The Epistemic Limits of NLP Models in Media Bias Detection: Toward a Framework for Context-Aware and Reflexive AI Systems

Francisco-Javier Rodrigo-Ginés*, Jorge Chamorro-Padial, Pablo Rodríguez-Díaz
T-Systems Iberia
francisco-javier.rodrigo@t-systems.com

***Abstract*— Natural language processing (NLP) models are increasingly used to identify media bias, yet their capacity to capture the interpretive and contextual nature of bias remains fundamentally limited. Most systems conceptualize bias as a linguistic deviation detectable through statistical regularities, overlooking the socio-political and pragmatic dimensions that shape news discourse. This paper examines the epistemic limits of such models and proposes a theoretical framework for context-aware and reflexive AI systems. The framework integrates four epistemic dimensions: representation, contextual grounding, interpretive plurality, and governance. To characterize how meaning and bias co-evolve in communicative contexts. Building on this foundation, we introduce the notion of reflexive AI: systems capable of articulating the provenance, uncertainty, and assumptions underlying their own outputs. Rather than aiming to eliminate bias, reflexive systems make explicit the interpretive processes through which bias is constructed and recognized. This epistemic reframing advances media bias detection from a task of classification toward one of governance, providing a principled basis for designing transparent, accountable, and contextually grounded NLP systems.**

I. Introduction

The increasing reliance on Natural Language Processing (NLP) models to detect media bias reflects a broader societal trend toward the automation of interpretive and evaluative tasks [1]. Automated systems are now expected to distinguish between factual reporting and biased discourse, often in multilingual, high-volume environments [2]. Despite their technical sophistication, these systems encounter profound epistemological limitations when attempting to model a phenomenon that is not purely linguistic but socio-cognitive and interpretive in nature [3], [4].

Media bias is not a simple deviation from objectivity, but a multidimensional construct that involves framing, selection, omission, and attribution within specific cultural and historical contexts [5]. Such complexity challenges the underlying assumptions of most NLP architectures, which operate on the premise that meaning can be inferred from surface-level text features and statistical co-occurrences. These models, optimized for linguistic regularity, often mistake correlation for causation and overlook the hermeneutic dimension of discourse, the way meaning is shaped through perspective, ideology, and intention [6], [7].

This epistemic gap has practical and ethical consequences. Systems that aim to detect bias without acknowledging their own interpretive standpoint risk reproducing or amplifying the very asymmetries they intend to mitigate. Bias detection models trained on partial datasets or homogeneous annotator populations may inadvertently encode culturally specific judgments as universal norms [8]. The epistemological opacity of such systems, where decisions appear objective but rest on unexamined assumptions, raises concerns for both trust and accountability in automated media analysis [9], [10].

This paper argues that these limitations are not merely technical shortcomings but epistemic constraints inherent to the current paradigm of data-driven NLP. It proposes that meaningful progress in media bias detection requires reframing the task as a problem of epistemic modeling rather than classification. Specifically, we contend that detecting bias necessitates the representation of contextual knowledge across four interdependent dimensions: linguistic, discursive, cultural-historical, and pragmatic.

To support this claim, the paper develops a conceptual framework that models these dimensions as layers of epistemic context. This framework does not aim to replace quantitative NLP methods but to complement them by clarifying their epistemic boundaries and suggesting avenues for reflexive, context-aware AI systems. Such systems would explicitly encode uncertainty, interpretive diversity, and source traceability; elements essential to the development of responsible and transparent media analysis tools [11], [12].

The contributions of this paper are threefold:

1) It articulates the epistemological limits of current NLP approaches to media bias detection, highlighting the reductionist assumptions underlying linguistic modeling.
2) It proposes a multidimensional framework for representing epistemic context in bias analysis, integrating insights from linguistics, philosophy of science, and communication theory.
3) It outlines principles for designing reflexive and context-aware AI systems that recognize their own interpretive constraints and make such limitations transparent to human users.

The remainder of this paper is organized as follows. Section II discusses the epistemological nature of bias and its incompatibility with purely statistical NLP methods. Section III presents the proposed multidimensional framework of epistemic context. Section IV explores implications for the design of reflexive AI systems. Section V concludes with reflections on how acknowledging epistemic limits can foster

*Corresponding author

trust, interpretability, and accountability in automated media analysis.

