Combating Disinformation: Is There No Alternative?

Davide Bassi[1*], **Søren Kirkegaard Fomsgaard**[2*], **Erik Bran Marino**[3*], **Katarina Laken**[1,4]
[1]CiTIUS - Universidade de Santiago de Compostela, Santiago de Compostela, Spain
[2]GREYC - University of Caen, Caen, France
[3]CIDEHUS - University of Évora, Évora, Portugal
[4]FBK - Fondazione Bruno Kessler, Trento, Italy
davide.bassi@usc.es, soren.fomsgaard@unicaen.fr, erik.marino@uevora.pt, alaken@fbk.eu

Abstract

This position paper critiques the dominance of detection-centered approaches in misinformation research. We argue that the prevailing paradigm treats information disorders as a content-level anomaly to be identified and suppressed, thereby obscuring the structural conditions under which different forms of information disorders emerge and resonate. Drawing on critical anthropology, we propose an alternative "clinical" model: information disorders should be understood not only as informational distortion, but as a syndrome with complex causes embedded in contexts of economic precarity, institutional distrust, and informational inequality. Treating detection as the ends rather than the means of intervention risks misguiding our efforts. Rather than positioning NLP primarily as a tool for boundary enforcement, we outline a reorientation toward structural diagnosis: diversifying data beyond WEIRD contexts, extracting socioeconomic and trust-related signals from discourse, and integrating computational outputs within interdisciplinary causal frameworks. Under this model, detection becomes a means for an epidemiology of discourse, subordinated to the broader objective of cultivating long-term epistemic resilience in our online environments.

Keywords: Disinformation, Natural Language Processing, Research Paradigm

1. NLP Research in the Age of Infodemics

Taken as a codification of medical ethics, the Hippocratic Oath famously frames the moral role of the physician in the practice of their craft. Until the turn of the 20th century, physicians under this oath treated their patients with a beneficent paternalistic attitude, trying to diagnose the causes of their ailments (Will, 2011). Much of the work in NLP based on what we will term the «Detection Paradigm» (cf. Section 2), takes a similar stance with respect to tackling information disorders (Wardle and Derakhshan, 2017). However, instead of focusing on diagnosing and treating causes, we argue that current work hyperfocuses on treating information disorders via detecting and eliminating, from a position of authority.

This paper argues for an uprooting and a reframing both of the paradigm of intervention in information disorders and of the role of the researcher within this paradigm. Instead of focusing exclusively on understanding mis- and disinformation in terms of content (and its production, dissemination and consumption[1]), with the aim of intervening on it through detection and moderation alone, we argue for an etiological re-orientation of the epistemics, practice of NLP research on information disorders.

One of the great epidemics of our time is arguably the epistemic crisis brought on by the combination of more or less regulated mass social media and the undermining of epistemic trust caused by the prevalence and political weaponization of synthetic content created using generative artificial intelligence. This so-called "infodemic"[2] is a multifaceted problem that must be tackled from multiple angles, for example as outlined in the five pillars of the High-Level Expert Group on Fake News and Online Disinformation (European Commission. Directorate General for Communications Networks, Content and Technology., 2018; European Parliament. Directorate General for Parliamentary Research Services., 2019). Our position is that a failure to correctly frame the nature of the problem of online mis- and disinformation as a syndrome of a broader disease, leads to a failure of treatment. Rather than focusing on providing palliative care, which is what relying on detection and moderation entails, we should pivot towards treating the root causes of the disease. Doing this requires understanding the underlying structural causes that lead to the production, dissemination and consumption lifecycle of information disorders. In this context, we argue that we should think of NLP researchers and practitioners as being akin to physicians or nurses. The role of researchers should in the first place be to diagnose and understand the disease based on the syndrome exhibited by the patient. The rest of the paper proceeds as follows. We first

[*]Equal contribution.
[1]cf. the C5 model (Kruijver et al., 2025).
[2]WHO.

Proceedings of the 1st Workshop on Information Disorder (InDor) @ LREC 2026, pages 1–15
12 May 2026.

characterize the Detection Paradigm in Section 2 and criticize examples of it when put into practice in Section 3. Section 4 looks at medical practice for a holistic understanding of disease treatment as a candidate model. Section 5 discusses sociological causes of disinformation. In Section 6, we turn our attention to pointing out possible avenues for reorienting future NLP research away from its focus on detection towards understanding the causes of information disorders, by drawing inspiration from clinical practices.

2. Defining the Detection Paradigm

What we refer to here as the Detection Paradigm (DP) is the predominant trend in NLP and computational social science research to focus on combating adverse social phenomena by *reducing* them to algorithmic detection problems. This ties in with a broader discussion of the 'taskification' of NLP, in which complex problems are reduced to tasks with clear benchmarks (Schlangen, 2021; Litschko et al., 2023). The Detection Paradigm is driven by reductionist epistemics, and it tends to favor solutionist interventions in practice.

Reductionist Epistemics. The DP is based on the tacit assumption that adverse complex phenomena such as misinformation can be captured by proxy of digital representations of language and media. The DP understands misinformation through a combination of two reductive epistemic components: propositionalism[3] and the reductive conceptual framework of the Conduit Metaphor (Reddy, 1979). The first reifies the problem of misinformation and construes it to be reducible to assertions about states of affairs, whose contents are verifiably false. Under this view, misinformation can be understood purely in terms of the truth value of its contents, without considering social, cultural, psychological or technological contexts within which information is produced, distributed and consumed. The second, originally inspired by information theory, frames communication as the unidirectional process of transmitting some propositional, informational content by a sender through a channel to a receiver. Miscommunication, and by extension, misinformation, is the result of imperfect transmission under this conceptual framework. Together, these two reductive components constrain both our understanding of the complex interrelated causes and effects of misinformation and our capacity to intervene on it.

[3]While propositionalism is relevant for information disorders such as mis- and disinformation, this aspect of DP does not necessarily extend to other harmful social phenomena typically modeled in NLP such as hate speech or propaganda.

Solutionist Interventionism. At the level of praxis, the DP is characterized by solutionist interventionism: the tendency to construe the social and political problem posed by misinformation as a technical problem that requires technical (engineering) solutions. Researching misinformation thus becomes the problem of accurately detecting it so that it can be eliminated, as opposed to, for instance, understanding what causes it or how and why it is spread.

The literature in NLP applied to adverse social online phenomena, broadly speaking, tends to reflect and favor the DP. For example, work has been done on detecting hate speech (Warner and Hirschberg, 2012; Djuric et al., 2015; Waseem and Hovy, 2016; Davidson et al., 2017), cyberbullying (Dinakar et al., 2011), trolling (Mihaylov and Nakov, 2016), offensive (Razavi et al., 2010) and abusive language (Nobata et al., 2016), propaganda (Amazon Alexa AI et al., 2021; Yoosuf and Yang, 2019), fake news (Ahmed et al., 2017; Ghosh and Shah, 2019; Pérez-Rosas et al., 2018), rumors (Derczynski and Zubiaga, 2020), conspiracy theories (Gupta, 2025), harassment (Bretschneider et al., 2014), political bias (Gangula et al., 2019), polarizing content (Graumas et al., 2019), to name some popular research topics of the last decade.

