

The Intent Gap in Disinformation Detection: Evidence from 84 Studies

Xinyu Wang¹, Brett Frischmann², Qianhui Dai¹, Sarah Rajtmajer¹

¹College of Information Sciences and Technology, Pennsylvania State University, USA

²Charles Widger School of Law, Villanova University, USA

xzw5184@psu.edu, brett.frischmann@law.villanova.edu, qpd5030@psu.edu, smr48@psu.edu

Abstract

Disinformation has become a central challenge in the digital information ecosystem, yet the defining property that distinguishes it from misinformation—intent—remains poorly represented in computational detection and moderation research. This paper synthesizes how intent is conceptualized and operationalized across the technical and policy literatures. Through a systematic review of 84 algorithmic papers on disinformation and information/influence operations (IO) detection, we assess how intent is treated in definitions, modeling, data provenance, and labeling protocols. Our analysis reveals four major gaps: (1) conceptual inconsistency; (2) operationalization deficiency; (3) institutional misalignment; and (4) governance frameworks. We propose a roadmap toward intent-aware disinformation detection and moderation that bridges definitional rigor with methodological practice, advancing both theoretical coherence and policy relevance in combating disinformation.

Introduction

The rapid growth of online social networks has amplified the scale and speed of information circulation, leading to widespread information disorder (Vosoughi, Roy, and Aral 2018). In response, a body of research has emerged proposing algorithmic approaches to detect and mitigate the spread of false, misleading, and inauthentic content across platforms. While these efforts broadly address the spread of false content, they often fail to disentangle two distinct phenomena: misinformation and disinformation. The key difference is *intent*—whether the information is shared inadvertently or deliberately (Aïmeur, Amri, and Brassard 2023; Erhardt and Pentland 2022). This distinction is not only of academic interest but has real implications for interventions, in terms of what to do and which actors to focus on. Misinformation typically calls for fact-checking and educational measures (Guess et al. 2020; Orosz et al. 2024; Hoes et al. 2024), whereas disinformation often mandates coordinated and policy-level actions aimed at identifying and disrupting intentional manipulation (Bateman and Jackson 2024; Bellantoni, Alfonsi, and Matasick 2020; Romanishyn, Malyska, and Goncharuk 2025). Misinformation may lead to a focus on intermediaries and consumers (i.e., audience), while

disinformation usually leads to a focus on the suppliers (e.g., speakers, manipulators). Thus, alongside identifying and addressing malicious intent is identifying and dealing with malicious actors and their actions.

Despite the fundamental role it plays, intent remains elusive to most algorithmic detection and moderation systems, at least based on the research we report below. The reason why, we suggest, is that intent is a particularly difficult thing to measure or capture in mere observation and analysis of the disinformation itself. After all, intent is a subjective state of mind attributable to an actual person. In cognitive science, ethics, law, and philosophy, intent is a contested and complicated concept. As Frischmann & Selinger (2018) explain: “Intention is a mental state that is part belief, part desire, and part value. My intention to do something—say to write th[is] explanatory text... or to eat an apple—entails (1) beliefs about the action, (2) desire to act, and (3) some sense of value attributable to the act (p. 364).” In pragmatic ethical and legal contexts, the focus often turns to evidence of intent. For example, a written signature is considered an objective manifestation that a person intends to enter into a contract (Frischmann & Vardi, 2024). In the disinformation context, the challenge is determining which types of evidence can be useful in divining an actor’s subjective state of mind. This challenge is particularly vexing given the scale of the problem in complex information ecosystems.

Our work is motivated by these challenges and focused around the following three research questions:

- **RQ1:** How is disinformation defined and conceptualized across academic, legal, and platform policy contexts, and how consistently is intent recognized as a defining property?
- **RQ2:** To what extent do current detection and moderation methods operationalize intent directly or indirectly?
- **RQ3:** What methodological and governance directions can better integrate intent-awareness into disinformation detection and moderation frameworks?

To address these questions, we systematically examine 84 disinformation-detection papers, platform policies, and governance frameworks. Our review reveals that although intent is central in rhetoric, it is weakly and inconsistently operationalized in practice, and rarely embedded in methodological design. In particular, most computational models

rely on text-based falsity detection, with limited integration of behavioral, contextual, or coordination-based signals that could capture purposeful manipulation. A similar disconnect appears at the institutional level: while platform policies increasingly target coordinated or inauthentic behavior, these enforcement frameworks remain opaque and largely inaccessible to researchers.

These findings raise a deeper question about the feasibility of measuring intent computationally. While metadata, behavioral traces, and coordination patterns can offer partial signals, they remain approximations rather than definitive measures of deliberate manipulation. If algorithmic systems cannot assess intent in a principled or transparent manner, their integration into platform moderation and policy frameworks should be treated with caution. By presenting where and how intent is (and is not) represented across research and policy, this paper reframes intent from an assumed property of disinformation to an empirical construct requiring explicit operationalization. Ultimately, we outline methodological and governance pathways toward intent-aware detection and moderation that balance definitional rigor with computational and ethical constraints.

Conceptual and Operational Context

Misinformation vs. disinformation

Scholars use two main framings to distinguish *misinformation* and *disinformation*. One strand treats them as *non-overlapping* categories separated by intentionality: misinformation refers to false or misleading content shared without an intent to deceive, whereas disinformation requires a deliberate intent to mislead or cause harm (Aïmeur, Amri, and Brassard 2023). A second, now common, strand adopts an *umbrella* view in which misinformation is defined by falsity regardless of intent, and disinformation is the intentional subset under that umbrella. Representative definitions include misinformation as “false or inaccurate information” that may spread with or without intent (Wu et al. 2019), and disinformation as “deliberately propagated false information” (Tucker et al. 2018) or “meant to deceive” (Guess and Lyons 2020).

Beyond falsity, some research emphasizes that disinformation may also involve the manipulation of context, accounts, or networks, rather than the content itself being untrue. For example, disinformation campaigns may involve content that is factually accurate but strategically disseminated through coordinated accounts to mislead or manipulate audiences (Starbird, Arif, and Wilson 2019; Wang et al. 2023). In such cases, the misrepresentation lies not in the factual accuracy of the claim, but in the strategic framing or the actor-network behind it (Erhardt and Pentland 2022). The information involved can itself be accurate, yet function as disinformation when applied or contextualized in ways that deceive, such as when recipients are misled about the identity or intent of the communicator (Lahmann 2020).