II. The Epistemological Problem of Bias Detection

The task of detecting media bias has often been approached as a problem of classification: given a text, an algorithm must determine whether it exhibits bias or not. This framing presupposes that bias is a stable, observable property of linguistic data. However, from an epistemological standpoint, bias cannot be treated as an inherent attribute of language but as a relational construct that emerges from the interaction between text, producer, and audience [13], [6]. As such, it is not directly observable in the same way as syntactic or semantic patterns; it depends on interpretation, perspective, and socio-historical positioning.

A. Bias as an Interpretive Construct

In traditional epistemology, knowledge is justified belief formed under conditions of evidence and inference. When applied to discourse, these conditions are mediated by cultural frameworks and communicative intentions. Media bias represents a deviation not from truth, but from neutrality as perceived within a given interpretive community [7]. It is a phenomenon of framing and emphasis rather than falsification. Entman's seminal theory defines framing as the selection of certain aspects of perceived reality to make them more salient, thereby promoting a particular interpretation [5].

From this perspective, bias is not a falsifiable statement about facts but a metalinguistic operation on how meaning is structured and received. This implies that bias cannot be exhaustively captured through surface-level linguistic cues or frequency distributions. What is often measured by NLP systems are statistical regularities of word choice, not the interpretive structures that organize meaning. This distinction marks the core epistemological limitation of bias detection by automated means.

B. The Reductionism of Data-Driven Models

Contemporary NLP systems, particularly those based on large-scale pretraining, rely on statistical induction: the approximation of meaning through patterns of co-occurrence. This approach assumes that truth conditions are inferable from textual form, effectively collapsing epistemology into correlation. Such reductionism reflects what Floridi terms the "data-centric fallacy", the idea that informational abundance compensates for conceptual inadequacy [11].

In practice, data-driven models exhibit three structural blind spots. First, they operate in decontextualized environments where cultural and pragmatic signals are either absent or encoded as noise. Second, they conflate representational regularity with epistemic validity, treating frequent forms as reliable indicators of truth. Third, they lack mechanisms for reflexivity: the capacity to represent and question their own epistemic boundaries. Consequently, these models may reproduce ideological asymmetries embedded in the data, normalizing them as linguistic conventions [14].

Attempts to mitigate these effects through debiasing or fairness constraints often address symptoms rather than causes. They correct imbalances in label distribution but leave untouched the deeper issue: the epistemic model underlying NLP remains ill-suited for interpretive phenomena such as bias.

C. Toward an Epistemology of Context

To advance beyond the limitations of current NLP paradigms, it is necessary to recognize that bias operates across multiple layers of context: linguistic, discursive, cultural, and pragmatic. Each layer introduces interpretive dependencies that cannot be collapsed into a single statistical representation. Meaning, in this view, is not a property of data but a relation among situated agents within communicative practices [15].

This insight connects with Popper's principle of falsifiability, which holds that scientific propositions must be framed such that they can, in principle, be refuted [16]. Bias detection systems, however, often operate under non-falsifiable assumptions: they presume that "bias" can be defined and labeled independently of the observer. This renders their epistemic claims circular. A reflexive model of media analysis must instead treat interpretive disagreement not as annotation error, but as a form of epistemic data, an expression of plural perspectives within a contested social reality.

Recognizing these epistemological limits is not a rejection of NLP but a redefinition of its scope. Automated methods can support bias analysis only when embedded within frameworks that make their interpretive assumptions explicit, measurable, and contestable. Such frameworks move from the illusion of objectivity toward the design of transparent, accountable, and context-aware AI systems.

III. Dimensions of Epistemic Context in Media Bias

If bias is not a discrete linguistic feature but a relational construct embedded in social communication, its detection depends on modeling the epistemic context in which meaning is produced and interpreted. To clarify this dependency, we propose a multidimensional framework comprising four interrelated dimensions of epistemic context: linguistic, discursive, cultural-historical, and pragmatic. Each dimension corresponds to a distinct level of abstraction in the generation and interpretation of meaning, and each introduces constraints that challenge the assumptions of conventional NLP systems.