3. Critique of the Detection Paradigm

3.1. The Logic of the Instrument

The solutionist character of the DP, as defined above, is not merely a policy preference but a structural consequence of the available technical inventory. As Maslow's maxim states: «It is tempting, if the only tool you have is a hammer, to treat everything as if it were a nail» (Maslow, 1966). Classification models enforce a reduction of output into discrete decision boundaries by their very design. Content is ultimately classified as safe or unsafe, true or false, spam or legit, to trigger automated removal or labeling. This reconfiguration is based on an epistemic reduction: the efficacy of the system relies on the stability of truth values, a condition that holds for only a subset of communicative acts. It is necessary to distinguish between two modes of epistemic discourse, meaning the way we frame knowledge through discourse:

Mode A (Empirical Verification): Discourse regarding verifiable facts, such as election dates, physical events, or geospatial data. These claims possess a binary truth value verifiable against a shared objective reality.

Mode B (Political Hermeneutics): Discourse regarding values, legal interpretations, historical narratives, or future predictions. These are subjective, normative, and dependent on the observer's

political framework.

The detection paradigm applies operational logic designed for Mode A to discourse belonging to Mode B. This constitutes a «category mistake» that attempts to purify political hybridity into binary facts (Latour, 2012). Automated Fact-Checking (AFC) systems exemplify this category mistake by treating interpretive discourse (Mode B) as empirically verifiable (Mode A). Because AFC architectures demand single veracity labels, they force underspecified claims, such as 'It is illegal in Illinois to record a conversation', into reductive binary classifications (e.g., True/False), completely ignoring unstated contextual variables (like whether it is done secretly or not) (Glockner et al., 2024). Furthermore, this strict true/false dichotomy proves epistemically inadequate even for scientific information. Given the inherently probabilistic and evolving nature of scientific knowledge, assessing gradients of accuracy is far more meaningful than enforcing absolute binary classifications. And, as La Barbera et al. (2025) argue, these binary labels do more than oversimplify reality; they act as authoritative judgments that delegitimize valid alternative interpretations and justify unwarranted interventions.

Crucially, this algorithmic flaw is the crystallization of human epistemic confusion. For instance, following U.S. House Resolution 894, which moved to equate anti-Zionism with antisemitism, Meta updated its technical moderation policy on the term «Zionist» (U.S. House of Representatives, 2023; Meta Transparency Center, 2024) to reflect this shift in political (mode B) discourse. When such political definitions are operationalized into moderation systems the infrastructure of safety automates the suppression of counter-hegemonic discourse (Gramsci, 1929/2020), censoring political debate under the designation of «misinformation correction».

The persistence of this category mistake serves a functional role in the preservation of the status quo. The exclusive focus on content moderation acts as a palliative measure, lowering the «temperature» of the public arena by suppressing the symptoms (e.g., toxic polarization and informational noise) without addressing the underlying infection.

3.2. The Detection Paradigm in Practice

The pervasiveness of such an approach is also evident from the institutional responses to the problem. Considering the European context, while the EU recognizes how the information crisis intersect with other crises, identifying it as one among multiple elements of the polycrises we are facing (F et al., 2025), when it comes to defining the actions to be put in place to address it, adopts a purely information based approach. This is particularly evident when considering the "EU Code of Conduct on Misinformation" (European Commission, 2025). This document outlines 6 macro strategies through which we should tackle misinformation:

- *Demonetization of disinformation*: cutting financial incentives by preventing ads and monetization next to disinformation content.
- *Political advertising transparency*: making political and issue ads clearly identifiable, traceable, and publicly accessible.
- *Integrity of services*: reducing manipulative behaviors (bots, fake accounts, coordinated inauthentic behavior, deepfakes).
- *User empowerment*: enable users to recognize, contextualize, and challenge disinformation (with labels and digital tools or fostering media literacy).
- *Empowering researchers & fact-checkers*: providing such roles with data access and structural support for independent scrutiny
- *Monitoring, transparency & accountability*: ensuring measurable commitments, reporting, audits, and public oversight of implementation.

Although articulated as distinct interventions, these measures share a conceptualization of information disorders as a detectable anomaly to be contained and eradicated from the information environment. In this sense, they constitute variations of the same logic, converging on the same instrumental objective: the detection and suppression of problematic content. Even media literacy initiatives aimed at empowering users can be understood as operating within this logic: instead of automating detection externally, they internalize it within users, reconfiguring them as self-monitoring nodes of the information environment. However, once examined beyond its immediate operational aims, the apparent coherence of this approach reveals a set of critical repercussions that unfold at multiple and interrelated levels.

Effectiveness: given the low base rate of misinformation consumption in the "Global North" (approximately 0.15–6% of total media consumption), intervention focused on its detection and suppression are bound to yield only marginal improvements in the general quality of the informational environment (Acerbi et al., 2022); thereby exposing a structural limit of detection-centered approaches.

Feasibility: detection-based governance presupposes sustained investment by private platforms in trust and safety infrastructures (e.g. technical systems, human moderators, and audits). Despite regulatory initiatives such as the EU Digital Services Act (DSA), which formally mandates systemic risk mitigation and transparency obligations,

recent developments showed how the viability of such paradigm rests on corporate incentives, which, in turn, are neither stable nor inherently aligned with long-term public epistemic resilience. In 2023, Meta laid off approximately 21,000 employees, with reported consequences for its Trust and Safety operations (Field and Vanian, 2023). Similarly, following Elon Musk's takeover, X dismissed nearly 80% of engineers dedicated to trust and safety functions (Brewster, 2023). In the same period, Google reduced the staff of *Jigsaw*, its internal unit focused on countering online toxicity and hate speech by roughly one third (Field and Vanian, 2023). These developments reveal a structural fragility: the effectiveness of detection-centered interventions depends on the continuous allocation of resources by profit-driven corporations. When economic priorities shift, the operational capacity of the paradigm contracts accordingly, regardless of formal regulatory frameworks. The scenario is then further complicated by initiatives like the portal `freedom.gov` hosted by the US government, intended to allow citizens in Europe and elsewhere to access material that European laws have sought to restrict — including alleged hate speech and content linked to terrorism — by bypassing local content restrictions. Although framed by U.S. officials as an effort to promote digital freedom, critics argue that the portal could undercut the regulatory aims of the Digital Services Act and analogous frameworks, and that it embodies a fundamentally different normative stance toward harmful content (Down, 2026).

Public Legitimacy: as shown by Rich et al. (2020), the support for social media fact checking is highly uneven and strongly structured by partisanship. While most of the people endorse fact-checking "in the abstract", support declines when interventions target politically salient actors or in-group figures, with fact-checks frequently perceived as partisan or punitive rather than corrective. As a result, fact-checking is effective mainly among those already inclined to trust the institutions performing it, thereby limiting its broader legitimacy in polarized information environments.

Explanatory scope: detection-centered approaches systematically under-specify the determinants of mis- and disinformation by privileging content-level properties over both individual and structural drivers. At the individual level, they largely ignore the role of attitudinal factors—such as partisanship, identity management, and motivated reasoning—that shape how users actively select, interpret, and engage with information (Altay et al., 2023; Robertson et al., 2023; Bakshy et al., 2015). At the structural level, they overlook the broader socio-economic and institutional conditions under which misinformative and conspiratorial narratives flourish, including corruption, economic insecurity, inequality, and deficits of institutional trust (Hornsey and Pearson, 2022; Adamus et al., 2024; Amazeen et al., 2024). By doing so, the DP obscures how engagement with misinformation functions as a response to unresolved societal needs, such as equal access to information, political representation and agency, institutional credibility and shared spaces for negotiating identity and social belonging (we elaborate more on this in Sec. 5). In this sense, the detection paradigm exemplifies what Morozov (2013) terms *technological solutionism*: the displacement of political and structural questions by ostensibly neutral technical fixes.

A physician that treats the symptom itself as the disease, however, cannot prescribe an adequate cure.