Many contemporary frameworks also situate disinformation within the landscape of information or influence operations (IO) (Radsch 2022; Wang et al. 2023; Van Benthem, Dias, and Hollis 2022). IO is defined as *the deployment of*

digital resources for cognitive purposes to change or reinforce attitudes or behaviors of targeted audiences in ways that align with the authors’ interests (Van Benthem, Dias, and Hollis 2022). Recent studies highlight that some disinformation efforts function as coordinated influence operations aimed at shaping opinions and amplifying targeted narratives (Smith et al. 2021; Bradshaw 2020). Importantly, manipulation within IO does not necessarily rely on falsehood, as much of the most effective disinformation consists of true facts arranged in misleading ways (Benkler, Faris, and Roberts 2018).

Disinformation in legal contexts

Intent plays an important role in regulating disinformation in various legal contexts (for brevity, we focus on the US). One prominent example is Section 10(b) of the Securities Exchange Act of 1934. Section 10(b) makes it unlawful to “use or employ, in connection with the purchase or sale of any security” a “manipulative or deceptive device or contrivance in contravention of such rules and regulations as the [SEC] may prescribe.” 15 U.S.C. § 78j(b). The SEC’s implementing regulation, Rule 10b-5, makes it unlawful, in connection with the purchase or sale of any security, to: “- Employ any device, scheme, or artifice to defraud; - Make any untrue statement of a material fact or to omit to state a material fact necessary in order to make the statements made not misleading; or - Engage in any act, practice, or course of business which operates or would operate as a fraud or deceit upon any person.” Notably, a plaintiff must demonstrate that the defendant acted with the relevant state of mind, referred to as *scienter*, which in this case is the intent to deceive, manipulate or defraud. Many judicial decisions and even legislative reform (Private Securities Litigation Reform Act of 1995, Pub. L. 104–67, 109 Stat. 737) have grappled with the difficulties of this evidentiary burden.

A related example is defamation law, which generally provides relief when publication of a false statement harms a person’s reputation and the speaker or publisher is at fault. Ordinarily, fault only requires negligence, but when the defamed person is a public figure, fault requires actual malice, which means “knowledge that it was false or with reckless disregard of whether it was false or not” (New York Times Co. v. Sullivan 1964). While knowledge is different from intent, both are subjective states of mind, proof of which presents similar difficulties.

There are many other laws that regulate disinformation in different contexts, ranging from elections (e.g., 52 U.S.C. §30124 makes it unlawful to fraudulently misrepresent oneself as speaking or acting on behalf of a candidate or political party; 52 U.S.C. §20511(2) punishes anyone who “knowingly and willfully” deprives or defrauds voters of a fair election through certain false practices) to broadcast communications (e.g., 47 C.F.R. §73.1217 prohibits broadcast radio/TV stations from knowingly broadcasting false information about a crime or catastrophe that causes substantial public harm).

In these and other related cases, persistent challenges include: How to determine whether someone had the relevant state of mind? What are reliable types of evidence to sup-

Platform	Misinformation?	Disinformation/IO?	Policy distinguishes intent?	Primary link(s)
Facebook (Meta)	Yes (Community Standards: Misinformation)	Yes (Coordinated Inauthentic Behavior (CIB) / Information Operations)	Yes, via behavior/coordination (CIB focuses on strategic manipulation using fake accounts)	https://transparency.fb.com/policies/community-standards/misinformation/ ; https://about.fb.com/news/2021/08/july-2021-coordinated-inauthentic-behavior-report/ ; https://about.fb.com/news/tag/coordinated-inauthentic-behavior/
Instagram (Meta)	Yes (False Information + fact-checking)	No standalone “disinformation” section; covered by Meta’s CIB/IO framework	Indirect (intent captured through Meta CIB; not in Instagram’s own policies)	https://help.instagram.com/388534952086572 ; https://help.instagram.com/2109682462659451
X (Twitter)	Yes (Civic Integrity: misleading information)	Yes (Platform Integrity & Authenticity includes <i>information operations</i>)	Partial–yes (explicit intent in civic suppression; IO rules target coordinated manipulation)	https://help.x.com/en/rules-and-policies/election-integrity-policy ; https://blog.x.com/en-us/topics/product/2020/updating-our-approach-to-misleading-information(no longer enforced) ; https://help.x.com/en/rules-and-policies#platform-integrity-and-authenticity
YouTube	Yes (Medical Misinformation; Elections Misinformation)	No; handled under “Manipulated content” on misinformation policy page	Generally no (content/harm-based rather than intent-based)	https://support.google.com/youtube/topic/10833358?hl=en&ref_topic=2803176&sjid=14780839018783515554-NA ;
TikTok	Yes (Community Guidelines; newsroom posts on misinformation)	Yes (discuss <i>disinformation</i> and coordinated inauthentic activities)	Partial–yes (rules against coordinated inauthentic activities imply intent to influence)	https://www.tiktok.com/community-guidelines/en/integrity-authenticity?cgversion=2025H2update#1 ; https://newsroom.tiktok.com/it-it/combatting-misinformation-and-election-interference-on-tiktok ; https://www.tiktok.com/transparency/en-us/combatting-misinformation/ ; https://redditinc.com/policies/reddit-rules ; https://redditinc.com/blog/how-reddit-supports-civic-engagement-and-election-integrity-in-2024 ; https://redditinc.com/blog/keeping-our-platform-safe ; https://redditinc.com/policies/transparency
Reddit	Partial–yes (No policy specifically addresses “misinformation”; election-related false info and manipulated/AI content addressed)	Mentioned in official blogs (e.g., AI-generated disinformation)	Partial–yes (focus on content manipulation and coordinated inauthentic behavior in transparency reports; but intent is not explicitly mentioned as a criterion)	
LinkedIn	Yes (False or misleading content; Misinformation and inauthentic behavior)	No; handled as false/misleading content and inauthentic behavior	No	https://www.linkedin.com/help/linkedin/answer/a1425416 ; https://www.linkedin.com/legal/professional-community-policies#be-trustworthy-policy ; https://www.linkedin.com/help/linkedin/answer/a1340752
WhatsApp (Meta)	Yes (User guidance on misinformation)	No	No	https://faq.whatsapp.com/431498999157251 ; https://faq.whatsapp.com/518562649771533

Table 1: Summary of public policy surfaces. Hyperlinks point to the primary public pages used to communicate each policy. *Note on terminology.* Platforms frequently use *information operations* and *influence operations* (IO) as operational labels for campaigns we would normally treat as disinformation; these terms emphasize organizational intent and coordinated deception.

port inferences about subjective states of mind such as intent and knowledge? These issues are pervasive in the law. As Crump (2009) explains, these challenges are complicated by the fact that intent definitions vary across legal contexts (e.g., intent as purpose/desire to perform an act or cause a consequence; intent as knowledge/belief of substantially certain consequence). Direct evidence of a person’s state of mind would include testimony by the person or witnesses who had conversations with the person. Most often, though, evaluating state of mind depends upon circumstantial evidence, which includes a wide array of contextual information that supports inferences about a person’s state of mind. In a securities fraud case, for example, intent is not proven by the mere existence of a false statement of material fact; in the absence of direct testimonial evidence, circumstantial evidence, such as the nature and timing of other actions by the defendant, will be necessary to infer intent.

