Media Bias Within Information Disorder: Bridging Two Research Communities Through a Systematic Review

Francisco-Javier Rodrigo-Ginés[*], **Jorge Chamorro-Padial**[†]
[*]Universidad Nacional de Educación a Distancia — NLP & IR Group, Madrid, Spain
[†]OpenCodice Research, Spain
frodrigo@invi.uned.es, jorge@opencodice.com

Abstract

Information disorder research overwhelmingly focuses on fabricated or manipulated content (fake news, deepfakes, propaganda) while comparatively neglecting the most pervasive form of distorted information: media bias. Unlike outright falsehoods, media bias operates within the boundaries of factual reporting, distorting public understanding through framing, omission, and word choice rather than fabrication. This makes it harder to detect, harder to regulate, and paradoxically more influential, since it originates from trusted mainstream sources rather than marginal actors. In this position paper, we argue that media bias should be recognized as a first-class category within information disorder frameworks. Drawing on the Wardle and Derakhshan (2017) taxonomy, communication theory, and a systematic review of over 100 studies on automated media bias detection, we demonstrate that current frameworks inadequately account for the systematic distortion of *true* content. We present a consolidated taxonomy of media bias types organized by linguistic level, compare detection paradigms across the information disorder and media bias communities, and identify four properties that make media bias uniquely dangerous: its scale, its source credibility, the invisibility of omission, and its cumulative normative effect. We conclude with an integrated research agenda grounded in specific gaps identified through the review.

Keywords: information disorder, media bias, framing, omission bias, position paper, systematic review, NLP

1. Introduction

When researchers and policymakers speak of "information disorder", they almost invariably mean content that is *false*: fabricated news articles, doctored images, conspiracy theories shared on social media (Lazer et al., 2018; Vosoughi et al., 2018). This conflation of information disorder with falsehood has shaped research agendas, detection tools, and regulatory frameworks alike. Fact-checking initiatives, automated fake news classifiers, and platform content moderation policies all target the same object: content that deviates from factual reality.

Yet the most widespread and arguably most influential form of distorted information is not false at all. **Media bias**, the systematic tendency of news outlets to present information in ways that favour particular perspectives, interpretations, or actors (Hamborg et al., 2019), operates entirely within the boundaries of factual reporting. A biased article need not contain a single false claim. Instead, it shapes understanding through what it emphasizes and what it omits, through the words it chooses and the frames it constructs (Entman, 1993). This concern extends to modern language technologies: large language models (LLMs), trained on vast corpora of news text, risk inheriting and amplifying the very biases present in their training data (Gallegos et al., 2024; Park and Kim, 2025), making the study of media bias urgent not only for understanding journalism but also for building trustworthy AI systems.

In this position paper, we argue that media bias is an underrecognized and inadequately addressed form of information disorder. Drawing on the influential framework of Wardle and Derakhshan (2017), on communication theory, and on a systematic review of over 100 studies on automated media bias detection (Hamborg et al., 2019; Rodrigo-Ginés et al., 2024), we make three claims, each grounded in evidence from the review:

1. Media bias is a form of information disorder that current frameworks acknowledge in principle but marginalize in practice.
2. The properties that distinguish media bias from other forms of information disorder (its scale, its origin in trusted sources, the invisibility of omission, and its cumulative normative effect) make it *more* dangerous than fabricated content, not less.
3. The NLP communities working on information disorder (fake news, propaganda, fact-checking) and media bias detection operate largely in isolation, to the detriment of both, a disconnect clearly visible in the citation patterns and methodological choices documented in the review.

We conclude by proposing an integrated research agenda that positions media bias at the

Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026, pages 45–54

centre of information disorder research, with specific calls for multilingual resources, perspectivist annotation frameworks, and evaluation paradigms that move beyond binary detection.

2. The Information Disorder Landscape

The term "information disorder" was popularized by Wardle and Derakhshan (2017) as an alternative to the imprecise and politically weaponized label "fake news". Their framework distinguishes three categories along two dimensions, *intent to harm* and *falsity*:

- **Misinformation**: false content shared without intent to harm (e.g., honest mistakes, misunderstood statistics).
- **Disinformation**: false content deliberately created and shared to cause harm (e.g., fabricated news, coordinated campaigns).
- **Malinformation**: genuine content shared with intent to harm (e.g., leaks, harassment, out-of-context sharing).