4. The Ideal Clinical Model

It thus becomes evident that the DP fosters a reductive view of mis- and disinformation, that remains blind to the underlying societal causes of the issue, by modeling it as a problem of the verification of context-free propositions. Staying with our medical metaphor, our critique for this "blind obsession with the symptom", (i.e. misinformation) closely parallels long-standing critiques advanced by critical medical anthropology. In that context, the reduction of illness to isolated biological dysfunctions, while neglecting their sociological, contextual, and phenomenological dimensions, has been shown to result in the medicalization of fundamentally social problems (Singer, 2009; Farmer, 2004).

An 'ideal clinical model' would instead treat misinformation as a diagnostic signal, not as a pathological object to be excised from the information environment. Just as critical medical approaches situate illness within broader configurations of social inequality, institutional power and lived experience, a structural approach to misinformation must situate communicative practices within the political–economic conditions that make certain narratives resonate. Engagement with misinformative content should therefore be analyzed as an action occurring within an arena structured by historically sedimented distrust, economic precarity, identity fragmentation, and transformations in the architecture of the public sphere. In this sense, mis- and disinformation become intelligible as relational phenomena emerging at the intersection between individual and group agencies and structural constraints. They are not reducible to epistemic deficit; rather, they reflect the dynamic interplay between the social construction of interpretative categories and the material conditions that shape everyday life. From this perspective, the task is not "merely" to correct claims, but to understand the processes through which meanings are produced,

circulated, and appropriated. Finally, reframing mis- and disinformation as a capability-depriving condition embedded in the political economy of the public sphere, the most effective intervention is the one that aims to restore the epistemic and social capabilities required for meaningful participation in collective sense-making, rather than solely at suppressing inaccurate content (Nussbaum, 2011).

5. Diagnosing the Sociological Causes of Disinformation

To operationalize the 'ideal clinical model' described in Sec. 4, this section aims at outlining (some of) the structural and contextual conditions under which misinformative narratives emerge, circulate, and acquire relevance. Disinformation is here understood as embedded in transformations of the public sphere (Jungherr and Schroeder, 2021): susceptibility and engagement are systematically patterned by social position and institutional credibility rather than by simple deficits in reasoning (Altay and Mercier, 2026; Schirmer et al., 2025).

Unequal Access to High-Quality Information: as professional journalism increasingly relies on subscription-based models, high-quality news becomes partially excludable, while sensationalist or low-quality content remains usually freely accessible. In high-income countries, only 17% of individuals pay for online news and subscribers are disproportionately affluent and highly educated (Newman et al., 2024). Informational quality thus becomes stratified along socioeconomic lines. Crucially, much information disorder research presupposes equal access to credible information, thereby overlooking how structural inequalities condition exposure opportunities in the first place (Schirmer et al., 2025). When everyday life is organized around financial instability and short-term survival concerns, cognitive resources are disproportionately allocated to immediate problem-solving, leaving diminished capacity for effortful tasks such as cross-checking sources or evaluating evidentiary quality (De Bruijn and Antonides, 2022; Burlacu et al., 2022). This makes information verification not simply a matter of motivation or competence, but a resource-intensive practice. Under these conditions, vulnerability to low-quality information reflects infrastructural asymmetry and structurally constrained cognitive bandwidth, rather than merely individual epistemic failure (Ronzani, 2025).

Structural Transformations and Gatekeeping: The digital public arena has eroded the influence of traditional gatekeepers, such as journalists and political parties, who previously filtered information flows (Jungherr and Schroeder, 2021). While this has lowered barriers to entry for marginalized topics and perspectives previously neglected by mainstream media, it has also produced an environment in which information visibility depends less on professional verification and more on algorithmic amplification and user engagement (Marino et al., 2024). Most major social media platforms operate on engagement-based business models in which clicks, shares and watch time drive advertising revenue (Napoli, 2019): where early internet content discovery was largely user-led, algorithmic recommendation feeds shifted this paradigm toward system-driven distribution, in which emotionally arousing and polarizing content is systematically amplified regardless of its truth value (Vosoughi et al., 2018). This represents an externalization of epistemic responsibility: the burden of filtering, once held by institutional gatekeepers, has been reallocated as a substantial cognitive load onto the individual. Such pressure is reflected in the finding that 39% of the global population actively avoids the news not because they consider it false, but because they find it overwhelming and emotionally draining (Newman et al., 2024). The resulting vulnerability is not primarily to the persuasion of false claims, but to the fragmentation of the shared frames through which collective reality is constructed.

Economic Uncertainty and Loss of Agency: socioeconomic dynamics can also shift how people make sense of information by undermining perceived control and institutional legitimacy. Higher levels of objective and perceived economic inequality increase openness to alternative explanatory frames and endorsement of conspiracy beliefs (Altay and Mercier, 2026). This effect is fully mediated by *anomie* (i.e. the perception that society and its institutions are breaking down) thereby motivating a search for meaning and control that conspiratorial narratives promise to restore (Casara et al., 2022). In this view, conspiracy beliefs are not merely incorrect beliefs, but compensatory explanations that become attractive under existential threat: when perceived agency is low, conspiratorial accounts offer coherent causal stories, identifiable culprits, and a sense of epistemic and moral leverage over events (Douglas et al., 2019). Crucially, this also means that counter-disinformation strategies focused exclusively on fact-checking fail to address the demand for alternative narratives, which frequently stems from a grain of truth generated by material inequality — what Benford and Snow (2000) call *frame resonance*: a narrative succeeds not simply because it is available, but because it resonates with the lived experience of those who feel abandoned by institutions (Waisbord, 2018).

Historical Roots of Systemic Distrust: distrust in official information is also historically structured. Communities that have experienced repeated neglect, discrimination, or harm at the hands of trusted

authorities develop durable patterns of distrust that constitute rational responses to structural conditions (Smith and Freyd, 2014). Cox (2024) finds that a majority of Black Americans believe that societal institutions are designed to disadvantage them, a perception that can reduce reliance on government-backed health or climate messaging (Jaiswal et al., 2020). Such attitudes are grounded in documented histories of discrimination and institutional betrayal (particularly in medical and scientific domains, epitomized by the legacy of the Tuskegee Syphilis Study) which have long shaped patterns of epistemic mistrust (Bogart et al., 2021). From this perspective, distrust operates as an adaptive response to historically sedimented expectations shaped by repeated experiences of exclusion and harm.

Corruption and Institutional Credibility: institutional distrust is not only historically sedimented but is continuously reproduced by contemporary governance quality. In contexts where corruption is salient, conspiratorial explanations become structurally less implausible. Cordonier and Cafiero (2024) shows that higher levels of public sector corruption are associated with stronger endorsement of conspiracy beliefs, even when controlling for democracy, press freedom, inequality, and human development. Relatedly, populist attitudes - often rooted in anti-elite and anti-institutional rhetoric - predict reduced trust in political and scientific institutions and increased reliance on alternative media, thereby indirectly fostering lower compliance with expert-backed guidance (Ehrke et al., 2023).

Identity and Social Signaling: online platforms should not be conceptualized merely as information distribution systems, but as arenas for identity performance. Bail (2022) argues that social media environments encourage users to publicly signal group belonging, moral commitments, and political identities, rewarding expressions that are distinctive, emotionally charged, or normatively aligned with in-group expectations. In this context, sharing or affirming contested claims can function less as a statement of epistemic conviction than as an act of social positioning. As Mercier (2020) and Bassi et al. (2025) contend, individuals often endorse or publicly defend ideas - including implausible ones - not because they are epistemically persuaded, but because doing so reinforces alliances and distinguishes adversaries. At its extreme, this dynamic converges with the phenomenon of *conspirituality* (Ward and Voas, 2011), in which the search for alternative spiritual truths merges with the quest for hidden political ones, forming self-reinforcing epistemic communities organized around identity rather than evidence. Empirical data confirms the depth of this motivated selective exposure: individuals with highly unfavorable views on vaccines are nine times more likely to visit vaccine-skeptical websites than their favorable peers; during the 2016 US election, Trump supporters were twice as likely to visit unreliable news sites compared to Clinton supporters (Altay and Mercier, 2026). For these users, the problem is not a lack of data but a lack of institutional trust, and misinformation circulates in part because it serves relational and identity-building functions that accurate information does not.

