. In: Goldberg Y, Kozareva Z, Zhang Y (eds) *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.731.

Plikynas D, Rizgelienė I, Korvel G (2025) Systematic review of fake news, propaganda, and disinformation: Examining authors, content, and social impact through machine learning. *IEEE Access* 13:17583–17629. https://doi.org/10.1109/ACCESS.2025.3530688.

Publications Office of the European Union (n.d.) EuroVoc: The EU's multilingual and multidisciplinary thesaurus. https://eur-lex.europa.eu/browse/eurovoc.html. Accessed 28 Aug 2026.

Sahitaj A, Sahitaj P, Solopova V et al (2025) Hybrid annotation for propaganda detection: Integrating LLM pre-annotations with human intelligence. In: Atwell K, Biester L, Borah A et al (eds) *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)*, pp 215–228, Vienna, Austria. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.nlp4pi-1.18.

Searle JR (1969) *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9781139173438.

Shahi GK (2024) FakeKG: A knowledge graph of fake claims for improving automated fact-checking (student abstract). *Proceedings of the AAAI Conference on Artificial Intelligence* 37(13):16320–16321. https://doi.org/10.1609/aaai.v37i13.27020.

Starbird K, Arif A, Wilson T (2019) Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations.

Proceedings of the ACM on Human-Computer Interaction 3(CSCW):1–26. https://doi.org/10.1145/3359229.

Stocking SH, Holstein LW (2009) Manufacturing doubt: Journalists' roles and the construction of ignorance in a scientific controversy. *Public Understanding of Science* 18(1):23–42. https://doi.org/10.1177/0963662507079373.

Tajfel H, Turner JC (1979) An integrative theory of intergroup conflict. In: Austin WG, Worchel S (eds) *The Social Psychology of Intergroup Relations*, pp 33–47. Brooks/Cole, Monterey, CA.

UKIP (2016) New UKIP poster launching today - Breaking Point. https://x.com/ukip/status/743382813542494212. Accessed 30 Mar 2026.

Wardle C, Derakhshan H (2017) Information disorder: Toward an interdisciplinary framework for research and policy making. Report DGI(2017)09, Council of Europe. https://rm.coe.int/information-disorder-report-2017/1680766412.

Westlake T, Lindelauf R, Čule B et al (2026) Early post-publication behavioural signals of misinformation on social media. *Computers in Human Behavior Reports* 24:101302. https://doi.org/10.1016/j.chbr.2026.101302.

Wimsatt WK, Beardsley MC (1946) The intentional fallacy. *The Sewanee Review* 54(3):468–488.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- appendix.pdf