Within this tripartite model, Wardle and Derakhshan identify seven types of problematic content, ranging from satire and parody (lowest harm) to fabricated content (highest harm). Type 3, "misleading content: misleading use of information to frame an issue or an individual", is the category most directly relevant to media bias.

The framework's blind spot. While the Wardle–Derakhshan taxonomy is rightly influential, it carries an implicit assumption: problematic content is *exceptional*. The seven types describe deviations from normal journalistic practice: satire that is mistaken for news, content that is improperly attributed, narratives that are fabricated. Media bias, however, is not exceptional. It is the *norm*. As communication scholars have long documented, all news production involves selection, emphasis, and framing; neutrality is an aspirational ideal rather than a default state (McQuail, 2010; Entman, 1993). When bias is the baseline rather than the anomaly, a framework designed to identify departures from good journalism fundamentally struggles to account for it.

Where media bias fits, and does not fit. Media bias maps most closely to the "misleading content" category, but this mapping is imperfect. Misleading content in the original framework implies a deliberate or at least identifiable act of misrepresentation: a photograph cropped to change its meaning, a headline that mischaracterizes the article body. Much media bias, by contrast, is structural and often unconscious: editorial choices about which stories to cover, which sources to quote, and which perspectives to include reflect institutional norms, audience expectations, and market incentives rather than individual acts of deception (Groeling, 2013; Shoemaker and Vos, 2009). The framework's reliance on intent (the mis/dis/mal distinction) maps poorly onto a phenomenon where intent is diffuse, institutional, and frequently invisible even to the journalists producing the content (Eberl et al., 2017).

Moreover, the "misleading content" category is one of seven, situated between "false connection" (type 2) and "imposter content" (type 4). This positioning implies rough equivalence in scope and importance. In reality, media bias dwarfs all other categories in volume: every news article carries framing choices, while fabricated content, propaganda, and imposter content are comparatively rare events in the broader information ecosystem (Vosoughi et al., 2018).

3. Media Bias: A Taxonomy from the Literature

To understand why media bias warrants special attention within information disorder frameworks, we must first clarify what media bias *is*. The term is used loosely in public discourse; in research, it encompasses a structured set of phenomena operating at multiple linguistic levels (Hamborg et al., 2019; Rodrigo-Ginés et al., 2024).

The underlying systematic review. The taxonomy and evidence presented throughout this paper draw on a systematic review of automated media bias detection (Rodrigo-Ginés et al., 2024) that followed the PRISMA guidelines. The review queried Scopus, Web of Science, and IEEE Xplore using search terms combining "media bias" with computational detection methods, covering the period 2000–2023. After applying inclusion criteria (peer-reviewed, English-language studies addressing automated detection or characterization of media bias in news text), the final corpus comprised over 100 studies. Each study was coded for bias type addressed, detection method, dataset used, language, and evaluation approach. We refer the reader to the original review for the full protocol and detailed results; here we synthesize the findings most relevant to positioning media bias within information disorder.

Drawing on this review, we present a consolidated taxonomy that organizes bias types by the linguistic level at which they manifest. Table 1 summarizes this taxonomy and maps each type to its closest parallel in information disorder research.

Table 1: A taxonomy of media bias types organized by linguistic level, drawn from a systematic review of automated media bias detection (Rodrigo-Ginés et al., 2024). The rightmost column maps each type to its closest parallel in information disorder research, revealing that many bias types lack a direct counterpart.

Level	Bias Type	Description	Info. Disorder Parallel
Word / Token	Word choice / Labelling	Evaluatively loaded terms ("regime" vs. "government")	Propaganda: loaded language
	Subjective intensifiers	Opinion-injecting adjectives and adverbs	Propaganda: exaggeration
	Attribution bias	Selective quoting, misattribution of statements	Imposter content (partial)
	Sensationalism	Dramatization, hyperbolic language	Clickbait, false connection
Sentence	Opinions as facts	Subjective judgments presented as established truth	Misleading content
	Mind reading	Claiming knowledge of actors' thoughts without evidence	Fabricated content (partial)
	Source selection bias	Quoting only sources aligned with a viewpoint	*No direct parallel*
Article / Discourse	Framing bias	Selective emphasis of particular aspects of an event	Misleading content
	Omission bias	Systematic exclusion of relevant facts or perspectives	*No direct parallel*
	Spin	Interpreting events with a particular evaluative slant	Misleading content
Outlet / Corpus	Gatekeeping	Editorial decisions about which stories to cover	*No direct parallel*
	Coverage bias	Disproportionate attention to certain topics or actors	*No direct parallel*

