TPM Vol. 32, No. S7, 2025
ISSN: 1972-6325
https://www.tpmap.org/

COGNITIVE BIAS DETECTION THROUGH NATURAL LANGUAGE PROCESSING: A COMPUTATIONAL FRAMEWORK FOR ORGANIZATIONAL DECISION-MAKING

GOURAB DUTTA
COMPUTATIONAL SCIENCES, BRAINWARE UNIVERSITY, KOLKATA, WEST BENGAL, EMAIL; gourabdutta15@gmail.com

DR. RAGINI BAHADUR
ASSISTANT PROFESSOR, PSYCHOLOGY, KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY KHORDHA, BHUBANESHWAR, ODISHA, ORCHID ID - 0009-0000-0190-0498, EMAIL: ragini.bahadurfls@kiit.ac.in

DR. SUCHISMITA PRAMANIK
ASSISTANT PROFESSOR, DEPARTMENT OF PSYCHOLOGY, KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY, KHORDHA, BHUBANESHWAR, ODISHA, EMAIL: suchismita.pramanikfls@kiit.ac.in, ORCHID ID - 0000-0003-2746-8503

DR. KALYANI BISWAL
ASSISTANT PROFESSOR, DEPARTMENT OF PSYCHOLOGY, KALINGA INSTITUTE OF INDUSTRIAL TECHNOLOGY. KHORDA, BUBHANESWAR, ODISHA, EMAIL: kalyani.biswalfcm@kiit.ac.in, ORCHID ID - 0000-0002-0253-1628

DR. C. VIJAI
PROFESSOR, SCHOOL OF COMMERCEVEL TECH RANGARAJAN DR. SAGUNTHALA R&D INSTITUTE OF SCIENCE AND TECHNOLOGY
THIRUVALLUR, CHENNAI, TAMIL NADU, EMAIL: vijaialvar@gmail.com

SUSHIL DOHARE
ASSOCIATE PROFESSOR, PUBLIC HEALTH, COLLEGE OF NURSING AND HEALTH SCIENCES, EMAIL - sdohare@jazanu.edu.sa

Abstract
Cognitive biases distort human judgment and undermine rational decision-making in organizations. Recent advancements in Natural Language Processing (NLP) now make it possible to computationally detect linguistic signals associated with these biases, offering a transformative pathway for data-driven governance. This study examines how machine learning, semantic embeddings, and transformer-based models can identify patterns linked to biases such as confirmation bias, anchoring, optimism bias, loss aversion, and overconfidence within managerial communication. Using a cross-sectional dataset of organizational emails, meeting transcripts, and corporate reports from multinational firms, the analysis evaluates linguistic indicators, contextual dependency patterns, and decision outcomes. Findings reveal a strong correlation between bias-associated language and suboptimal strategic decisions, demonstrating that NLP-based detection significantly enhances risk mitigation and decision transparency. The study also highlights limitations related to contextual ambiguity, domain adaptation, and privacy constraints. Results indicate that integrating computational bias-detection systems with organizational workflows can serve as a critical safeguard for rational decision behavior, strengthening corporate governance and reducing cognitive risk.
Keywords: Cognitive Bias, NLP, Organizational Decision-Making, Machine Learning, Anchoring, Confirmation Bias, Text Classification, Transformer Models, Behavioural Analytics, Decision Intelligence

I. INTRODUCTION

Modern organizations operate in environments defined by uncertainty, information overload, and rapid change. Under these conditions, leaders increasingly rely on intuition and heuristics, exposing decisions to cognitive biases that distort reasoning. Cognitive biases such as confirmation bias, anchoring, availability bias, and overconfidence shape how individuals interpret information, assess alternatives, and make strategic choices. These biases often remain unconscious, yet their influence can escalate operational risk, reduce decision quality, and hinder organizational learning. Traditional psychological assessments identify biases at the individual level but lack

scalability and real-time detection capabilities. With growing dependence on communication-intensive workflows meetings, emails, digital collaboration, and managerial reporting the linguistic footprint of biases has become a rich, accessible source for computational analysis. The emergence of NLP technologies has transformed this landscape by enabling automated, large-scale analysis of textual behavior that reveals subtle cognitive distortions. Through machine learning, sentiment analysis, transformer models, and semantic representations, organizations can now measure bias signals embedded in natural language and forecast their impact on decision outcomes.