5.1. Two Epistemic Positions: The Uninformed and the Disinformed

The structural conditions described above do not produce a single, homogeneous 'misinformed public'. Following a classical methodological distinction between ideal-typical categories (Weber, 1949), we argue that they instead generate two structurally distinct epistemic positions that are too often conflated in the literature.

The *uninformed* are those passively excluded from accurate information by structural barriers (e.g. access costs, digital skill deficits, cognitive bandwidth constraints) rather than by any motivated rejection of it. Their relationship to misinformation is one of exposure without resistance, driven by infrastructural asymmetry rather than epistemic choice.

The *disinformed*, by contrast, occupy an active position: they encounter accurate information but reject it, because acceptance would threaten group identity, undermine social alliances, or destabilize the compensatory narrative frameworks that give meaning to experiences of institutional betrayal and loss of agency.

This distinction carries direct implications for intervention design. Strategies premised on correcting an information deficit (e.g. fact-checking, content labeling, source literacy training) are well-suited to the uninformed, but structurally misaligned with the disinformed, for whom the problem is not epistemic but relational and political. Treating the disinformed as merely uninformed ignores the profound moral and social dimensions of resistance to institutional truth claims (Morozov, 2013). Effective intervention must instead engage with the structural conditions we outlined that make conspiratorial and misinformative frameworks not just available, but meaningful.

6. Reorienting NLP Research

If misinformation is not merely anomalous content to be removed, but a symptom of deeper structural transformations, research efforts must shift toward modeling the contextual and systemic conditions under which such narratives emerge and acquire meaning. While the macro-level determinants discussed in Section 5 have traditionally been investi-

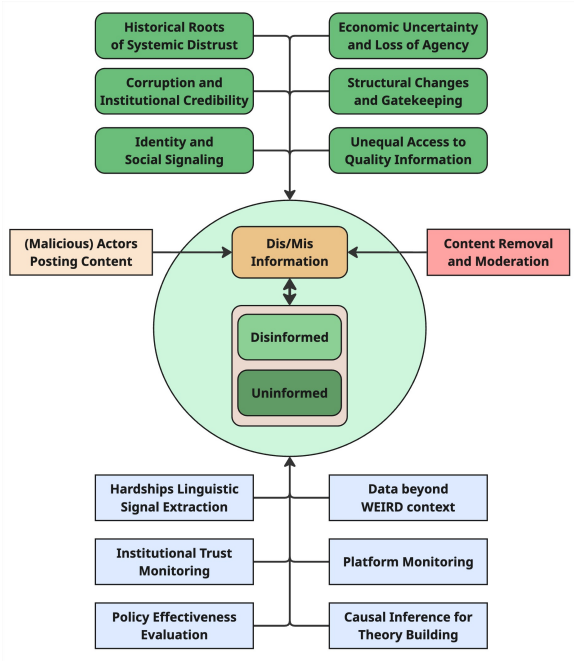

Figure 1: Structural model of mis- and disinformation. Dark green boxes represent structural causes shaping the socio-epistemic context (light green) in which disinformation (orange) is consumed by two publics: the *disinformed* and the *uninformed*. Blue boxes indicate proposed NLP contributions to tackle the issue.

gated within sociology, political science, and psychology, this raises a distinct question for computational research: what is the specific contribution that NLP can provide within a structurally informed framework? We contend that this contribution lies in moving beyond content-level classification toward the large-scale operationalization of structural variables. In the following, we outline concrete research directions through which the NLP community can support such a reorientation.

Data and Sampling: a central limitation of the current literature concerns its sampling bias. More than 80% of misinformation research is conducted in high-income Western democracies and typically relies on young, educated, urban, and digitally literate participants (Badrinathan et al., 2025; Blair et al., 2024). As emphasized by Ronzani (2025), this narrow empirical base risks generating globally generalized conclusions from demographically restricted populations, thereby obscuring the role of structural variance in misinformation dynamics. This issue is not merely geographical but socioeconomic. Schirmer et al. (2025) highlights how digital exclusion, economic precarity, language barriers, and historically rooted institutional mistrust shape differential exposure to and engagement with misinformation. Yet such dimensions remain underrepresented in large-scale computational studies. The main challenge in reorienting NLP research beyond Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations lies on the fact that in many contexts, misinformation circulates not primarily through public social media feeds but via encrypted messaging platforms (e.g., WhatsApp, Telegram), local radio broadcasts, SMS chains, community gatherings, and religious networks (Badrinathan et al., 2025; Blair et al., 2024). However, recent methodological innovations provide viable pathways for accessing these environments responsibly, including chatbot-based recruitment strategies (Bowles et al., 2020) and structured WhatsApp data-donation infrastructures (Garimella and Chauchard, 2025).

Sociological Signal Extraction from Discourse: NLP can contribute by operationalizing *structural variables as first-class prediction targets*, i.e., by devising models that detect when discourse expresses specific sociological dimensions. Rather than detecting misinformation for removal, models can be trained to identify explicit and implicit hardship narratives such as inability to afford basic goods, instability of work and housing, and perceived unfairness of the economic system. Operationally, this implies building datasets and taxonomies for categories like material deprivation, economic insecurity, downward mobility, and resource-strain complaints, enabling supervised or weakly supervised detection at scale. For instance, hardship detection targets discourse reporting scarcity and struggle; identity-threat detection can target rhetorical strategies used to signal belonging and mark disaffiliation (Bassi et al., 2025); and political alienation can be operationalized via stigmatizing and exclusionary language detection (Sap et al., 2020; Straton et al., 2020; Bassi et al., 2024). This approach aligns with (i) text-as-data methods that treat language as measurement of latent social conditions (Puklavec et al., 2023), and (ii) recent domain-specific efforts to formalize poverty-related social phenomena in NLP, including datasets and taxonomies targeting poverty-related bias and discourse about poverty (Curto et al., 2025). The resulting signals can be linked to regional socioeconomic indicators (e.g., unemployment, inequality), educational access, or infrastructure metrics, enabling multi-level modeling of how material constraints shape informational engagement and vulnerability pathways. Although modeling structural vulnerability in NLP entails non-trivial ethical risks related to surveillance and stigmatization (cf. Section 6.1), these must be weighed against the long-term benefits of making such constraints publicly legible. From a capabilities-based perspective (Nussbaum, 2011), rendering these constraints visible informs interventions, largely outside NLP itself, aimed at improving the struc-

tural environments in which epistemic vulnerability arises, thereby promoting full human functioning and meaningful participation in collective sense-making.

Institutional Trust Monitoring Systems: A complementary direction concerns longitudinal monitoring of institutional legitimacy via construct-level detection of trust-related social phenomena. Specifically, NLP systems can be designed to detect specific delegitimization and distrust frames (e.g., corruption, betrayal, capture, manipulation), as well as experiences of institutional abandonment and credibility rupture (Djouvas et al., 2023; Lattmann, 2025). In this sense, NLP tools become a diagnostic dashboard for epistemic infrastructure: not just tracking isolated claims, but monitoring structural erosion in legitimacy and belonging across time, communities, and platforms. NLP systems can be designed to track fluctuations in trust-related discourse across time, communities, and platforms.