Word- and sentence-level bias. The most frequently studied bias types operate at the surface level of text. Recasens et al. (2013) demonstrated that even single-word substitutions ("claimed" vs. "stated", "regime" vs. "government") can measurably shift reader perception. At the sentence level, presenting opinions as facts or attributing unverified mental states to public figures ("mind reading") introduces bias without any overtly evaluative language. These types are the most accessible to current NLP methods, as they leave explicit textual traces that can be detected through word-level or span-level analysis (Spinde et al., 2021a,b). Importantly, word-level bias types have clear parallels in propaganda detection: loaded language, exaggeration, and name-calling are among the persuasion techniques catalogued by Da San Martino et al. (2019). Yet, as we discuss in Section 6, these overlapping phenomena are studied by separate communities using separate datasets and separate terminology.

Article- and outlet-level bias. At higher linguistic levels, bias becomes progressively harder to detect and progressively more consequential. Framing bias, the selection and emphasis of particular aspects of a story while downplaying others (Entman, 1993), requires understanding how a full article structures its narrative. The same event (a protest, a policy change, an economic report) can be framed as a triumph or a failure, a step forward or a crisis, depending on which facts are foregrounded. Framing bias has received growing computational attention, notably through SemEval-2023 Task 3 (Piskorski et al., 2023) and dedicated media frames corpora (Card et al., 2015). Omission bias, what is *not* said, is structurally invisible: there is no textual trace to detect because the biasing element is the missing text (Puglisi and Snyder Jr., 2015). At the outlet level, gatekeeping and coverage bias reflect the editorial decision of what constitutes "news", itself an act of selection and omission (Shoemaker and Vos, 2009; McCombs and Shaw, 1972). The systematic review found that the vast majority of detection methods target word- or sentence-level bias, with article-level and outlet-level methods representing a small and growing but still underdeveloped area of research (Rodrigo-Ginés et al., 2024).

Key observations from the taxonomy. Three findings from this mapping are particularly relevant to our argument. First, the bias types most studied in NLP (word choice and framing) have partial parallels in propaganda detection (loaded language, persuasion techniques), yet the two communities

rarely share methods or datasets. Second, several consequential bias types (omission, gatekeeping, coverage bias, source selection) have *no direct parallel* in information disorder taxonomies. These are precisely the types that operate through absence rather than presence, making them invisible to verification-based approaches. Third, the taxonomy reveals a *hierarchy of subtlety*: word-level bias is the most detectable and the most studied, while outlet-level bias is the most consequential but the least amenable to current NLP methods (Baly et al., 2020).

The perspectivist nature of bias. A crucial complication, amply documented in the review, is that bias perception is inherently subjective. The same article may be judged as "balanced" by one reader and "biased" by another, depending on their prior beliefs, political identity, and expectations of journalistic norms (Eberl et al., 2017). This subjectivity has important methodological implications: datasets annotated by majority vote may systematically suppress minority perspectives, and evaluation metrics that treat bias as a binary ground truth may misrepresent the phenomenon (Basile et al., 2021; Cabitza et al., 2023). The systematic review found that the majority of existing datasets rely on aggregated labels, with perspectivist annotation remaining the exception rather than the norm (Rodrigo-Ginés et al., 2024).

4. Detection Methods: Two Parallel Tracks

The systematic review reveals a striking pattern: the information disorder and media bias communities have developed parallel but largely independent detection ecosystems. This section compares these two tracks, drawing on the method and dataset landscape documented in the review (Rodrigo-Ginés et al., 2024).

Information disorder detection. The computational study of information disorder has developed mature pipelines for fake news detection (Lazer et al., 2018; Zannettou et al., 2019), rumour verification, automated fact-checking (Thorne et al., 2018; Augenstein et al., 2019; Nakov et al., 2021), and propaganda detection (Da San Martino et al., 2019; Dimitrov et al., 2021). These tasks share a common assumption: there exists a ground truth (a claim is true or false, a source is reliable or unreliable) against which system output can be evaluated. The dominant paradigm is *verification*, that is, checking content against reality. Detection methods rely heavily on external knowledge sources (knowledge graphs, fact-checking databases) and on network-level signals (propagation patterns, source credibility scores) (Baly et al., 2020). Shared tasks have produced large-scale benchmarks: FEVER (Thorne et al., 2018) for fact verification, SemEval-2020 Task 11 and SemEval-2021 Task 6 for propaganda detection (Dimitrov et al., 2021), and numerous fake news detection corpora. Evaluation is typically based on standard classification metrics (accuracy, precision, recall, F1) applied to binary or multi-class labels with clear ground truth.