A. Linguistic Dimension: Lexical and Syntactic Framing

The linguistic dimension encompasses the micro-level features through which bias manifests: lexical choice, syntactic structure, and semantic association. At this level, bias is often expressed through evaluative adjectives, agent-patient asymmetries, or presuppositional constructions. NLP models

are most effective here, as such phenomena are textually observable and quantifiable [17]. Beyond lexical or syntactic cues, recent studies have explored the use of persuasive techniques as operational proxies for bias manifestation [18]. However, linguistic indicators alone cannot disambiguate whether a lexical choice reflects authorial bias, stylistic convention, or contextual necessity. Without external reference, frequency-based methods conflate expression with intention. This dimension is thus necessary but insufficient for epistemic modeling of bias.

B. Discursive Dimension: Narrative and Ideological Structure

The discursive dimension refers to how texts organize meaning across sentences and genres, forming coherent narratives that promote particular interpretations of reality. Discourse operates through selection, omission, and framing, what Entman described as "the salience of aspects of perceived reality" [5]. Bias at this level arises not from isolated lexical choices but from systematic asymmetries in representation, such as the prioritization of certain actors, sources, or causal explanations [6]. Computationally, this dimension corresponds to macro-level features such as topic distribution, argumentation structure, and narrative framing. Yet current NLP models rarely integrate such structures explicitly, relying instead on token-level embeddings that obscure ideological organization. Capturing discursive bias requires models that represent not only words but the inferential relations among them.

C. Cultural-Historical Dimension: Situated Knowledge and Shared Memory

Bias is also conditioned by the socio-historical environment in which communication occurs. The cultural-historical dimension captures how shared narratives, collective memories, and ideological schemas shape the interpretation of media content [19]. What constitutes bias in one cultural context may be perceived as neutrality in another. This relativity challenges the universalist orientation of most machine learning pipelines, which presume that ground truth labels are stable across annotators and contexts. Moreover, datasets often lack metadata about the temporal, political, or cultural conditions of text production, rendering them epistemically incomplete. Integrating this dimension requires encoding provenance, temporality, and socio-political context as part of model inputs or evaluation protocols, aligning with emerging approaches to data documentation and contextual AI governance [3], [12].

D. Pragmatic Dimension: Intention and Reception

Finally, the pragmatic dimension concerns the interactional context of communication, the intentions of the author and the interpretive stance of the audience. From a pragmatic perspective, bias cannot be inferred solely from text, because meaning is co-constructed through use. Wittgenstein's notion that "meaning is use" underscores the dependency of interpretation on communicative purpose [15]. A statement that appears biased in isolation may be ironic, rhetorical, or counterfactual within its communicative setting. Modeling this dimension requires systems capable of representing speaker intent, social roles, and audience expectations, potentially through multi-agent or simulation-based approaches [20]. This is the least tractable dimension computationally, yet the most critical for avoiding category errors in automated bias detection.

E. Synthesis: The Epistemic Context Model

Table I summarizes the four dimensions and their implications for computational modeling. Together, they define an epistemic space within which bias operates as an emergent property rather than a measurable attribute. The closer a system approaches the pragmatic and cultural-historical layers, the greater its need for contextual knowledge and reflexivity. Conversely, systems limited to the linguistic layer remain constrained to descriptive analytics. This framework thus provides a theoretical basis for designing reflexive and context-aware AI systems that acknowledge the partiality of their own representations.

IV. Toward Reflexive and Context-Aware AI

The four-dimensional framework outlined in Section III implies that responsible media bias detection cannot rely solely on textual regularities but must integrate awareness of its epistemic conditions. This requirement motivates the concept of reflexive and context-aware AI: systems that explicitly represent the limitations, provenance, and interpretive assumptions underlying their own outputs. Unlike traditional explainable AI, which focuses on transparency after decision-making, reflexive AI embeds epistemic self-awareness into the inference process itself [21].

A. Defining Reflexivity in AI Systems

Reflexivity, in this context, refers to an AI system's capacity to model the conditions of its own knowledge production. A reflexive system does not merely output predictions but includes meta-level information about the epistemic scope and confidence of those predictions. This includes:

- **Provenance awareness:** explicit encoding of data origin, temporal range, and annotator diversity;
- **Uncertainty expression:** calibrated confidence measures that reflect epistemic rather than purely statistical uncertainty [22];
- **Interpretive plurality:** capacity to represent multiple plausible interpretations of the same input, rather than enforcing a single authoritative label [23];
- **Self-limitation:** explicit declaration of contexts in which the system's inference is likely invalid or incomplete.