Turning to the online context and regulation of and by platforms, Section 230 of the Communications Decency Act provides online platforms immunity from lawsuits that would treat them as speakers or publishers of content produced by others (e.g., user-generated content). So, for example, if someone posts a fraudulent stock tip, defamatory statement, or false information about a crime or catastrophe on a social media feed, the social media company is not legally responsible. Yet, for various reasons (beyond the scope of this article), companies often have detailed policies and moderation systems to identify and deal with these and various other forms of disinformation. Consequently, they must grapple with the persistent challenges associated with assessing intent and other states of mind that arise in the legal context.

Disinformation in platform policies and actions

For major social media platforms, Table 1 summarizes: (i) whether they publish policies targeting misinformation; (ii) whether they provide dedicated sections or pages on disinformation/information operations; and (iii) whether their information-disorder policies explicitly distinguish intent. We observe a common pattern—platforms have clear, enforceable misinformation policies (e.g., topic-specific policies on medical or election claims) but reserve “disinformation” for information/influence operations and coordinated inauthentic behaviors. This approach largely bypasses consideration, much less evaluation, of intent at the level of individual users, which facilitates scalable enforcement (for instance, bulk removal of accounts involved in a detected operation) but leaves content-level guidance and enforcement intent-agnostic. In practice, intent is addressed mainly through the manual or automated detection of coordinated network behaviors rather than within content moderation itself. This implies that disinformation detection could and should incorporate *intent signals* (coordination, deception about identity/origin, campaign goals) and not only content truthfulness.

The contrast between policy orientation and content enforcement raises broader questions for detection and moderation research. If intent is expressed and enforced primarily through behavioral coordination, then relying solely on textual or content-based cues may be insufficient to capture disinformation dynamics. Motivated by this observation, we next review existing studies to examine how current detection and moderation systems conceptualize and operationalize intent. Specifically, we want to investigate whether they rely on text alone or integrate contextual and behavioral sig-

nals to infer coordinated or strategic manipulation.

Data Collection

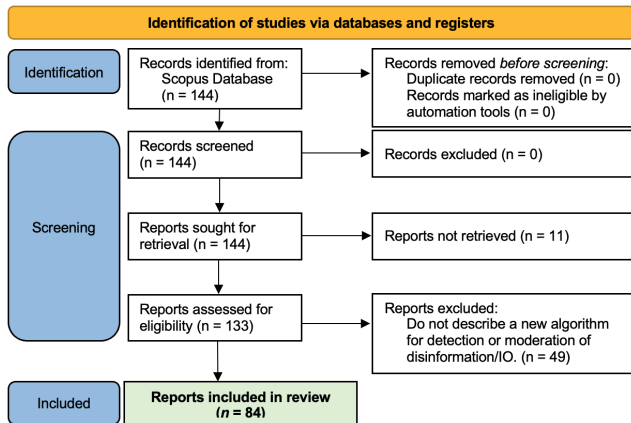


Figure 1: PRISMA Diagram summarizing the identification, screening, and selection of papers for the disinformation and information operations review.

We adopted a precision-oriented retrieval strategy designed to identify studies that explicitly position themselves as disinformation or information/influence operation detection work. Our goal was not to exhaust the broader misinformation or fake-news detection literature, but to assess whether studies that self-identify with the disinformation construct operationalize its intent throughout the research pipeline. We queried Scopus for studies on disinformation detection algorithms and moderation available under search prior to October 2025. To improve precision, we restricted the search to the title field rather than keywords or abstracts, since many papers mention “disinformation” in metadata without making it the primary object of analysis. By limiting results to titles that explicitly include “disinformation,” “information operations,” or “influence operations,” we target works that self-identify with this construct, which fall within the intentional, actor-driven domain of false information propagation, recognizing that we sacrifice recall in service to precision. Our initial search query is provided below, and yields 144 papers in total.

```
TITLE ( ( disinformation OR ( "influence operation" ) OR ( "information operation" ) ) AND ( moderat* OR detect* ) ) AND ( LIMIT-TO ( LANGUAGE , "English" ) )
```

We then conducted a second-level filtering, retaining only those papers for which full text was retrievable¹ and which **describe an algorithm/computational methodology for detection or moderation of disinformation or IO**. Inclusion in the review was based on task framing rather than

¹Retrievable indicates full text is either open access, accessible through our University’s library or otherwise available through online resources.

definitional correctness. Papers were retained if they explicitly presented themselves as disinformation or IO detection/moderation studies and met our computational-method criterion, even when their internal definitions later treated disinformation interchangeably with fake news or misinformation. After filtering, we obtained 84 papers for subsequent analysis. A detailed PRISMA diagram for this process is shown in Figure 1. Two researchers independently analyzed the reviewed papers along four intent-relevant dimensions: disinformation *definition*, intent *operationalization*, *data provenance*, and *labeling protocol*. Inter-coder agreement was assessed using percent agreement, yielding scores of 0.98, 0.95, 0.90, and 0.94 across the four dimensions, respectively (average = 0.9425). Discrepancies were resolved through discussion, with final labels assigned by consensus. The full coding scheme is presented in Table 2.

Results

Disinformation definition

We examine how each paper defines disinformation. The definitional framing is foundational: if a paper’s definition does not encode intent, subsequent choices in data, labeling, and modeling are unlikely to incorporate intent as a construct. We code definitions into two categories:

- Intent–Present (64/84)**: the definition conveys intentionality (e.g., uses “intent,” “intentionally,” “deliberately,” “strategically,” or references IO that presuppose purposeful manipulation) (e.g., (Kramer, Golovchenko, and Hjorth 2025; George 2025)).
- Intent–Agnostic (20/84)**: the term *disinformation* is used interchangeably with “misinformation” or “fake news,” or it is defined solely by falsity without any mention or proxy of intent or purposefully ignores intent (e.g., (Nwaiwu, Jongsawat, and Tungkasthan 2025; Ljubi et al. 2025)).

Intent operationalization

We next examine how each paper operationalizes intent within its methodological design, focusing on the features and proxies used to approximate or infer intentional deception. Across the corpus, we observe substantial variation in how intent is captured, ranging from being embedded in model architecture to being entirely unaddressed. To systematize these differences, we classify all papers into five mutually exclusive categories that reflect how explicitly intent is embedded or inferred: **Inherent/Simulated**, **Coordination/Graph-based**, **Contextual/Metadata/Behavioral**, **Linguistic/Stylistic/Rhetorical**, and **Text-only/No operationalization**. Each paper is assigned to a single category based on the strongest form of operationalization present.