Media bias detection. The computational study of media bias, as documented in the systematic review, follows a different trajectory. Methods focus on bias detection at the lexical, sentence, and article levels (Hamborg et al., 2019; Recasens et al., 2013; Fan et al., 2019; Gangula et al., 2019), framing analysis (Card et al., 2015; Piskorski et al., 2023), and increasingly, the use of large language models as bias detectors (Wen and Younes, 2024; Maab et al., 2024; Lin et al., 2024; Trhlík and Stenetorp, 2024). This community operates with a different assumption: bias is not a binary factual property but a continuous, perspective-dependent characteristic. The dominant paradigm is *interpretation*, that is, analyzing how content is constructed rather than whether it is true. Detection relies primarily on textual features extracted from the content itself, with comparatively less use of external knowledge or network signals. The review documented a methodological evolution from early lexicon-based and feature-engineering approaches to transformer-based classifiers that achieve state-of-the-art performance on sentence-level bias detection (Rodrigo-Ginés et al., 2024).

Shared techniques, separate ecosystems. Despite these different paradigms, the two communities employ remarkably similar technical machinery: transformer-based classifiers, attention mechanisms for span-level annotation, and multi-task learning architectures. SemEval-2023 Task 3 (Piskorski et al., 2023), which addressed news genre, framing, and persuasion detection in a multilingual setup, represents a rare point of convergence, but remains the exception rather than the rule. The review found that media bias detection papers rarely cite propaganda detection work, and vice versa, even when they address overlapping phenomena such as loaded language or persuasive framing (Rodrigo-Ginés et al., 2024). This isolation extends to the use of LLMs: recent work on using GPT-based models for bias detection (Wen and Younes, 2024; Maab et al., 2024) and for propaganda detection employs similar prompting strategies but develops them independently.

The dataset landscape. The dataset gap is equally revealing. The systematic review identified a pronounced concentration of resources: the vast majority of media bias datasets are in English, with very few resources available for other languages (Hamborg et al., 2019; Fan et al., 2019; Spinde et al., 2021a; Lim et al., 2020). Key English datasets, including MBIC (Spinde et al., 2021a), BASIL (Fan et al., 2019), and BABE (Spinde et al., 2021b), provide word-level and sentence-level annotations but differ substantially in their annotation schemes, bias definitions, and granularity. MBIC includes annotator characteristics (demographics, political leaning), enabling analysis of how background affects bias perception, while BABE uses expert annotators for higher inter-annotator agreement. This fragmentation makes cross-dataset evaluation difficult and limits the generalizability of detection methods. By contrast, the information disorder community benefits from large-scale, standardized benchmarks (FEVER alone contains over 185,000 claims) that enable direct comparison across methods and have driven rapid methodological progress. The absence of comparable standardized resources for media bias detection is, in our view, one of the most significant practical consequences of the community's relative isolation.

5. Why Media Bias is Uniquely Dangerous

We identify four properties that distinguish media bias from other forms of information disorder and that, taken together, make it a uniquely powerful force for shaping public understanding. These properties are not merely theoretical assertions; each is supported by empirical evidence. Empirical studies have shown that biased mainstream media measurably shifts voting behaviour (DellaVigna and Kaplan, 2007) and drives political polarization (Martin and Yurukoglu, 2017), while exposure to fabricated news is far more limited than commonly assumed (Allcott and Gentzkow, 2017).

Scale. Fabricated content, while attention-grabbing, represents a small fraction of the information ecosystem. Vosoughi et al. (2018) found that false news stories on Twitter were shared by far fewer users than true stories, despite spreading faster within networks. Media bias, by contrast, is omnipresent: every article published by every news outlet carries framing choices, emphasis decisions, and omissions. The cumulative volume of biased-but-factual content vastly exceeds the volume of outright fabrication. If information disorder is defined by its capacity to distort public understanding, then the aggregate effect of daily biased reporting, across thousands of outlets, millions of articles, and billions of reader impressions, is likely far greater than that of viral fake news.