Such systems align with the broader paradigm of epistemic humility in AI design, which emphasizes the articulation of boundaries over the illusion of omniscience. Reflexive AI thus represents a shift from optimizing predictive accuracy to managing interpretive validity.

TABLE I

FOUR DIMENSIONS OF EPISTEMIC CONTEXT IN MEDIA BIAS AND THEIR IMPLICATIONS FOR NLP MODELING.

Dimension	Epistemic Definition	Implication for NLP
Linguistic	Observable textual features such as lexical choice, syntax, and sentiment.	Enables pattern-based bias indicators but lacks interpretive grounding.
Discursive	Narrative, framing, and causal organization of meaning across text.	Requires macro-level modeling of coherence and ideology.
Cultural-Historical	Shared memory, ideology, and socio-political situatedness of communication.	Demands contextual metadata and provenance-aware datasets.
Pragmatic	Intention, reception, and communicative function of discourse acts.	Calls for models integrating social roles, purpose, and interactional feedback.

B. Architectural Principles for Context-Aware Systems

To operationalize reflexivity, we propose three complementary design principles: contextualization, pluralization, and traceability. These principles can be instantiated as modular components in NLP pipelines or LLM-based architectures.

(1) Contextualization. Models must condition their outputs on metadata describing communicative and sociopolitical context. This includes source origin, publication date, and known ideological orientation. Context conditioning can be achieved through structured prompts, context embeddings, or dynamic knowledge retrieval modules [24]. The aim is not to "de-bias" language models but to situate their inferences relative to context.

(2) Pluralization. Interpretive diversity should be preserved rather than collapsed into a single prediction. Instead of enforcing consensus labels, systems can produce ensembles of interpretations weighted by annotator ideology, linguistic variety, or cultural perspective. From a perspectivist standpoint, such multiplicity reflects the coexistence of partial yet complementary viewpoints that together enrich epistemic validity [25]. This approach aligns with recent developments in learning with disagreements (LeWiDi), which treat annotator divergence not as noise but as a signal of inherent subjectivity in the data [26]. By formalizing disagreement as structured evidence, pluralization mechanisms can be implemented through multi-agent architectures, where independent reasoning agents generate and negotiate alternative perspectives. This approach transforms disagreement from noise into epistemic signal [9].

(3) Traceability. Each decision must be linked to an interpretable causal and contextual chain: which data influenced the output, which assumptions were applied, and under what contextual boundaries. Traceability aligns with standards such as IEEE 7001 on transparency in autonomous systems [12] and recent efforts in algorithmic accountability emphasizing explainability and bias mitigation [27]. At the computational level, this can be implemented through provenance graphs or structured model cards documenting contextual dependencies.

These principles formalize reflexivity as an operational property of AI systems: they make explicit what a system knows, under which conditions it knows it, and when it cannot know. This represents an epistemic counterpart to conventional performance optimization.

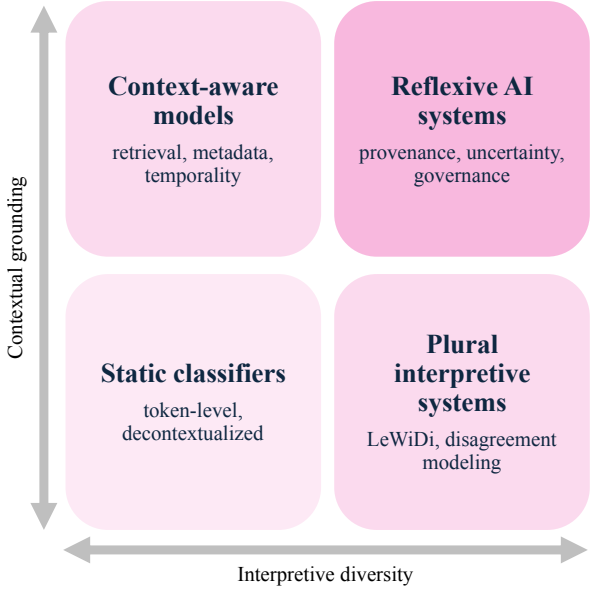

Fig. 1. Epistemic continuum of interpretive diversity and contextual grounding in media bias analysis.