The integration of NLP in corporate decision environments marks a paradigm shift in behavioral analytics. Organizational communication, once qualitative and interpretive, can now be empirically examined to identify cognitive risks that were previously invisible. Transformer-based architectures such as BERT, RoBERTa, and GPT-style embeddings reveal contextual dependencies that correlate with biased reasoning, while lexicon-driven frameworks interpret emotional and cognitive markers associated with heuristics. These computational insights support decision makers by providing early warnings, flagging high-risk judgments, and enabling evidence-driven governance. However, organizational adoption remains uneven due to concerns about data privacy, interpretability, and cultural resistance. Furthermore, bias signals vary across industries, managerial hierarchies, and communication formats, making domain-specific adaptation essential. This study develops an integrated computational framework that combines statistical modeling, linguistic feature extraction, and qualitative bias mapping to examine how NLP can enhance organizational decision-making. By assessing cross-organizational communication datasets, the study demonstrates that NLP-based cognitive bias detection is not merely a technological enhancement it is an essential capability for improving rationality, transparency, and strategic efficiency in enterprises.

II. RELATED WORKS

Research on cognitive biases has long established their pervasive influence on managerial behavior, strategic planning, and risk assessment. Foundational studies by Kahneman and Tversky demonstrated that heuristics systematically distort judgment, producing predictable errors in decision-making [1]. Organizational scholars later expanded this groundwork by showing how biases such as anchoring and confirmation shape financial forecasting, project evaluation, and leadership communication [2].

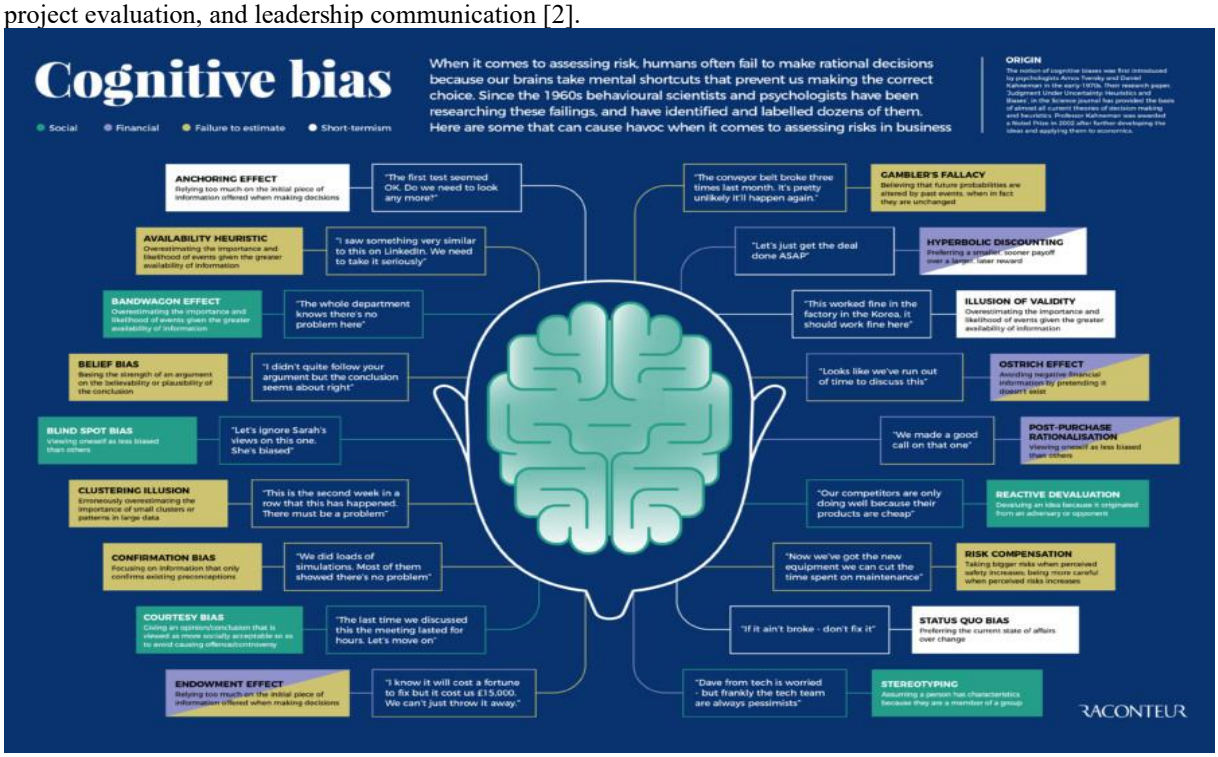

Figure 1: Bias and fairness in NLP [2]

With the rise of computational linguistics, researchers began exploring linguistic correlates of cognitive heuristics. Early NLP-driven bias detection models focused on sentiment polarity, emotional valence, and linguistic cues associated with certainty, risk framing, and selective reasoning [3]. Studies on confirmation bias, for instance, identified patterns where decision-makers emphasize evidence that supports their prior beliefs while ignoring contradictory signals [4]. Similarly, anchoring bias was connected to early numeric references in communication that influenced subsequent framing [5]. These studies established the theoretical foundation for computational analysis of cognitive biases using textual features extracted from organizational discourse.