Source credibility. Disinformation typically originates from anonymous accounts, fringe websites, or state-backed troll operations, that is, sources that audiences have learned to distrust, at least in principle (Lazer et al., 2018). Media bias, however, is produced by the most trusted institutions in the information ecosystem: established newspapers, public broadcasters, and major digital news platforms. This is the paradox at the core of our argument: the very credibility that makes mainstream media valuable as information sources also makes their biases more influential. Readers apply less critical scrutiny to content from trusted sources, accepting framing choices and omissions as part of "the way things are" rather than as editorial decisions that could have been made differently (Ecker et al., 2022).

The invisibility of omission. Fact-checkers can identify false claims. Propaganda detection tools can flag loaded language and rhetorical manipulation (Da San Martino et al., 2019). Deepfake detectors can analyze visual artifacts (Vaccari and Chadwick, 2020). But no tool can flag what is *not there*. Omission bias, the systematic exclusion of perspectives, sources, or facts, leaves no trace in the published text. A reader cannot know what they were not told, and an automated system cannot detect the absence of content it was never given. As our taxonomy shows (Table 1), omission and gatekeeping bias have no parallel in information disorder frameworks precisely because those frameworks are built on the assumption that disordered content *exists* in a detectable form. This makes omission bias the most durable and least accountable form of information distortion: it cannot be "fact-checked" because no individual claim is false, and it cannot be detected by systems trained on textual features because the relevant signal is extratextual.

Cumulative normativity. Each individual biased article may appear unremarkable. The danger lies in accumulation: when audiences are consistently exposed to news framed from a particular perspective, that perspective becomes the perceived default, the "normal" way of understanding an issue (Scheufele, 1999; McCombs and Shaw, 1972). This normative effect is qualitatively different from the acute shock of encountering a false claim. False claims can be refuted; biased framing, absorbed over years of media consumption, reshapes the cognitive frameworks through

which audiences interpret all subsequent information. The systematic review found evidence of this asymmetry in computational terms as well: while fake news detection has converged on increasingly effective methods, media bias detection remains a significantly harder task, in part because the target itself (what counts as "biased") shifts with audience and context (Rodrigo-Ginés et al., 2024).

6. The Disconnect Between Research Communities

Despite the conceptual proximity of media bias and information disorder, the NLP communities working on these problems are remarkably disconnected, a pattern the systematic review makes quantitatively visible (Rodrigo-Ginés et al., 2024).

Divergent citation networks. Shared tasks treat fake news detection (SemEval-2019 Task 7), propaganda detection (SemEval-2020 Task 11, SemEval-2021 Task 6), and news framing (SemEval-2023 Task 3) as separate problems with separate datasets and separate evaluation metrics. Survey papers on information disorder rarely cite the media bias literature in depth, and vice versa. The Wardle–Derakhshan framework (Wardle and Derakhshan, 2017), widely cited in information disorder research, is rarely referenced in NLP papers on media bias detection, even when the detected bias types map directly to the framework's categories. Conversely, foundational media bias work such as Hamborg et al. (2019) and Recasens et al. (2013) is seldom cited in propaganda or fact-checking papers, despite addressing overlapping linguistic phenomena.

Missed methodological synergies. This fragmentation has practical consequences documented in the review. Methods developed for propaganda detection (e.g., persuasion technique classifiers) could directly inform framing bias detection, since persuasion and framing share rhetorical mechanisms: both involve selecting which aspects of reality to make salient and which to suppress. Perspectivist annotation methods developed in the media bias community (Basile et al., 2021; Cabitza et al., 2023) could improve the handling of subjectivity in fact-checking and credibility assessment, where annotator disagreement is typically treated as noise rather than signal. Source-level analysis (Baly et al., 2020) could provide contextual features for article-level bias detection, since the political orientation and editorial line of a news outlet constitute strong priors for the bias expected in individual articles. The LLM-based methods now being explored for bias detection (Wen and Younes, 2024; Maab et al., 2024) use prompting strategies remarkably similar to those employed for propaganda detection, yet the two lines of work develop these strategies independently.

The cost of disconnection. These synergies remain largely unexploited. The review identified only a handful of studies that explicitly bridge the two communities, most notably SemEval-2023 Task 3 (Piskorski et al., 2023), which combined genre classification, framing detection, and persuasion technique identification in a single multilingual shared task. The rarity of such bridging efforts suggests that the disconnect is not merely a citation gap but reflects deeper differences in how the two communities conceptualize their objects of study: verification vs. interpretation, truth vs. perspective, binary vs. spectral evaluation. Overcoming this divide requires more than cross-referencing papers; it requires rethinking shared tasks, evaluation metrics, and even the definition of what constitutes "disordered" information.