C. From Explainability to Epistemic Governance

While XAI aims to make model decisions intelligible, reflexive AI extends this goal toward epistemic governance: the systematic management of how, why, and under what assumptions knowledge claims are produced by machines [28]. In the domain of media bias detection, this shift reframes the system from a detector of deviance to a mediator of interpretation. The system's function becomes not to declare texts as biased or unbiased, but to expose the interpretive frameworks through which such judgments emerge.

Reflexive architectures also enable institutional accountability. When systems record interpretive pathways and confidence profiles, human overseers can audit not only outcomes but epistemic processes. This creates a verifiable link between algorithmic behavior and epistemic responsibility, an essential feature for trustworthy AI in communicative domains [29].

Ultimately, the transition from static classification to reflexive interpretation marks a paradigm shift in NLP research. It acknowledges that understanding bias requires not only detecting linguistic signals but modeling the social

processes through which meaning and perspective are co-constructed. Reflexive AI does not aspire to eliminate bias, but to render its conditions of possibility explicit and contestable.

V. Discussion and Implications

The proposed framework for reflexive and context-aware AI expands the analytical scope of media bias detection from linguistic regularities to epistemic structures. Its implications extend beyond the specific domain of NLP, raising broader questions about how artificial systems produce, justify, and communicate knowledge. This section discusses three domains of consequence: epistemological, methodological, and institutional.

A. Epistemological Implications: From Objectivity to Reflexivity

Traditional computational models implicitly adopt a positivist epistemology, assuming that bias can be measured as deviation from a neutral linguistic baseline. The reflexive framework challenges this assumption by reframing objectivity as a relational construct rather than an intrinsic property of data. Bias is not eliminated through algorithmic refinement; it is situated, negotiated, and context-dependent [30].

This shift aligns with post-positivist conceptions of knowledge in science and technology studies, where understanding emerges through reflexive engagement with uncertainty rather than its elimination [31]. The epistemic task for AI systems, therefore, is not to discover objective truth, but to represent the interpretive processes that give rise to competing truths. Reflexive AI recognizes that interpretive diversity is not noise but an indicator of epistemic vitality within a pluralistic information ecosystem.

B. Methodological Implications: Evaluating Interpretive Systems

From a methodological perspective, the framework suggests that standard evaluation metrics (accuracy, F1-score, or fairness indices) are insufficient to assess systems operating under epistemic plurality. Reflexive systems require new forms of validation that measure interpretive alignment, uncertainty calibration, and contextual transparency [22].

Future evaluation protocols could incorporate measures such as:

- **Epistemic alignment:** quantifying the degree of overlap between model-generated explanations and human interpretive rationales;
- **Interpretive diversity:** assessing whether the system preserves or suppresses alternative perspectives in its output;
- **Context traceability:** evaluating the granularity and completeness of provenance metadata.

Such metrics shift the focus from predictive success to epistemic responsibility. This reorientation parallels the evolution of interpretability research, where the goal is not explanation alone but the justification of model behavior within a broader normative framework [32]. In this sense, reflexive AI demands evaluation methodologies that are as much hermeneutic as statistical.

C. Institutional Implications: Epistemic Governance in AI Systems

At the institutional level, reflexive architectures provide a foundation for epistemic governance: the deliberate management of how knowledge claims are generated, contextualized, and audited within AI systems. Governance mechanisms must address not only technical performance but epistemic legitimacy, the degree to which automated analyses are recognized as credible within their social contexts [33].

Embedding reflexivity into design aligns with current regulatory trends emphasizing traceability, human oversight, and explainability [29]. However, the framework also points beyond compliance: it envisions organizations as hybrid epistemic entities where humans and machines co-produce meaning. This requires institutional infrastructures for continuous auditing, plural review, and interpretive dialogue between system designers, users, and affected stakeholders.

Reflexive AI thus redefines the relationship between automation and responsibility. Instead of delegating judgment to algorithms, it distributes interpretive agency across socio-technical assemblages. In doing so, it transforms media analysis from a process of detection into one of negotiation, where the transparency of epistemic assumptions becomes as important as the accuracy of results.