Platform and Amplification Auditing NLP and computational methods can also contribute by modeling how platform architectures shape exposure and amplification. Rather than presenting the detection of misinformative content using its semantic properties as a solution to the epistemic crisis, research can investigate how engagement-based ranking systems, recommender algorithms, and network structures condition visibility, diffusion velocity, and cross-group exposure. Combining discourse analysis with network and ranking data would enable systematic auditing of amplification dynamics, identifying structural asymmetries in content propagation independent of truth value. Such approaches shift attention from "what is false" to "what becomes visible and why."

Policy and Intervention Evaluation: NLP-based measurement systems can support the evaluation of structural interventions (Nannini et al., 2025). Computational indicators of institutional trust, polarization dynamics, and exposure asymmetries can be tracked before and after regulatory changes (e.g., platform transparency mandates, media subsidy policies, or moderation reforms), enabling systematic assessment of their downstream epistemic effects. Rather than relying on platform self-reporting or short-term engagement metrics, such infrastructures would allow independent monitoring of how policy decisions reshape informational ecosystems over time; configuring NLP as a component of democratic accountability.

From Measurement to Causal Inference: finally, the value of these tools lies not in their descriptive power alone but in their integration within interdisciplinary causal frameworks. Computational outputs—structural discourse indicators, trust trajectories, network diffusion patterns—can be combined with socioeconomic data, policy interventions, and institutional metrics to support causal inference. Under this reorientation, detection would be subordinated to a broader epistemic objective: understanding the structural conditions under which misinformation emerges, circulates, and acquires meaning.

6.1. Methodological Considerations

To fully realize the clinical model proposed above, it is crucial to recognize the epistemological limits of text-based computational methods. If the root causes of information disorders are structural, then NLP cannot serve as the sole instrument of detection and intervention. Instead, NLP should be conceptualized as a complementary diagnostic tool. It provides a scalable layer for measuring discursive symptoms, but the actual interventions, aimed at restoring epistemic resilience and addressing socioeconomic marginalization, must inherently take place in other domains, such as public policy, institutional reform, and civic education.

Given the possible avenues for reorienting NLP outlined above, we now discuss some general methodological considerations to be kept in mind to correctly operationalize this shift:

Integrating Textual Signals with Socio-Institutional Data NLP-derived indicators acquire diagnostic value only when anchored to external data. Aforementioned discourse signals about institutional distrust, hardship narratives, identity-threat framing should be linked to regional socioeconomic indicators (e.g., unemployment, Gini coefficients, poverty indices) and survey-based trust measures (e.g., Eurobarometer, World Values Survey) to identify whether structural conditions predict discursive vulnerability patterns at scale (Khoun et al., 2025). Bor et al. (2026), for instance, combining cross-national survey data on online political hostility with macro-level democracy and inequality indicators across 30 countries, demonstrates that discourse patterns are structured by societal conditions rather than platform affordances alone.

Hybridizing Detection through Mixed-Methods First, the use of NLP for detection should be reoriented from a standalone classification task into a component of a hybrid, mixed-methods research architecture. Because the core determinants of misinformation lie largely outside the text, computational analysis is intended to be triangulated with external methodologies. This includes, but is not limited to: integrating NLP pipelines with digital ethnography to understand the localized cultural context of online communities (Kozinets and Gretzel, 2024); utilizing sociolinguistic models to distin-

guish expressions of material deprivation (Muneton-Santa et al., 2022); combining semantic classification with platform metadata to audit algorithmic amplification (Nannini et al., 2025); and leveraging LLM-driven swarm intelligence engines to simulate the sociological spread of these narratives *in silico* (i.e. *MiroFish*). In this hybrid model, automated detection is no longer the terminus of the research pipeline, but rather a sensor that informs broader qualitative, quantitative, and computational social investigations.

Reliability and Riskiness of Extracting Social Variables All measuring instruments have error margins, and NLP tools are no different. Modern LLMs are often overestimated in terms of how reliable their outputs are, because they produce very fluent and natural text. However, LLMs often produce factually incorrect responses (Ji et al., 2023) and poorly deal with social context (Hovy and Yang, 2021). This becomes especially salient in cross-cultural contexts (Moayeri et al., 2024), or when dealing with data from non-standard language varieties (often used by minoritized demographics) (Sap et al., 2019; Nguyen et al., 2025).

A large part of NLP research is currently focused on improving model performances on pre-established benchmarks (Schlangen, 2021). But when actually using NLP tools, it is at least equally critical to understand the error margins, both quantitatively and qualitatively. Error patterns are not random, they are the result of systemic biases in the training data and inherent ambiguities of natural language in use. One of the main tasks of NLP practitioners should be to help scholars and policy makers from other fields understand how to responsibly interpret results of NLP tools.

Extracting social variables from text also poses ethical risks in terms of profiling. We do not advocate for the development of systems to profile individuals for social class, wealth, or similar factors. However, we believe that traditional sociological research can benefit from NLP research by revealing aggregate patterns of, among others, social stratification, linguistic variation, and cultural capital.

7. Conclusions

This paper has argued that the prevailing *Detection Paradigm* configures misinformation as a pathological object to be excised from the informational body. Drawing on critical anthropology, we suggested that this mirrors earlier reductions of illness to isolated biological dysfunctions: symptoms are managed, while the social conditions that generate vulnerability remain intact. In clinical contexts, such reductionism has long been criticized for depoliticizing suffering, transforming historically produced inequalities into technical problems of correction. Analogously, when misinformation is framed primarily as a content-level anomaly, structural determinants such as economic precarity, institutional distrust and identity fragmentation are displaced from the analytic foreground. We do not deny that misinformation can exert causal effects (Tay et al., 2024). Yet recognizing multicausality does not imply symmetry of intervention. Even if misinformation contributes to downstream harms, treating detection as the primary endpoint risks confusing proximal triggers with generative conditions. In an 'ideal clinical model,' symptoms are diagnostically meaningful signals of deeper configurations of power, exclusion, and legitimacy. Hence, detection should function as an instrument of structural diagnosis: an '*epidemiology of discourse*', rather than as a self-sufficient cure. This reorientation carries institutional implications. Detection-centered governance presupposes sustained alignment between public epistemic goods and platform incentives, a fragile and politically contingent condition. The proposed approach, instead, seeks to equip public institutions with computational tools capable of identifying patterns of distrust, marginalization, and socio-economic stress that shape the resonance of misinformative narratives. Under this view, NLP shifts from boundary enforcement to infrastructural analysis, contributing to a diagnostic framework that reconnects epistemic disorder to its political–economic substrates.

Concluding, we emphasize how our contribution is intentionally programmatic. The determinants and research directions outlined are illustrative rather than comprehensive, and detection can remain operationally valuable in circumscribed contexts. However, if misinformation is understood not only as distortion but as expression, its analysis must be repoliticized: embedding it within the structural transformations of the public sphere that condition both belief formation and institutional legitimacy. Only within such a framework can mitigation move beyond symptom management toward the cultivation of long-term epistemic resilience.