- Inherent/Simulated (14/84)**: Intent is *embedded in the design or simulated through controlled construction of deception*. In this category, intentional manipulation is treated as an intrinsic property of the modeling setup rather than inferred post hoc. Systems in this group

Category	Definition	Example Indicators
Definition	How the paper defines disinformation/IO and, in particular, whether intent is stated as a required property.	Uses terms such as “intentional,” “deliberate,” “strategic”.
Operationalization	How the methodology captures or proxies intent beyond content falsity.	Coordination/network features, contextual features, etc.
Data provenance	What data types and sources enable inference about actors and coordination.	publicly available datasets or self-collected data, etc.
Labeling protocol	How ground truth is constructed and whether labels encode intent cues.	Manual annotation with intent guidelines, expert curation, source-attribution, etc.

Table 2: Coding scheme for intent-aware evaluation of disinformation and information/influence operation (IO) research.

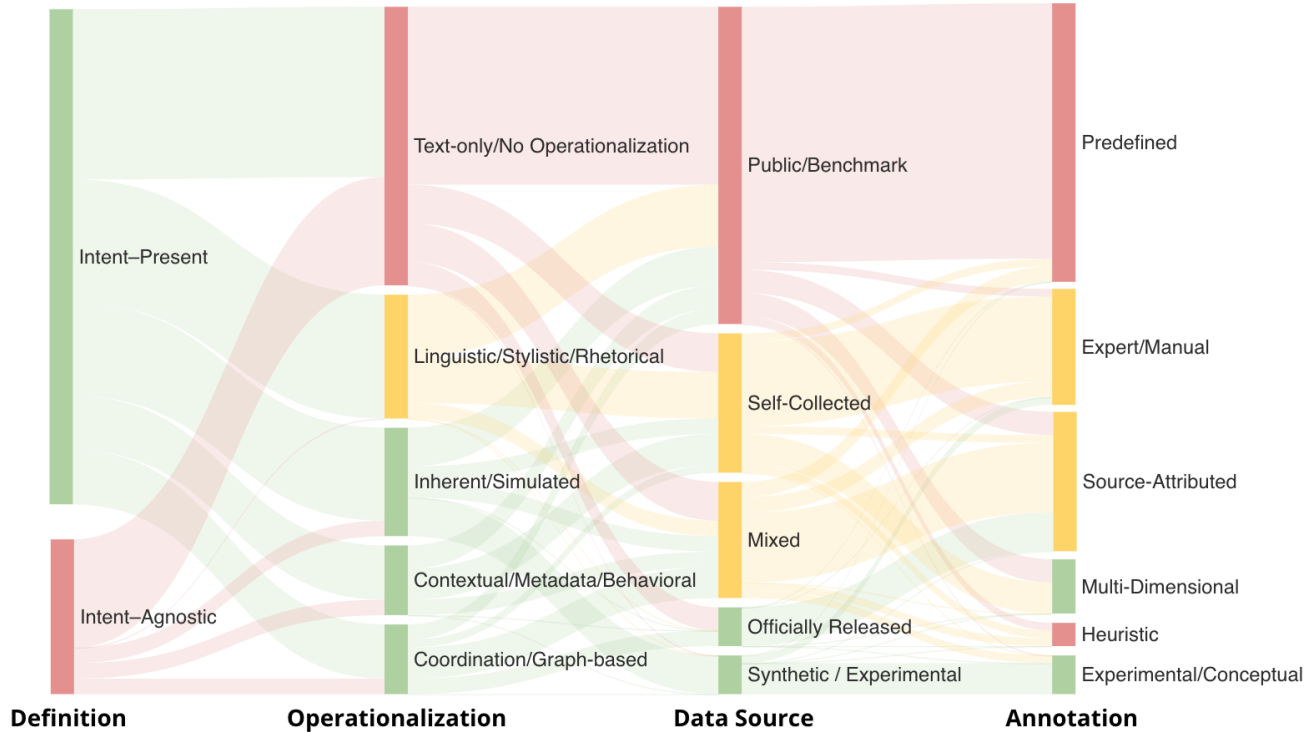


Figure 2: Sankey diagram presenting the co-occurrence of intent representation across four components of the research pipeline. Green: Intent relatively well-captured; Yellow: Intent moderately captured; Red: Intent weakly or not captured.

model deception through architectural or generative assumptions—such as adversarial generation (Yoo et al. 2024; Nathanson et al. 2024), image–text mismatch (Pan et al. 2024), or headline–body mismatch used as deliberate deception logic (Sepúlveda-Torres et al. 2021). These studies operationalize intent by embedding it directly into the data generation or system design process.

- **Coordination/Graph-based (9/84)**: Intent is *inferred from relational and collective behavior*. Models in this category operationalize intent through *network- and graph-based coordination signals*, treating intent as a property of collective rather than individual behavior. Typical features include diffusion structure (Gabriel, Broniatowski, and Johnson 2023), temporal coordination and reaction dynamics (Kong et al. 2023), and cross-network causal patterns (Smith et al. 2021).

- **Contextual/Metadata/Behavioral (9/84)**: Intent is *inferred from contextual, user, or behavioral cues* that extend beyond the textual features. These models leverage auxiliary signals, such as user or account-level metadata (Smith, Ehrett, and Warren 2025), past behavior (Kumar, West, and Leskovec 2016), or multimodal features (Chung, Zhang, and Pan 2023), to approximate the intentional nature of disinformation.
- **Linguistic/Stylistic/Rhetorical (16/84)**: Intent is *approximated through deep textual patterning rather than surface-level veracity or stance cues*. Although these models rely solely on textual inputs, they make deliberate efforts to infer intent such as stylistic cues (Bonet-Jover et al. 2024), subjectivity (Nwaiwu, Jongsawat, and Tungkasthan 2025; Bonet-Jover et al. 2023), and rhetorical persuasion strategies (Sosnowski et al. 2024) that sig-

nal purposeful manipulation.

- **Text-only/No Operationalization (36/84)**: Intent is *not operationalized empirically*. Although many papers adopt intent-aware definitions of disinformation, their modeling objectives reduce to veracity classification (e.g., true vs. false, real vs. fake) using textual embeddings or lexical features (e.g., (Putra, Sibaroni, and Ihsan 2023; Tarczewska, Marciniak, and Giełczyk 2021)). No behavioral, contextual, or stylistic proxies for intent are implemented. These studies therefore reflect a *construct gap* between the conceptual definition of disinformation as intentional and its empirical treatment as mere falsity.

Overall, we find that while intent is frequently referenced conceptually in definitions of disinformation, it is often inconsistently operationalized, or absent entirely in algorithmic detection models.