7. Toward an Integrated Research Agenda

We propose five directions for integrating media bias into information disorder research, each motivated by a specific gap identified through the systematic review:

1. Extend information disorder frameworks. The Wardle–Derakhshan taxonomy should be revised to include media bias as a primary category, not a subcategory of "misleading content". Our taxonomy (Table 1) demonstrates that at least four major bias types (omission, gatekeeping, coverage bias, and source selection) have no parallel in current information disorder classifications. We propose distinguishing between *content-level disorder* (fabrication, manipulation, impersonation) and *framing-level disorder* (bias, selective emphasis, omission), with explicit recognition that the latter operates on true content and requires different detection paradigms. This distinction is not merely taxonomic: it implies different annotation guidelines, different evaluation criteria, and different intervention strategies.

2. Develop multilingual media bias resources. The systematic review confirmed a severe English-language concentration: the vast majority of media bias datasets and detection methods are developed for English (Hamborg et al., 2019; Fan et al., 2019; Spinde et al., 2021a). Information disorder, however, is a global phenomenon, and media bias

manifests differently across linguistic and cultural contexts (Piskorski et al., 2023). SemEval-2023 Task 3 demonstrated both the feasibility and the value of multilingual framing analysis, but comparable resources for media bias detection beyond English remain scarce. We call for the creation of media bias resources in underrepresented languages, with annotation schemes sensitive to local journalistic norms and political contexts. Such resources would not only expand the geographic coverage of media bias research but also reveal culturally specific bias patterns that monolingual studies cannot capture.

3. Embrace perspectivist annotation. The review found that the majority of media bias datasets resolve annotator disagreement through majority vote, effectively treating bias as a binary factual property (Rodrigo-Ginés et al., 2024). This contradicts the inherently subjective nature of bias perception (Eberl et al., 2017). Rather than resolving disagreement through aggregation, annotation frameworks should preserve individual judgments, enabling models to predict distributions of opinion rather than single labels (Basile et al., 2021; Cabitza et al., 2023). This perspectivist approach is especially important for media bias, where the "ground truth" is not a fact to be verified but a judgment to be understood. Datasets such as MBIC (Spinde et al., 2021a), which record annotator characteristics alongside annotations, point toward this direction but remain exceptions in the field.

4. Move beyond binary detection. The review documented that current media bias detection is dominated by binary classification (biased vs. unbiased), with few systems modelling bias as a spectrum (Rodrigo-Ginés et al., 2024). A more nuanced approach would characterize the *type* and *direction* of bias simultaneously, following the multi-level taxonomy presented in Table 1. Each level of the taxonomy calls for different evaluative frameworks: word-level bias lends itself to span-extraction and token-classification methods akin to named entity recognition; sentence-level bias requires document-contextual classification; article-level framing demands discourse-aware models that capture narrative structure; and outlet-level bias necessitates corpus-comparative approaches that contrast coverage patterns across sources. Evaluation metrics should reflect this complexity, moving from accuracy and F1 toward measures that capture calibration, ranking quality, and agreement with diverse annotator populations. The information disorder community's experience with multi-label propaganda detection (Da San Martino et al., 2019), where a single text span may exhibit multiple persuasion techniques, offers a useful model for multi-type bias annotation.

5. Interrogate LLMs as both tools and vectors. Large language models present a dual challenge for media bias research. As detection tools, they show promising capabilities for identifying bias (Wen and Younes, 2024; Maab et al., 2024; Lin et al., 2024), but they also carry their own biases, shaped by training data that inevitably reflects the biases of the media it was drawn from (Gallegos et al., 2024; Park and Kim, 2025). An LLM trained on biased news may reproduce and even amplify the biases present in its training corpus. The review noted a rapid growth in LLM-based media bias studies, but these rarely address whether LLMs' own biases compromise their utility as detectors (Rodrigo-Ginés et al., 2024). Research on LLMs and media bias must address both directions: using LLMs to detect bias in content, and detecting bias in the LLMs themselves. The fact that these models are now being integrated into newsroom workflows for summarization, translation, and even content generation makes this question urgent.