D. Limitations and Future Perspectives

The proposed framework remains conceptual and requires empirical validation. Implementing reflexive architectures will involve trade-offs between interpretive fidelity and computational efficiency, as well as new challenges for scalability and annotation design. Nevertheless, the theoretical basis established here provides a foundation for experimental inquiry.

Future work should explore how reflexivity can be operationalized through: (i) multimodal representations combining linguistic and contextual metadata; (ii) interactive systems that allow human reviewers to interrogate and adjust model assumptions; (iii) agent-based simulations that model interpretive negotiation among heterogeneous AI agents; and (iv) longitudinal studies evaluating how reflexive transparency influences user trust and interpretive behavior over time.

Ultimately, the transition toward reflexive AI is not a purely technical innovation but an epistemic reorientation. It calls for AI systems that not only perform linguistic analysis but understand and communicate the limits of their own understanding.

VI. Conclusion

This paper has examined the epistemological limitations of current NLP models in detecting media bias and proposed a framework for reflexive and context-aware AI. The central argument is that bias cannot be exhaustively represented as a linguistic or statistical property because it emerges from

interpretive processes distributed across linguistic, discursive, cultural, and pragmatic dimensions. As such, systems that treat bias as a measurable deviation from neutrality operate under epistemic reductionism: they quantify expression without modeling interpretation.

The proposed framework contributes to the scientific understanding of AI in three main ways. First, it introduces a multidimensional model of epistemic context that formalizes how meaning and bias co-evolve across levels of communication. Second, it defines the concept of reflexive AI as a class of systems capable of representing their own epistemic assumptions, provenance, and interpretive uncertainty. Third, it advances the notion of epistemic governance, the deliberate management of how knowledge is produced, contextualized, and audited within AI systems.

Recognizing bias as a relational and situated phenomenon does not preclude automation but redefines its purpose. Reflexive AI shifts the objective from detecting bias to explicating the interpretive processes that generate it. This reorientation transforms the problem from classification to governance, where the epistemic transparency of AI systems becomes a measurable design parameter.

Future research should pursue three directions. (i) Empirical studies to evaluate how reflexive architectures affect user trust, interpretive diversity, and accountability. (ii) Development of benchmarks that assess epistemic metrics, such as interpretive alignment and contextual traceability. (iii) Exploration of hybrid human–AI systems where interpretive negotiation is modeled as a dynamic process.

Ultimately, the framework presented here reframes the epistemological foundation of NLP for media analysis. It calls for AI systems that do not merely analyze language but participate responsibly in the co-production of meaning. Such systems would embody an epistemic ethics: the capacity to know, and to know the limits of what they know.

REFERENCES

[1] F.-J. Rodrigo-Ginés, "Automated Media Bias Detection: Challenges and Opportunities," in *Proceedings of the Doctoral Symposium on Natural Language Processing from the Proyecto ILENIA (PLN-DS-2023) held as part of the XXXIX edition of the International Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), Jaén, Spain, September 28, 2023* (M. T. Martín-Valdivia, E. Martínez-Cámara, M. D. Molina-González, and S. M. J. Zafra, eds.), vol. 3625 of *CEUR Workshop Proceedings*, pp. 86–94, CEUR-WS.org, 2023.

[2] S. K. Hamed, M. J. Ab Aziz, and M. R. Yaakub, "A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion," *Heliyon*, vol. 9, p. e20382, Oct. 2023.

[3] E. M. Bender and B. Friedman, "Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science," *Transactions of the Association for Computational Linguistics*, vol. 6, pp. 587–604, Dec. 2018.

[4] B. Mittelstadt, "Principles alone cannot guarantee ethical AI," *Nature Machine Intelligence*, vol. 1, pp. 501–507, Nov. 2019.

[5] R. M. Entman, "Framing: Toward Clarification of a Fractured Paradigm," *Journal of Communication*, vol. 43, pp. 51–58, Dec. 1993.

[6] T. A. Van Dijk, *Discourse and power*. Bloomsbury Publishing, 2017.

[7] T. Mccarthy, ed., *The Theory of Communicative Action*. Wiley, 1991.

[8] M. Sap, D. Card, S. Gabriel, Y. Choi, and N. A. Smith, "The Risk of Racial Bias in Hate Speech Detection," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, (Florence, Italy), pp. 1668–1678, Association for Computational Linguistics, 2019.