8. Ethical Considerations

Current NLP Ethics is largely informed by principalist and applied-ethics frameworks, most notably, Value Sensitive Design (Friedman and Kahn, 2002; Friedman et al., 2013). Attention has also been paid to applying explicitly deontological (Prabhumoye et al., 2021) or consequentialist (Card and Smith, 2020) analyses to NLP practices, but even so, ethical considerations in the NLP community have primarily been oriented towards design and deployment practices of researchers and developers (cf. the ACM Code of Ethics and IEEE Global Initiative

on Ethics of Autonomous and Intelligent Systems (Chatila and Havens, 2019)). Assessment of social impact (Hovy and Spruit, 2016), such as harm, risk, and algorithmic fairness, is crucial, but if we are concerned with understanding the mechanisms that drive the expressions of information disorders, we need to re-conceptualize the role of NLP researchers not just as builders of potentially dangerous tools, or as moderators of undesirable content, but as curious caretakers who play a role in improving the structural conditions that shape our information environments for the better. If we instead think of the role of NLP as facilitating the diagnoses of diseases and disorders, it would be appropriate to move away from the beneficent paternalism implied by the Detection Paradigm, towards an attitude that favors autonomy of and care for the populations that we model. Biomedical ethics changed from framing the doctor-patient relationship as one marked by the well-meaning authority of the doctor over their patient to one increasingly characterized by the respect for the autonomy of and interpersonal care for the patient (Will, 2011), largely informed by the Ethics of Care (Gilligan, 1984; Noddings, 1984). NLP research should adopt a comparable attitude towards the individuals and populations we model with our data and tools. Discussions of social harms (Hovy and Spruit, 2016) and ethics-by-design (Leidner and Plachouras, 2017) pushed the ethics of NLP in necessary directions. We believe the Ethics of Care could offer a similar path forward for improving our ethical framework of NLP, just as it has been suggested for AI Ethics for decision making (Villegas-Galaviz and Martin, 2023). Precisely how to faithfully support the autonomy and flourishing of the participants and populations modeled in NLP research applied to phenomena like misinformation remains for now an open question for future research.

Limitations

This paper is intentionally programmatic and therefore subject to several limitations. The structural conditions outlined in Section 5 and the NLP research directions proposed in Section 6 are illustrative rather than exhaustive: both represent theoretically grounded selections, not comprehensive inventories.

Additionally, although the paper critiques the WEIRD bias of existing misinformation research, its own argumentation draws predominantly on evidence and policy examples from European and North American contexts; the structural factors identified may differ substantially in salience and configuration elsewhere.

At the methodological level, the considerations raised in Section 6.1 apply with particular force to the sociological signal extraction direction. NLP tools operate with non-trivial error margins and underperform systematically on non-standard language varieties and cross-cultural contexts. Indicators of institutional distrust, hardship, or identity threat should therefore be treated as noisy proxies requiring external validation before informing policy.

Finally, extracting socioeconomic signals from discourse at scale carries substantive ethical risks. Even when intended for aggregate analysis, such systems can be repurposed for individual profiling or surveillance of vulnerable populations. The boundary between aggregate monitoring and individual-level stigmatization is technically fragile and institutionally contingent — a risk that any operationalization of the proposed framework must explicitly address.

Acknowledgments

This work was supported by the HYBRIDS project, which has received funding from the European Union's Horizon Europe research and innovation programme under the Marie Skłodowska-Curie Grant Agreement No. 101073351 and from the UK Research and Innovation (UKRI) Horizon Europe funding guarantee (Grant Number: EP/X036758/1).

References

Alberto Acerbi, Sacha Altay, and Hugo Mercier. 2022. Research note: Fighting misinformation or fighting for information? *Harvard Misinformation Review*.

Magdalena Adamus, Eva Ballová Mikušková, Pavol Kačmár, Martin Guzi, Matuš Adamkovič, Maria Chayinska, and Jais Adam-Troian. 2024. The mediating effect of institutional trust in the relationship between precarity and conspiracy beliefs: A conceptual replication of adam-troian et al.(2023). *British Journal of Social Psychology*, 63(3):1207–1225.

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments*, pages 127–138, Cham. Springer International Publishing. TLDR: A fake news detection model that use n-gram analysis and machine learning techniques is proposed, which investigates and compares two different features extraction techniques and six different machine classification techniques.

Sacha Altay, Manon Berriche, Hendrik Heuer, Johan Farkas, and Steven Rathje. 2023. A survey of expert views on misinformation: Definitions, determinants, solutions, and future of the field. *Harvard Kennedy School Misinformation Review*, 4(4):1–34.

Sacha Altay and Hugo Mercier. 2026. Misinformation is a symptom: Commentary on Ecker et al. (2025). *American Psychologist*, 81(2):287–289.

Michelle A Amazeen, Rosalynn A Vasquez, Arunima Krishna, Yi Grace Ji, Chao Chris Su, and James J Cummings. 2024. Missing voices: examining how misinformation-susceptible individuals from underrepresented communities engage, perceive, and combat science misinformation. *Science communication*, 46(1):3–35.

Amazon Alexa AI, Seunghak Yu, Giovanni Da San Martino, Department of Mathematics, University of Padova, Italy, Mitra Mohtarami, MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA, James Glass, MIT Computer Science and Artificial Intelligence Laboratory, Cambridge, MA, USA, Preslav Nakov, and Qatar Computing Research Institute, HBKU, Qatar. 2021. Interpretable Propaganda Detection in News Articles. In *Proceedings of the Conference Recent Advances in Natural Language Processing - Deep Learning for Natural Language Processing Methods and Applications*, pages 1597–1605. INCOMA Ltd. Shoumen, BULGARIA.

Sumitra Badrinathan, Simon Chauchard, and Niloufer Siddiqui. 2025. Misinformation and support for vigilantism: An experiment in india and pakistan. *American Political Science Review*, 119(2):947–965.

Chris Bail. 2022. *Breaking the social media prism: How to make our platforms less polarizing*. Princeton University Press.

Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.

Davide Bassi, Giovanni Da San Martino, Renata Vieira, and Martín Pereira-Fariña. 2025. Drawing digital lines: pattern analysis of divisive rhetoric in social network discussions. *Humanities and Social Sciences Communications*, 12(1):2009.

Davide Bassi, Luisa Orrù, Christian Moro, Davide Salvarani, and Gian Piero Turchi. 2024. Investigating avhs narratives through text analysis: the proposal of dialogic science for tackling stigmatization. *BMC psychology*, 12(1):434.

Robert D Benford and David A Snow. 2000. Framing processes and social movements: An overview and assessment. *Annual review of sociology*, 26(2000):611–639.

Robert A Blair, Jessica Gottlieb, Brendan Nyhan, Laura Paler, Pablo Argote, and Charlene J Stainfield. 2024. Interventions to counter misinformation: Lessons from the global north and applications to the global south. *Current Opinion in Psychology*, 55:101732.

Laura M Bogart, Bisola O Ojikutu, Keshav Tyagi, David J Klein, Matt G Mutchler, Lu Dong, Sean J Lawrence, Damone R Thomas, and Sarah Kellman. 2021. Covid-19 related medical mistrust, health impacts, and potential vaccine hesitancy among black americans living with hiv. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 86(2):200–207.

Alexander Bor, Antoine Marie, Lea Pradella, and Michael Bang Petersen. 2026. Social media users experience more political hostility in less economically equal and less democratic societies. *Nature Human Behaviour*.

Jeremy Bowles, Horacio Larreguy, and Shelley Liu. 2020. Countering misinformation via whatsapp: Preliminary evidence from the covid-19 pandemic in zimbabwe. *PloS one*, 15(10):e0240005.

Uwe Bretschneider, Thomas Wöhner, and Ralf Peters. 2014. Detecting Online Harassment in Social Networks. *ICIS 2014 Proceedings*.

T Brewster. 2023. Google cuts company protecting people from surveillance to a 'skeleton crew,' say laid off workers. *Forbes*.