Data Provenance

We examine where the training datasets for algorithmic detection originate and whether their construction enables inference of intent. We group the papers into five mutually exclusive categories based on dataset sources and their capacity to capture intent: **public/benchmark**, **officially released**, **self-collected**, **synthetic/experimental**, and **mixed**.

- **Public/Benchmark (41/84)**: Open, standardized corpora such as LIAR (Wang 2017), PHEME (Zubiaga et al. 2016), FakeNewsNet (Shu et al. 2020), include binary labels (e.g., “fake” vs. “real”) or fact-checker judgments of the veracity of the text, and in most cases lack actor-level or temporal metadata. Intent is thus only assumed through the labeling framework or generation protocols of the original data source and cannot be empirically examined.
- **Officially Released (5/84)**: Datasets released by social-media companies or monitoring institutions, such as Twitter’s Information Operations archives (no longer available), Reddit suspicious accounts archive², Real411³, and EUvsDisinfo⁴.
- **Self-Collected (18/84)**: Corpora assembled directly by researchers via platform APIs or custom web crawlers, often tailored to specific regional, linguistic, or thematic contexts. These datasets typically allow stronger alignment between labels and the conceptual framing of disinformation, although this alignment is not guaranteed and may not be implemented when intent is not treated as a central construct. Examples include self-collected Twitter datasets (e.g., (Kramer, Golovchenko, and Hjorth 2025; Rastogi and Bansal 2022)) and YouTube API crawls (Bajaj et al. 2016).
- **Synthetic/Experimental (5/84)**: Human- or AI-generated data constructed by researchers to simulate deceptive intent under controlled conditions. Examples include GPT-based generation pipelines (Jiang et al.

2024) and human-constructed disinformation corpus (Williams, Aleroud, and Zimmerman 2023). Because intent is embedded directly in the creation process rather than inherited from an external source, these datasets offer a strong but artificial proxy for deliberate manipulation.

- **Mixed (15/84)**: Datasets combining multiple sources or construction strategies, such as merging fact-checked corpora with self-collected or synthetic data. These hybrid datasets aim to balance external validity with control over intent representation. However, differences in labeling standards and data provenance often introduce inconsistencies, making it difficult to isolate how intent is encoded or inferred across sources.

We find that most studies rely on text-based public corpora originally created for misinformation or fake-news classification rather than intent-focused analysis. Using such datasets for training and testing disinformation detection algorithms assumes that intent was encoded during collection or labeling, yet few papers document how or whether verification has been conducted based on intent. This distinction is crucial for validity: datasets that only encode claim veracity support misinformation detection, whereas datasets that preserve behavioral or contextual evidence allow operationalization of intent.

Labeling Protocol

We analyze how each paper constructs or inherits ground-truth labels and whether these labeling procedures allow intent to be inferred. The construction of labels determines whether a study captures deliberate manipulation or merely content falsity. Across the corpus, most papers adopt predefined binary labels reflecting factual accuracy, while only a few incorporate expert judgment, contextual inference, or design features that encode intentional deception.

- **Predefined (36/84)**: Labels are inherited from existing fact-checking or benchmark datasets. Intent is assumed conceptually through the use of these inherited labels.
- **Source-Attributed (18/84)**: Labels are defined by authoritative entities such as governments or platforms. Intent is implicitly embedded through these sources’ definitions of disinformation as coordinated or state-linked activity.
- **Expert/Manual (15/84)**: Labels are produced by trained annotators or domain experts using interpretive criteria to judge deception or manipulation. Intent is captured implicitly through agreement.
- **Heuristic (3/84)**: Labels are automatically generated through heuristics or weak supervision methods. No human annotation involved.
- **Experimental/Conceptual (5/84)**: Labels are embedded during dataset creation. Examples include experiments in which humans or LLM deliberately craft disinformation text or generate falsified images. Because deception is intentionally produced, intent is captured by construction.

²<https://www.reddit.com/r/reddit.com/wiki/suspiciousaccounts/>

³<https://www.real411.org/>

⁴<https://euvsdisinfo.eu/>

- **Multi-Dimensional (7/84)**: Labels integrate several dimensions in addition to binary labels. These multi-level frameworks allow partial inference of manipulation or persuasion.

The Sankey diagram in Figure 2 visualizes how intent is represented across the four components of the research pipeline—definition, operationalization, data provenance, and labeling protocol. We apply a three-level color scheme to each category: green denotes cases where intent is relatively well-captured, yellow denotes moderate or implicit representation, and red denotes weak or absent representation. As shown in the diagram, only 4.76% of the 84 papers demonstrate strong intent representation across all four components, and these cases occur exclusively in studies where intent is embedded by design through experimental or synthetic data generation. At the opposite end, 15.5% of papers show consistently weak or absent intent representation across all components, reflecting a persistent construct gap in how disinformation is empirically modeled. The broader pattern in the flows shows considerable fragmentation, as many papers recognize intent at the definitional level yet rarely carry this recognition into methodological components, creating a disconnect across the pipeline that sets the stage for potential interventions presented next.

Intervention

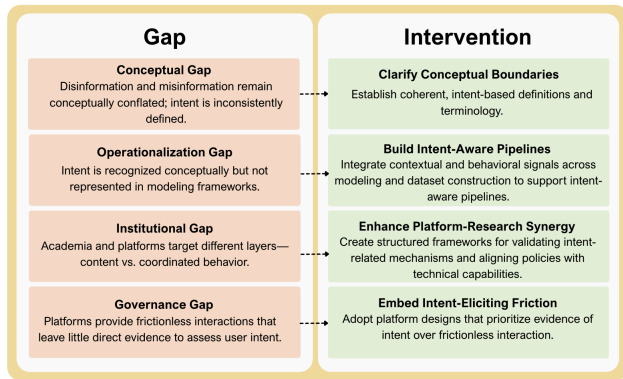


Figure 3: Mapping core gaps in disinformation research to targeted interventions.

Our review reveals that intent embedded in disinformation detection and moderation is constrained by gaps that exist across multiple domains: academic research, platform practice, and legal/policy frameworks. Together, these create a fragmented ecosystem where concept, operationalization, and enforcement remain misaligned. We outline four major categories of challenges and propose corresponding interventions.

Intervention 1: Clarifying conceptual boundaries

As presented above, many studies use the term “disinformation detection” while primarily assessing content veracity or factual correctness without evaluating whether the information was shared deliberately or strategically. This weakens

the construct validity of their claims and contributes to a fragmented conceptual foundation across research communities. The research community should pursue greater definitional precision and transparency. Future studies should explicitly declare their definitional scope and state whether intent is represented in the design, data, or evaluation. Several initiatives already demonstrate the feasibility of structured reporting for conceptual alignment: Datasheets (Gebru et al. 2021) and Model Cards (Mitchell et al. 2019) could be adapted to indicate intent representation and labeling rationale. Initiatives such as The International Fact Checking Network (IFCN) have established standards for source attribution and accountability that can inform dataset and model documentation⁵. Interdisciplinary collaboration among social scientists, computer scientists, and policy researchers would help ensure that definitions of disinformation remain conceptually rigorous and empirically grounded.