[9] S. Cranefield, N. Oren, and W. W. Vasconcelos, "Accountability for Practical Reasoning Agents," in *Agreement Technologies* (M. Lujak, ed.), vol. 11327, pp. 33–48, Cham: Springer International Publishing, 2019. Series Title: Lecture Notes in Computer Science.

[10] F.-J. Rodrigo-Ginés, J. Carrillo-de Albornoz, and L. Plaza, "A systematic review on media bias detection: What is media bias, how it is expressed, and how to detect it," *Expert Systems with Applications*, vol. 237, p. 121641, Mar. 2024.

[11] L. Floridi, *The logic of information: A theory of philosophy as conceptual design*. Oxford University Press, 2019.

[12] "Ieee standard for transparency of autonomous systems," *IEEE Std 7001-2021*, pp. 1–54, 2022.

[13] R. Capurro, "Hermeneutics and the phenomenon of information," *Metaphysics, epistemology, and technology. Research in philosophy and technology*, vol. 19, pp. 79–85, 2000.

[14] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, (Virtual Event Canada), pp. 610–623, ACM, Mar. 2021.

[15] L. Wittgenstein, *Philosophical investigations*. Blackwell, 1953.

[16] K. Popper, *The logic of scientific discovery*. Routledge, 1959.

[17] M. Recasens, C. Danescu-Niculescu-Mizil, and D. Jurafsky, "Linguistic models for analyzing and detecting biased language," in *Proceedings of the 51st annual meeting of the Association for Computational Linguistics (volume 1: long papers)*, pp. 1650–1659, 2013.

[18] F. J. Rodrigo-Ginés, J. Carrillo-de Albornoz, and L. Plaza, "Identifying media bias beyond words: Using automatic identification of persuasive techniques for media bias detection," *Procesamiento del Lenguaje Natural*, vol. 71, pp. 179–190, 2023.

[19] J. Assmann, *Cultural memory and early civilization: Writing, remembrance, and political imagination*. Cambridge University Press, 2011.

[20] M. Wooldridge, *An introduction to multi-agent systems*. Wiley, 2021.

[21] J. Clune, "AI-GAs: AI-generating algorithms, an alternate paradigm for producing general artificial intelligence," Feb. 2020. arXiv:1905.10985 [cs].

[22] A. Niculescu-Mizil and R. Caruana, "Predicting good probabilities with supervised learning," in *Proceedings of the 22nd international conference on Machine learning*, pp. 625–632, 2005.

[23] A. Davani, M. Diaz, D. Baker, and V. Prabhakaran, "D3code: Disentangling disagreements in data across cultures on offensiveness detection and evaluation," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18511–18526, 2024.

[24] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, *et al.*, "Retrieval-augmented generation for knowledge-intensive nlp tasks," *Advances in neural information processing systems*, vol. 33, pp. 9459–9474, 2020.

[25] S. Mitchell, *Unsimple Truths: Science, Complexity, and Policy*. Chicago, IL: University of Chicago Press, 2009.

[26] S. Uma, S. Suresh, A. B. Arjun, and P. K. Maji, "Learning with disagreements," in *Proc. 2021 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3475–3485, 2021.

[27] S. Ali, T. Abuhmed, S. El-Sappagh, K. Muhammad, J. M. Alonso-Moral, R. Confalonieri, R. Guidotti, J. Del Ser, N. Díaz-Rodríguez, and F. Herrera, "Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence," *Information fusion*, vol. 99, p. 101805, 2023.

[28] L. Floridi, "Translating principles into practices of digital ethics: Five risks of being unethical," *Philosophy & Technology*, vol. 32, pp. 185–193, 2019.

[29] NIST, "Artificial intelligence risk management framework (ai rmf 1.0)," *URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai*, pp. 100–1, 2023.

[30] S. Jasanoff, "The idiom of co-production," in *States of Knowledge: The Co-Production of Science and Social Order*, Routledge, 2004.

[31] H. Collins and R. Evans, *Rethinking expertise*. University of Chicago press, 2019.

[32] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[33] V. Dignum, *Responsible artificial intelligence: how to develop and use AI in a responsible way*, vol. 2156. Springer, 2019.