Sergiu Burlacu, Austėja Kažemekaitytė, Piero Ronzani, and Lucia Savadori. 2022. Blinded by worries: sin taxes and demand for temptation under financial worries. *Theory and Decision*, 92(1):141–187.

Dallas Card and Noah A. Smith. 2020. On Consequentialism and Fairness. *Frontiers in Artificial Intelligence*, 3.

Bruno Gabriel Salvador Casara, Caterina Suitner, and Jolanda Jetten. 2022. The impact of economic inequality on conspiracy beliefs. *Journal of Experimental Social Psychology*, 98:104245.

Raja Chatila and John C. Havens. 2019. The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. In Maria Isabel Aldinhas Ferreira, João Silva Sequeira, Gurvinder Singh Virk, Mohammad Osman Tokhi, and Endre E. Kadar, editors, *Robotics and Well-Being*, volume 95, pages 11–16. Springer International Publishing,

Cham. Series Title: Intelligent Systems, Control and Automation: Science and Engineering.

Laurent Cordonier and Florian Cafiero. 2024. Public sector corruption is fertile ground for conspiracy beliefs: A comparison between 26 western and non-western countries. *Social Science Quarterly*, 105(3):843–861.

Kiana Cox. 2024. *Most Black Americans believe US institutions were designed to hold Black people back*. Pew Research Center.

Georgina Curto, Svetlana Kiritchenko, Muhammad Hammad Fahim Siddiqui, Isar Nejadgholi, and Kathleen C. Fraser. 2025. Tackling social bias against the poor: a dataset and a taxonomy on aporophobia. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7010–7031, Albuquerque, New Mexico. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael W Macy, and Ingmar Weber. 2017. Automated Hate Speech Detection and the Problem of Offensive Language. In *ICWSM*.

Ernst-Jan De Bruijn and Gerrit Antonides. 2022. Poverty and economic decision making: a review of scarcity theory. *Theory and Decision*, 92(1):5–37.

Leon Derczynski and A. Zubiaga. 2020. Detection and Resolution of Rumors and Misinformation with NLP. In *COLING*. ECC: No Data (logprob: -123.494) tex.ids: derczynskiDetectionResolutionRumors2020a, derczynskiDetectionResolutionRumors2020b.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the Detection of Textual Cyberbullying. *Proceedings of the International AAAI Conference on Web and Social Media*, 5(3):11–17.

Constantinos Djouvas, Christos Christodoulou, Antonis Charalampous, and Nikandros Ioannidis. 2023. Identifying euroscepticism using a text-as-data approach: An experimental study employing parliamentary speeches. In *18th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP 2023)*, pages 1–6.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate Speech Detection with Comment Embeddings. In *Proceedings of the 24th International Conference on World Wide Web*, WWW '15 Companion, pages 29–30, New York, NY, USA. Association for Computing Machinery.

Karen M Douglas, Joseph E Uscinski, Robbie M Sutton, Aleksandra Cichocka, Turkay Nefes, Chee Siang Ang, and Farzin Deravi. 2019. Understanding conspiracy theories. *Political psychology*, 40:3–35.

A Down. 2026. Us builds website that will allow europeans to view blocked content. *The Guardian*.

Franziska Ehrke, Gloria Grommisch, Emma Penelope Busch, and Magdalena C. Kaczmarek. 2023. Populist attitudes predict compliance-related attitudes and behaviors during the covid-19 pandemic via trust in institutions. *Social Psychology*.

European Commission. 2025. The code of conduct on disinformation.

European Commission. Directorate General for Communications Networks, Content and Technology. 2018. *A Multi-Dimensional Approach to Disinformation: Report of the Independent High Level Group on Fake News and Online Disinformation.* Publications Office, LU.

European Parliament. Directorate General for Parliamentary Research Services. 2019. *Regulating Disinformation with Artificial Intelligence: Effects of Disinformation Initiatives on Freedom of Expression and Media Pluralism.* Publications Office, LU.

Cataldo F, Redecker C, Montanari E, Galariotis I, Greidanus H, and Quetel C. 2025. Rethinking societal resilience in a time of polycrisis. Technical report, Joint Research Center, Ispra (Italy).

Paul Farmer. 2004. An anthropology of structural violence. *Current anthropology*, 45(3):305–325.

H Field and J Vanian. 2023. Tech layoffs ravage the teams that fight online misinformation and hate speech. *CNBC*.

Batya Friedman and Peter H. Kahn. 2002. Human Values, Ethics, and Design. In *The Human-Computer Interaction Handbook*, pages 1177–1201. CRC Press.

Batya Friedman, Peter H. Kahn, Alan Borning, and Alina Huldtgren. 2013. Value Sensitive Design and Information Systems. In Neelke Doorn, Daan Schuurbiers, Ibo van de Poel, and Michael E. Gorman, editors, *Early Engagement and New Technologies: Opening up the Laboratory*, Philosophy of Engineering and Technology, pages 55–95. Springer Netherlands, Dordrecht.

Rama Rohit Reddy Gangula, Suma Reddy Duggenpudi, and Radhika Mamidi. 2019. Detecting Political Bias in News Articles Using Headline Attention. In *Proceedings of the 2019 ACL Workshop*

BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 77–84, Florence, Italy. Association for Computational Linguistics.

Kiran Garimella and Simon Chauchard. 2025. Whatsapp explorer: A data donation tool to facilitate research on whatsapp. *Mobile Media & Communication*, 13(3):481–503.

Souvick Ghosh and Chirag Shah. 2019. Toward Automatic Fake News Classification. In *HICSS*.

Carol Gilligan. 1984. *In a different voice*. Harvard University, [Place of publication not identified]. Section: 174 pages.

Max Glockner, Ieva Staliūnaitė, James Thorne, Gisela Vallejo, Andreas Vlachos, and Iryna Gurevych. 2024. Ambifc: Fact-checking ambiguous claims with evidence. *Transactions of the Association for Computational Linguistics*, 12:1–18.

Antonio Gramsci. 1929/2020. Selections from the prison notebooks. In *The applied theatre reader*, pages 141–142. Routledge.

Leon Graumas, R Latulippe David, and Tommaso Caselli. 2019. Twitter-based Polarised Embeddings for Abusive Language Detection. *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pages 1–7.

K. Gupta. 2025. Comparative Analysis of Encoder-Based and Decoder-Based Architectures for Automatic Conspiracy Theory Identification. In *2025 International Conference on Computational, Communication and Information Technology (ICCCIT)*, pages 394–398.

Matthew J Hornsey and Samuel Pearson. 2022. Cross-national differences in willingness to believe conspiracy theories. *Current Opinion in Psychology*, 47:101391.

Dirk Hovy and Shannon L. Spruit. 2016. The Social Impact of Natural Language Processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Dirk Hovy and Diyi Yang. 2021. The importance of modeling social factors of language: Theory and practice. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human language technologies*, pages 588–602.

Jessica Jaiswal, Caleb LoSchiavo, and David C Perlman. 2020. Disinformation, misinformation and inequality-driven mistrust in the time of covid-19: lessons unlearned from aids denialism. *AIDS and Behavior*, 24(10):2776–2780.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Andreas Jungherr and Ralph Schroeder. 2021. Disinformation and the structural transformations of the public arena: Addressing the actual challenges to democracy. *Social Media+ Society*, 7(1):2056305121988928.

Theara Khoun, Ate Poortinga, Nyein Soe Thwal, Iván González de Alba, Andrea McMahon, and Carlos Mendez. 2025. Mapping the dimensions of poverty through big data, socioeconomic surveys and machine learning in cambodia. *Social Indicators Research*, 180(3):1593–1618.