8. Conclusion

Information disorder is not only about what is false. It is also, and perhaps primarily, about what is *true but distorted*: factual content presented through selective framing, loaded language, and strategic omission. Media bias is this distortion operating at scale, produced by the most trusted institutions in the information ecosystem, and rendered invisible by the very ordinariness of its mechanisms.

Current information disorder frameworks acknowledge the existence of misleading content but treat it as one category among many, equivalent in scope to satire or imposter content. We have argued that this is a fundamental mischaracterization, grounding our claims in a systematic review of over 100 studies that reveals both the breadth of media bias phenomena (Table 1) and the depth of the disconnect between the communities that study information disorder and media bias.

Media bias is not a peripheral form of information disorder; it is the most pervasive, the most credible, and in many ways the most dangerous, precisely because it leaves no false claims to fact-check and no fabricated content to debunk. Recognizing media bias as a first-class category of information disorder is not merely a taxonomic exercise. It has practical implications for how we build detection systems (moving beyond verification to interpretation), how we annotate data (embracing perspectivism rather than enforcing consensus), how we

evaluate progress (replacing binary metrics with spectral ones), and how we study the role of language technologies that are simultaneously tools for detecting bias and vectors for propagating it.

The inaugural edition of the InDor workshop represents an opportunity to define the scope of information disorder research broadly enough to include the invisible layer that has been hiding in plain sight. We urge the community to seize it.

Ethics Statement

This position paper advocates for greater attention to media bias within information disorder research. We acknowledge that defining what constitutes "bias" is inherently normative and culturally situated. Any operationalization of bias detection carries the risk of reflecting the perspectives and blind spots of its designers. We do not advocate for automated censorship or content removal based on bias scores; rather, we call for tools that increase transparency about framing choices and support media literacy.

Limitations

This paper is a position paper grounded in a systematic review rather than an empirical study. While we draw on evidence from over 100 reviewed studies, our arguments about the relative importance of media bias within information disorder are interpretive rather than experimentally verified. The taxonomy presented in Table 1 is a synthesis that necessarily simplifies the diversity of bias phenomena documented in the literature. Additionally, our analysis of the disconnect between research communities is based on citation patterns and shared task participation observed in the review, which may not capture all forms of cross-pollination (e.g., informal collaborations or unpublished work).

9. Bibliographical References

Hunt Allcott and Matthew Gentzkow. 2017. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236.

Isabelle Augenstein, Christina Lioma, Dongsheng Wang, Lucas Chaves Lima, Casper Hansen, Christian Hansen, and Jakob Grue Simonsen. 2019. MultiFC: A real-world multi-domain dataset for evidence-based fact checking of claims. In *Proceedings of EMNLP-IJCNLP*, pages 4685–4697. Association for Computational Linguistics.

Ramy Baly, Georgi Karadzhov, Jisun An, Haewoon Kwak, Yoan Muhammed, and Preslav Nakov. 2020. What was written, by whom, and when? analyzing news through source-level factuality and bias. In *Proceedings of EMNLP*, pages 5765–5781. Association for Computational Linguistics.

Valerio Basile, Federico Cabitza, Andrea Campagner, and Michael Fell. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21. Association for Computational Linguistics.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):6860–6868.

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of ACL-IJCNLP*, pages 438–444. Association for Computational Linguistics.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of EMNLP-IJCNLP*, pages 5636–5646. Association for Computational Linguistics.

Stefano DellaVigna and Ethan Kaplan. 2007. The fox news effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Dalvi, Hassan Sajjad, Alex Nikolov, Yordan Atadjanov, and Preslav Nakov. 2021. SemEval-2021 task 6: Detection of persuasion techniques in texts and images. In *Proceedings of SemEval*, pages 70–98. Association for Computational Linguistics.

Jakob-Moritz Eberl, Hajo G. Boomgaarden, and Markus Wagner. 2017. One bias fits all? three types of media bias and their effects on party preferences. *Communication Research*, 44(8):1125–1148.

Ullrich K. H. Ecker, Stephan Lewandowsky, John Cook, Philipp Schmid, Lisa K. Fazio, Nadia Brashier, Panayiota Kendeou, Emily K. Vraga, and Michelle A. Amazeen. 2022. The psychological drivers of misinformation belief and its resistance to correction. *Nature Reviews Psychology*, 1(1):13–29.