Intervention 2: Operationalizing intent through system design

Because disinformation is intent-dependent, models that assess only content falsity are in effect detecting misinformation rather than disinformation. Approaches that proxy intent lack unified validation: intent remains assumed rather than inferred from actor-level or contextual evidence. Methodological innovation can further support a shift beyond falsity-based classification toward intent-aware modeling that integrates textual, behavioral, and contextual information, enabling inferences about strategic coordination and deliberate influence through which intentionality manifests (Wang et al. 2025; Erhardt and Pentland 2022). For instance, graph-based or retrieval-augmented architectures can encode network ties, temporal dependencies, and contextual information (Shi et al. 2024; Lakzaei, Haghiri Chehrehgani, and Bagheri 2024).

Dataset design must support these modeling advances. Building datasets and annotation protocols that encode dimensions of intent will help bridge conceptual clarity and methodological implementation (Wang et al. 2025). Future datasets should integrate multi-layered contextual information, including message provenance, actor roles, and coordination signals (François 2020). Annotation should move beyond binary truth labels toward capturing why and how content circulates, such as identifying initiators, amplifiers, and behavioral cues. Together, these design principles bring contextual modeling and intent-encoded data into alignment, forming a technical foundation for operationalizing intent in disinformation detection systems.

In practice, these components should be implemented as a layered system rather than in parallel. Content-based signals can support fast initial ranking, and behavioral and coordination-based signals can be incorporated in downstream analysis for escalation or enforcement. This design also acknowledges that intent-related signals may arrive later than initial content spread, making them more suitable for downstream intervention and campaign-level analysis to complement real-time detection.

⁵<https://ifcncodeofprinciples.poynter.org/>

Intervention 3: Promoting transparency and collaboration to enhance synergy

Academic research typically focuses on the falsity of content and, when intent is considered, relies heavily on linguistic or textual features. In contrast, platform moderation systems concentrate on coordinated and actor-level manipulation beyond the textual layer. These efforts, however, depend on internally maintained or officially released datasets whose construction, selection, and verification processes are not transparent.

At the same time, access to platform-level operational data is increasingly constrained. Platforms such as “X” have scaled back public data-sharing programs and key initiatives such as the Twitter Moderation Research Consortium (TMRC) were eliminated⁶. Moreover, the X Transparency Center now provides aggregated statistics, with no releases of public datasets⁷. This contraction in data availability not only limits academic validation of real-world practices but also widens the gap between research and platform efforts. As platforms restrict access and internalize detection pipelines, academic studies remain confined to text-based data, modeling intent abstractly and drifting further from the behavioral and network-level dynamics.

To address this fragmentation, future interventions should move beyond case-by-case collaborations toward system-level data governance models that institutionalize transparency as an infrastructural norm. Instead of expecting full data releases, platforms and researchers could co-design embedded validation protocols where model outputs are evaluated against platform-side ground truth on coordinated manipulation without revealing sensitive user-level data. Similar principles have been reflected in IO and industry-academic partnership models that propose structured, privacy-preserving frameworks for researcher access to platform data (Shapiro, Thompson, and Wanless 2020; King and Persily 2020). Establishing such frameworks would allow transparency to function through verified access and reproducible validation rather than public data exposure, aligning institutional incentives while safeguarding privacy and operational integrity.

Intervention 4: Embedding intent-eliciting friction in platform design

Platforms purport to govern disinformation in stated policies and via state-of-the-art detection and moderation systems. The previous interventions we discussed focused on integrating intent into detection and moderation systems. While these interventions could improve those systems, to an extent, we believe a governance gap would remain because in many cases, the evidence needed to evaluate the intent of disinformation suppliers would remain elusive. For the reasons familiar in law and ethics, which we briefly touched on, divining intent generally depends on circumstantial evidence about the specific person in a particular context. Direct evidence, such as an admission, is usually hard to obtain. This

⁶<https://www.cnn.com/2023/02/09/tech/musk-twitter-transparency-researchers>

⁷<https://transparency.x.com/en>

is true for disinformation online too. At least, this is likely the case given the current design of most digital platforms that make speech frictionless.

Yet, digital technologies afford platforms with alternative design options to shape and even govern the speech affordances provided to users, and here we see an opportunity for platforms to think more broadly about the relationships between their core platform services/system design and their content moderation systems (Wang et al. 2025). A rather simple mechanism to generate direct evidence of a person’s intentions is to ask them for testimonial evidence about their intentions when performing the relevant action, for example, when posting or sharing information. Of course, such “friction-in-design” would be overwhelming if implemented for all interactions online (Frischmann and Benesch 2023; Gordon-Tapiero, Ohm, and Ramaswami 2023). Yet we need not go so far. A platform could randomize deployment of the intention prompt or set up triggers, as some platforms have begun to do already. For example, when a person hits the share button without reading an article, some platforms will ask whether the person would prefer to read before sharing (Porterfield 2020). An alternative prompt could ask about who is the person’s intended audience and why the person wanted to share the specific content. Such direct evidence of intent could be quite useful to disinformation detection and moderation systems.

We emphasize that such friction-based mechanisms are not intended to serve as reliable ground-truth signals of intent, nor as a universal component of platform design. In practice, friction is better understood as a limited and targeted design instrument. For example, it can be deployed to complement behavioral and coordination-based signals. When used in this way, friction can contribute to understanding patterns of information sharing and user decision-making without relying on self-reports as standalone evidence for enforcement.

Conclusion

Our systematic review reveals that while the majority of studies conceptually acknowledge intent, only a minority operationalize intent in detection pipelines. Publicly available datasets that are used in these research efforts overwhelmingly capture content falsity rather than deliberate manipulation, while platform-level enforcement, where coordinated manipulations could be captured at scale, remains opaque and inaccessible to researchers. Bridging these gaps calls for both methodological and institutional interventions. We propose actionable pathways and identify the key stakeholders responsible for addressing the multi-layered limitations that currently fragment the field.

Intent is conceptually central to disinformation yet remains difficult to infer empirically. Intent reflects a state of mind that must typically be inferred from indirect evidence, whether behavioral, contextual, or temporal. These signals are often incomplete or ambiguous, and their interpretation depends on assumptions about actors, goals, and environments. Such uncertainty complicates efforts to integrate intent into computational systems, yet it also highlights the importance of treating intent as a meaningful component of dis-

information. As computational models, data infrastructures, and platform systems continue to advance, the opportunities for more coherent alignment expand as well. Stronger modeling frameworks, richer datasets, and improved transparency mechanisms can collectively move the field toward representations of intent that are more reliable and empirically grounded.