Robert V Kozinets and Ulrike Gretzel. 2024. Netnography evolved: New contexts, scope, procedures and sensibilities. *Annals of Tourism Research*, 104:103693.

Kimberley Kruijver, Neill Bo Finlayson, Beatrice Cadet, and Sico van der Meer. 2025. The disinformation lifecycle: An integrated understanding of its creation, spread and effects. *Discover Global Society*, 3(1):58.

David La Barbera, G Milanese, Georgios Peikos, Gabriella Pasi, Marco Viviani, et al. 2025. Beyond binary classification: ranking for information access in misinformation contexts. In *CEUR WORKSHOP PROCEEDINGS*, volume 4121, pages 1–7. CEUR-WS.

Bruno Latour. 2012. *We have never been modern*. Harvard university press.

Johannes Lattmann. 2025. Detecting eu sentiment in texts: A llm machine learning application for euroscepticism research. *OSF*.

Jochen L Leidner and Vassilis Plachouras. 2017. Ethical by Design: Ethics Best Practices for Natural Language Processing. In *Proceedings of the First Workshop on Ethics in Natural Language Processing*, pages 30–40, Valencia, Spain. ACM.

Robert Litschko, Max Müller-Eberstein, Rob Van Der Goot, Leon Weber-Genzel, and Barbara

Plank. 2023. Establishing trustworthiness: Rethinking tasks and model evaluation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 193–203.

Erik Bran Marino, Jesus M Benitez-Baleato, and Ana Sofia Ribeiro. 2024. The polarization loop: How emotions drive propagation of disinformation in online media—the case of conspiracy theories and extreme right movements in southern europe. *Social Sciences*, 13(11):603.

Abraham Harold Maslow. 1966. *The psychology of science*. Harper and Row.

Hugo Mercier. 2020. *Not born yesterday: The science of who we trust and what we believe*. Princeton University Press.

Meta Transparency Center. 2024. Update on Meta's policy on the term 'Zionist'.

Todor Mihaylov and Preslav Nakov. 2016. Hunting for Troll Comments in News Community Forums. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 399–405, Berlin, Germany. Association for Computational Linguistics.

Mazda Moayeri, Elham Tabassi, and Soheil Feizi. 2024. Worldbench: Quantifying geographic disparities in llm factual recall. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1211–1228.

Evgeny Morozov. 2013. *To save everything, click here: The folly of technological solutionism*. PublicAffairs.

Guberney Muneton-Santa, Daniel Escobar-Grisales, Felipe Orlando López-Pabón, Paula Andrea Pérez-Toro, and Juan Rafael Orozco-Arroyave. 2022. Classification of poverty condition using natural language processing. *Social Indicators Research*, 162(3):1413–1435.

Luca Nannini, Eleonora Bonel, Davide Bassi, and Michele Joshua Maggini. 2025. Beyond phase-in: assessing impacts on disinformation of the eu digital services act. *AI and Ethics*, 5(2):1241–1269.

Philip M Napoli. 2019. *Social media and the public interest: Media regulation in the disinformation age*. Columbia university press.

Nic Newman, Richard Fletcher, Craig T. Robertson, Amy Ross Arguedas, and Rasmus Kleis Nielsen. 2024. Reuters institute digital news report 2024. Report, Reuters Institute for the Study of Journalism, Oxford, UK.

Ilana Nguyen, Harini Suresh, and Evan Shieh. 2025. Representational harms in llm-generated narratives against nationalities located in the global south. In *HEAL Workshop, CHI*, volume 2025.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*, WWW '16, pages 145–153, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Nel Noddings. 1984. *Caring: a feminine approach to ethics and moral education*. University of California Press.

Martha C Nussbaum. 2011. Creating capabilities: The human development approach and its implementation. *Hypatia*, 24(3):211–215.

Shrimai Prabhumoye, Brendon Boldt, Ruslan Salakhutdinov, and Alan W Black. 2021. Case Study: Deontological Ethics in NLP. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3784–3798, Online. Association for Computational Linguistics.

Žiga Puklavec, Christoph Kogler, Olga Stavrova, and Marcel Zeelenberg. 2023. What we tweet about when we tweet about taxes: A topic modelling approach. *Journal of Economic Behavior & Organization*, 212:1242–1254.

Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. 2018. Automatic Detection of Fake News. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3391–3401, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Amir H. Razavi, Diana Inkpen, Sasha Uritsky, and Stan Matwin. 2010. Offensive Language Detection Using Multi-level Classification. In *Advances in Artificial Intelligence*, pages 16–27, Berlin, Heidelberg. Springer.

Michael J Reddy. 1979. *The Conduit Metaphor: A Case of Frame Conflict in Our Language about Language*. Cambridge: Cambridge UP, Cambridge.

Timothy S Rich, Ian Milden, and Mallory Treece Wagner. 2020. Research note: Does the public support fact-checking social media? it depends who and how you ask. *The Harvard Kennedy School Misinformation Review*, 1(8).

Ronald E Robertson, Jon Green, Damian J Ruck, Katherine Ognyanova, Christo Wilson, and David Lazer. 2023. Users choose to engage with more partisan news than they are exposed to on google search. *Nature*, 618(7964):342–348.

Piero Ronzani. 2025. Towards the study of world misinformation. *Harvard Kennedy School Misinformation Review*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 1668–1678.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Miriam Schirmer, Nathan Walter, and Emőke-Ágnes Horvát. 2025. Disparities by design: Toward a research agenda that links science misinformation and socioeconomic marginalization in the age of ai. *Harvard Kennedy School Misinformation Review*.

David Schlangen. 2021. Targeting the benchmark: On methodology in current natural language processing research. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 670–674.

Merrill Singer. 2009. *Introduction to syndemics: A critical systems approach to public and community health*. John Wiley & Sons.

Carly Parnitzke Smith and Jennifer J Freyd. 2014. Institutional betrayal. *American psychologist*, 69(6):575.

Nadiya Straton, Hyeju Jang, and Raymond Ng. 2020. Stigma annotation scheme and stigmatized language detection in health-care discussions on social media. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1178–1190, Marseille, France. European Language Resources Association.

Li Qian Tay, Stephan Lewandowsky, Mark J Hurlstone, Tim Kurz, and Ullrich KH Ecker. 2024. Thinking clearly about misinformation. *Communications Psychology*, 2(1):4.

U.S. House of Representatives. 2023. H.res.894 - Strongly condemning and denouncing the drastic rise of antisemitism in the United States and around the world. Technical report, 118th Congress. Accessed: 2026-02-10.

Carolina Villegas-Galaviz and Kirsten Martin. 2023. Moral distance, AI, and the ethics of care. *AI & society*, pages 1–12.

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

Silvio Waisbord. 2018. Truth is what happens to news: On journalism, fake news, and post-truth. *Journalism studies*, 19(13):1866–1878.

Charlotte Ward and David Voas. 2011. The emergence of conspirituality. *Journal of Contemporary Religion*, 26(1):103–121.

Claire Wardle and Hossein Derakhshan. 2017. *Information disorder: Toward an interdisciplinary framework for research and policymaking*, volume 27. Council of Europe Strasbourg.

William Warner and Julia Hirschberg. 2012. Detecting Hate Speech on the World Wide Web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, pages 88–93.

Max Weber. 1949. *Max Weber on the methodology of the social sciences*. Free Press.

Jonathan F. Will. 2011. A Brief Historical and Theoretical Perspective on Patient Autonomy and Medical Decision Making. *Chest*, 139(3):669–673.

Shehel Yoosuf and Yin Yang. 2019. Fine-Grained Propaganda Detection with Fine-Tuned BERT. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 87–91, Hong Kong, China. Association for Computational Linguistics.