Robert M. Entman. 1993. Framing: Toward clarification of a fractured paradigm. *Journal of Communication*, 43(4):51–58.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of EMNLP-IJCNLP*, pages 6343–6349. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Rama Rohit Reddy Gangula, Suma Dunn, and Jacob Eisenstein. 2019. Detecting media bias in news articles using gaussian bias distributions. In *Proceedings of the 2nd Workshop on NLP for Internet Freedom*, pages 48–55. Association for Computational Linguistics.

Tim Groeling. 2013. Media bias by the numbers: Challenges and opportunities in the empirical study of partisan news. *Annual Review of Political Science*, 16:129–151.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2019. Automated identification of media bias in news articles: An interdisciplinary literature review. *International Journal on Digital Libraries*, 20(4):391–415.

David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. 2018. The science of fake news. *Science*, 359(6380):1094–1096.

Sora Lim, Adam Jatowt, Masatoshi Yoshikawa, and Antoine Doucet. 2020. Annotating and analyzing biased sentences in news articles using crowdsourcing. In *Proceedings of the 12th Language Resources and Evaluation Conference (LREC)*, pages 1478–1484.

Liqiang Lin, Pengfei Li, and Frank Ferraro. 2024. IndiVec: An exploration of leveraging large language models for media bias detection with fine-grained bias indicators. In *Proceedings of EACL*. Association for Computational Linguistics.

Fatima Maab, Wasim Afzal, and Muhammad Kamran Malik. 2024. Media bias detection across language model families. In *Proceedings of NAACL*. Association for Computational Linguistics.

Gregory J. Martin and Ali Yurukoglu. 2017. Bias in cable news: Persuasion and polarization. *American Economic Review*, 107(9):2565–2599.

Maxwell E. McCombs and Donald L. Shaw. 1972. The agenda-setting function of mass media. *Public Opinion Quarterly*, 36(2):176–187.

Denis McQuail. 2010. *McQuail's Mass Communication Theory*, 6th edition. SAGE Publications.

Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of IJCAI*, pages 4551–4558.

Sihyeon Park and Kunwoo Kim. 2025. Is the source reliable? the effect of media outlet names on bias detection in LLMs. In *Proceedings of EMNLP*. Association for Computational Linguistics.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of SemEval*, pages 2343–2361. Association for Computational Linguistics.

Riccardo Puglisi and James M. Snyder Jr. 2015. Empirical studies of media bias. *Handbook of Media Economics*, 1:647–667.

Marta Recasens, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2013. Linguistic models for analyzing and detecting biased language. *Proceedings of ACL*, pages 1650–1659.

Francisco-Javier Rodrigo-Ginés, Jorge Carrillo-de Albornoz, and Laura Plaza. 2024. A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it. *Expert Systems with Applications*, 237:121641.

Dietram A. Scheufele. 1999. Framing as a theory of media effects. *Journal of Communication*, 49(1):103–122.

Pamela J. Shoemaker and Tim P. Vos. 2009. Gatekeeping theory. *Routledge*. Book.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021a. MBIC – a media bias annotation dataset including annotator characteristics. In *Proceedings of the iConference*, pages 399–407.

Timo Spinde, Manuel Plank, Jan-David Krieger, Terry Ruas, Bela Gipp, and Akiko Aizawa. 2021b. Neural media bias detection using distant supervision with BABE – bias annotations by experts. In *Findings of EMNLP*, pages 1166–1177. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: A large-scale dataset for fact extraction and VERification. In *Proceedings of NAACL-HLT*, pages 809–819. Association for Computational Linguistics.

Vojtěch Trhlík and Pontus Stenetorp. 2024. Generative models for media bias detection. https://arxiv.org/abs/2406.10773. ArXiv:2406.10773.

Cristian Vaccari and Andrew Chadwick. 2020. Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society*, 6(1).

Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The spread of true and false news online. *Science*, 359(6380):1146–1151.

Claire Wardle and Hossein Derakhshan. 2017. Information disorder: Toward an interdisciplinary framework for research and policy making. Technical Report DGI(2017)09, Council of Europe. Report prepared for the Council of Europe.

Lucie-Charlotte Wen and Lina Younes. 2024. Can GPT-3.5 detect media bias? an evaluation on the MBIB benchmark. https://arxiv.org/abs/2403.20158. ArXiv:2403.20158.

Savvas Zannettou, Michael Sirivianos, Jeremy Blackburn, and Nicolas Kourtellis. 2019. The web of false information: Rumors, fake news, hoaxes, clickbait, and various other shenanigans. *Journal of Data and Information Quality*, 11(3):1–37.