Achieving this alignment requires cooperation across the research community, platforms, and policy institutions. Researchers can clarify how intent is represented in definitions, data construction, and modeling choices. Platforms can provide structured and privacy-preserving access to behavioral and coordination signals that are essential for inferring purposeful manipulation. Institutions can establish validation and accountability expectations that promote more consistent incorporation of intent across detection and moderation practices. When these stakeholders work together, the field gains the capacity to move beyond surface-level proxies of falsity and toward systems that capture the strategic and coordinated properties that define disinformation.

Finally, our argument is not that intent should replace veracity, especially in settings where ground truth is incomplete or evolving. Instead, we treat intent as an additional analytical dimension that becomes especially relevant when content-level truth assessment alone is insufficient to characterize strategic manipulation.

Limitation

First, our search strategy likely excludes some relevant adjacent work, especially papers framed primarily as fake-news or misinformation detection that incorporate intent-aware features. We therefore treat the present corpus as a bounded review of self-identified disinformation/IO research rather than an exhaustive map of all computational work relevant to intent.

In addition, although the corpus includes some multimodal work, the field and our analysis remain weighted toward text-centric operationalizations of disinformation. This imbalance may partly reflect the current literature and partly the retrieval strategy used here. As a result, our conclusions are stronger for text-dominant detection pipelines than for rapidly growing multimodal settings.

Ethical Considerations

Inferring intent in disinformation detection is inherently uncertain, as intent is subjective and may be misattributed. Models that approximate intent risk falsely labeling unintentional behavior as malicious, which could lead to potential harms in moderation decisions. At the same time, excluding intent may oversimplify disinformation and create a false sense of consensus about what constitutes harmful information. We therefore argue that intent should be treated as a necessary component in disinformation detection.

Acknowledgment

This work was supported by funding from the National Science Foundation award #2537350.

References

- Aïmeur, E.; Amri, S.; and Brassard, G. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1): 30.
- Bajaj, P.; Kavidayal, M.; Srivastava, P.; Akhtar, M. N.; and Kumaraguru, P. 2016. Disinformation in multimedia annotation: Misleading metadata detection on YouTube. In *Proceedings of the 2016 ACM workshop on Vision and Language Integration Meets Multimedia Fusion*, 53–61.
- Bateman, J.; and Jackson, D. 2024. Countering disinformation effectively: An evidence-based policy guide.
- Bellantoni, A.; Alfonsi, C.; and Matasick, C. 2020. Governance responses to disinformation: How open government principles can inform policy options.
- Benkler, Y.; Faris, R.; and Roberts, H. 2018. *Network propaganda: Manipulation, disinformation, and radicalization in American politics*. Oxford University Press.
- Bonet-Jover, A.; Sepúlveda-Torres, R.; Saquete, E.; and Barco, P. M. 2023. Annotating reliability to enhance disinformation detection: annotation scheme, resource and evaluation. *Procesamiento del Lenguaje Natural*, 70: 15–26.
- Bonet-Jover, A.; Sepúlveda-Torres, R.; Saquete, E.; Martínez-Barco, P.; and Nieto-Pérez, M. 2024. RUN-AS: a novel approach to annotate news reliability for disinformation detection. *Language Resources and Evaluation*, 58(2): 609–639.
- Bradshaw, S. 2020. Influence operations and disinformation on social media. *Modern conflict and artificial intelligence*, 41–47.
- Chung, W.; Zhang, Y.; and Pan, J. 2023. A theory-based deep-learning approach to detecting disinformation in financial social media. *Information Systems Frontiers*, 25(2): 473–492.
- Crump, D. 2009. What does intent mean. *Hofstra L. Rev.*, 38: 1059.
- Erhardt, K.; and Pentland, A. 2022. Disambiguating disinformation: Extending beyond the veracity of online content. *arXiv preprint arXiv:2206.12915*.
- François, C. 2020. Actors, behaviors, content: A disinformation ABC. *Algorithms*.
- Frischmann, B.; and Benesch, S. 2023. Friction-in-design regulation as 21st century time, place, and manner restriction. *Yale JL & Tech.*, 25: 376.
- Gabriel, N. A.; Broniatowski, D. A.; and Johnson, N. F. 2023. Inductive detection of influence operations via graph learning. *Scientific Reports*, 13(1): 22571.
- Gebru, T.; Morgenstern, J.; Vecchione, B.; Vaughan, J. W.; Wallach, H.; Iii, H. D.; and Crawford, K. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12): 86–92.
- George, J. F. 2025. Political affiliation or need for cognition? It depends on the post: Comparing key factors related to detecting health disinformation in the US. *PLoS One*, 20(8): e0315259.
- Gordon-Tapiero, A.; Ohm, P.; and Ramaswami, A. 2023. Fact and Friction: A Case Study in the Fight Against False News. *UC Davis L. Rev.*, 57: 171.

- Guess, A. M.; Lerner, M.; Lyons, B.; Montgomery, J. M.; Nyhan, B.; Reifler, J.; and Sircar, N. 2020. A digital media literacy intervention increases discernment between mainstream and false news in the United States and India. *Proceedings of the National Academy of Sciences*, 117(27): 15536–15545.
- Guess, A. M.; and Lyons, B. A. 2020. Misinformation, disinformation, and online propaganda. *Social media and democracy: The state of the field, prospects for reform*, 10: 10–33.
- Hoes, E.; Aitken, B.; Zhang, J.; Gackowski, T.; and Wojcieszak, M. 2024. Prominent misinformation interventions reduce misperceptions but increase scepticism. *Nature Human Behaviour*, 8(8): 1545–1553.
- Jiang, B.; Tan, Z.; Nirmal, A.; and Liu, H. 2024. Disinformation detection: An evolving challenge in the age of llms. In *Proceedings of the 2024 siam international conference on data mining (sdm)*, 427–435. SIAM.
- King, G.; and Persily, N. 2020. A new model for industry–academic partnerships. *PS: Political Science & Politics*, 53(4): 703–709.
- Kong, Q.; Calderon, P.; Ram, R.; Boichak, O.; and Rizoiu, M.-A. 2023. Interval-censored transformer hawks: Detecting information operations using the reaction of social systems. In *Proceedings of the ACM web conference 2023*, 1813–1821.
- Kramer, M.; Golovchenko, Y.; and Hjorth, F. 2025. Detecting pro-kremlin disinformation using large language models. *Research & Politics*, 12(2): 20531680251351910.
- Kumar, S.; West, R.; and Leskovec, J. 2016. Disinformation on the web: Impact, characteristics, and detection of wikipedia hoaxes. In *Proceedings of the 25th international conference on World Wide Web*, 591–602.
- Lahmann, H. 2020. Information operations and the question of illegitimate interference under international law. *Israel Law Review*, 53(2): 189–224.
- Lakzaei, B.; Haghiri Chehrehgani, M.; and Bagheri, A. 2024. Disinformation detection using graph neural networks: a survey. *Artificial Intelligence Review*, 57(3): 52.
- Ljubi, I.; Grgić, Z.; Vuković, M.; and Gledec, G. 2025. Detecting Disinformation in Croatian Social Media Comments. *Future Internet*, 17(4): 178.
- Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*, 220–229.
- Nathanson, S.; Yoo, Y.; Na, D.; Cao, Y.; and Watkins, L. 2024. A Step Towards Modern Disinformation Detection: Novel Methods for Detecting LLM-Generated Text. In *MILCOM 2024-2024 IEEE Military Communications Conference (MILCOM)*, 615–620. IEEE.
- New York Times Co. v. Sullivan. 1964. 376 U.S. 254.
- Nwaiwu, S.; Jongsawat, N.; and Tungkaathan, A. 2025. Decoding Disinformation: A Feature-Driven Explainable AI Approach to Multi-Domain Fake News Detection. *Applied Sciences*, 15(17): 9498.
- Orosz, G.; Faragó, L.; Paskuj, B.; and Krekó, P. 2024. Strategies to combat misinformation: Enduring effects of a 15-minute online intervention on critical-thinking adolescents. *Computers in Human Behavior*, 159: 108338.
- Pan, Z.; Mao, Y.; Xiong, L.; Pang, T.; and Ping, P. 2024. Mfae: Multimodal fusion and alignment for entity-level disinformation detection. *Pattern Recognition Letters*, 184: 59–65.
- Porterfield, C. 2020. Twitter begins asking users to actually read articles before sharing them. *Forbes*. <https://www.forbes.com/sites/carlieporterfield/2020/06/10/twitter-begins-asking-users-to-actually-read-articles-before-sharing-them>.
- Putra, A. B. Y. A.; Sibaroni, Y.; and Ihsan, A. F. 2023. Disinformation detection on 2024 indonesia presidential election using indobert. In *2023 International Conference on Data Science and Its Applications (ICoDSA)*, 350–355. IEEE.
- Radsch, C. 2022. Ai and disinformation: state-aligned information operations and the distortion of the public sphere. *OSCE Representative on Freedom of the Media, Organization for Security and Co-operation in Europe*.
- Rastogi, S.; and Bansal, D. 2022. Disinformation detection on social media: An integrated approach. *Multimedia Tools and Applications*, 81(28): 40675–40707.
- Romanishyn, A.; Malytska, O.; and Goncharuk, V. 2025. AI-driven disinformation: policy recommendations for democratic resilience. *Frontiers in Artificial Intelligence*, 8: 1569115.
- Sepúlveda-Torres, R.; Vicente, M.; Saquete, E.; Lloret, E.; and Palomar, M. 2021. HeadlineStanceChecker: Exploiting summarization to detect headline disinformation. *Journal of Web Semantics*, 71: 100660.
- Shapiro, J. N.; Thompson, N.; and Wanless, A. 2020. *Research collaboration on influence operations between industry and academia: A way forward*. Carnegie Endowment for International Peace.
- Shi, K.; Sun, X.; Li, Q.; and Xu, G. 2024. Compressing long context for enhancing rag with amr-based concept distillation. *arXiv preprint arXiv:2405.03085*.
- Shu, K.; Mahudeswaran, D.; Wang, S.; Lee, D.; and Liu, H. 2020. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. *Big data*, 8(3): 171–188.
- Smith, D. H.; Ehrett, C.; and Warren, P. 2025. Unsupervised detection of coordinated information operations in the wild. *EPJ Data Science*, 14(1): 26.
- Smith, S. T.; Kao, E. K.; Mackin, E. D.; Shah, D. C.; Simek, O.; and Rubin, D. B. 2021. Automatic detection of influential actors in disinformation networks. *Proceedings of the National Academy of Sciences*, 118(4): e2011216118.
- Sosnowski, W.; Modzelewski, A.; Skorupska, K.; Otterbacher, J.; and Wierzbicki, A. 2024. Eu disinfotest: a benchmark for evaluating language models’ ability to detect disinformation narratives. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 14702–14723.

- Starbird, K.; Arif, A.; and 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.
- Tarczewska, M.; Marciniak, A.; and Gielczyk, A. 2021. Fake or Real? The Novel Approach to Detecting Online Disinformation Based on Multi ML Classifiers. In *International Conference on Computational Science*, 18–27. Springer.
- Tucker, J. A.; Guess, A.; Barberá, P.; Vaccari, C.; Siegel, A.; Sanovich, S.; Stukal, D.; and Nyhan, B. 2018. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*.
- Van Benthem, T.; Dias, T.; and Hollis, D. B. 2022. Information Operations under International Law. *Vand. J. Transnat'l L.*, 55: 1217.
- Vosoughi, S.; Roy, D.; and Aral, S. 2018. The spread of true and false news online. *science*, 359(6380): 1146–1151.
- Wang, W. Y. 2017. ”liar, liar pants on fire”: A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.
- Wang, X.; Koneru, S.; Venkit, P. N.; Frischmann, B.; and Rajtmajer, S. 2025. The unappreciated role of intent in algorithmic moderation of abusive content on social media. *Harvard Kennedy School Misinformation Review*.
- Wang, X.; Li, J.; Srivatsavaya, E.; and Rajtmajer, S. 2023. Evidence of inter-state coordination amongst state-backed information operations. *Scientific reports*, 13(1): 7716.
- Williams, J. A.; Aleroud, A.; and Zimmerman, D. 2023. Detecting science-based health disinformation: a stylometric machine learning approach. *Journal of Computational Social Science*, 6(2): 817–843.
- Wu, L.; Morstatter, F.; Carley, K. M.; and Liu, H. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD explorations newsletter*, 21(2): 80–90.
- Yoo, Y.; Na, D.; Nathanson, S.; Cao, Y.; and Watkins, L. 2024. Disinformation at Scale: Detecting AI-Human Composite Images via Convolution Ensembles. In *MILCOM 2024-2024 IEEE Military Communications Conference (MILCOM)*, 621–626. IEEE.
- Zubiaga, A.; Liakata, M.; Procter, R.; Wong Sak Hoi, G.; and Tolmie, P. 2016. Analysing how people orient to and spread rumours in social media by looking at conversational threads. *PloS one*, 11(3): e0150989.