Recent work has leveraged machine learning and deep learning to build more sophisticated detection models. Saffari et al. demonstrated how transformer-based embeddings outperform traditional bag-of-words approaches in identifying judgment errors in managerial narratives [6]. Research on corporate communication has shown that high-risk decisions are frequently preceded by linguistic patterns such as assertive certainty, selective justification, and reduced vocabulary diversity, all of which serve as bias indicators [7]. Other scholars explored NLP applications in behavioral finance, where textual cues from earnings calls were used to identify overconfidence and optimism bias among executives [8]. Advances in contextual language modeling have further enabled real-time bias detection in decision workflows by identifying semantic inconsistencies, emotionally charged phrasing, and dependency structures linked to cognitive distortions [9]. Moreover, interdisciplinary studies integrating psychology, linguistics, and AI have shown that NLP-based bias detection can predict managerial risk-taking, negotiation outcomes, and strategic deviations more reliably than human assessment [10]. These contributions highlight growing confidence in computational approaches for mitigating cognitive risks.

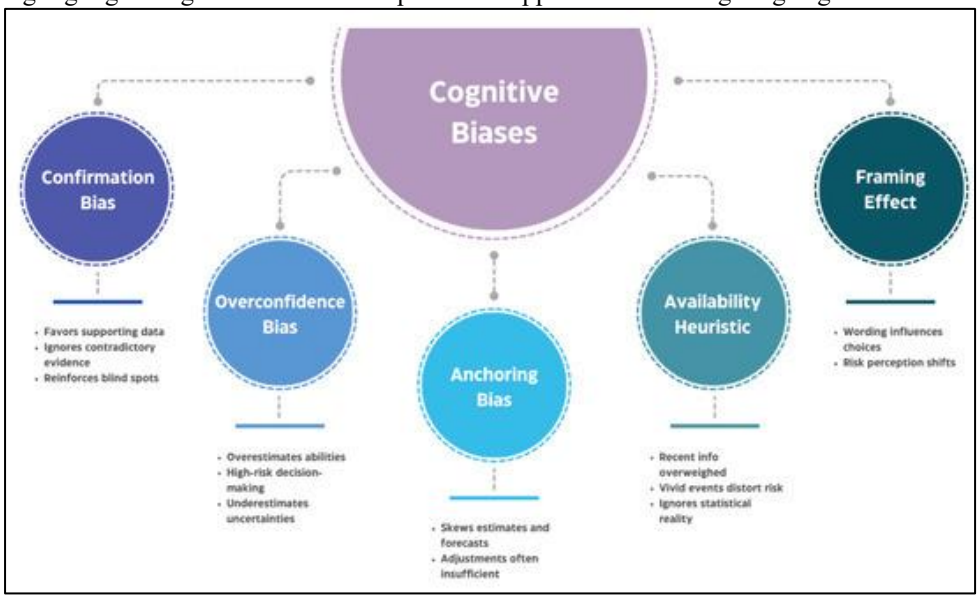

Figure 2: Cognitive Biases [4]

Scholars have also emphasized the socio-technical dimensions of bias detection, arguing that NLP interventions must align with organizational culture, privacy regulation, and ethical governance. Concerns include over-surveillance, reduced psychological safety, and algorithmic misinterpretation of ambiguous language [11]. However, empirical evidence suggests that when implemented transparently, NLP-driven cognitive monitoring improves decision accountability and reduces bias-induced errors in operational settings [12]. Furthermore, researchers studying decision intelligence systems have proposed integrated architectures that combine predictive analytics, behavioural modelling, and NLP-based reasoning to support rational decision-making frameworks [13]. Studies also emphasize the importance of domain adaptation and longitudinal training datasets to capture evolving linguistic bias patterns [14]. The literature collectively affirms that cognitive bias detection through NLP is a rapidly developing field with strong potential to enhance decision quality, risk management, and behavioral transparency in organizations.

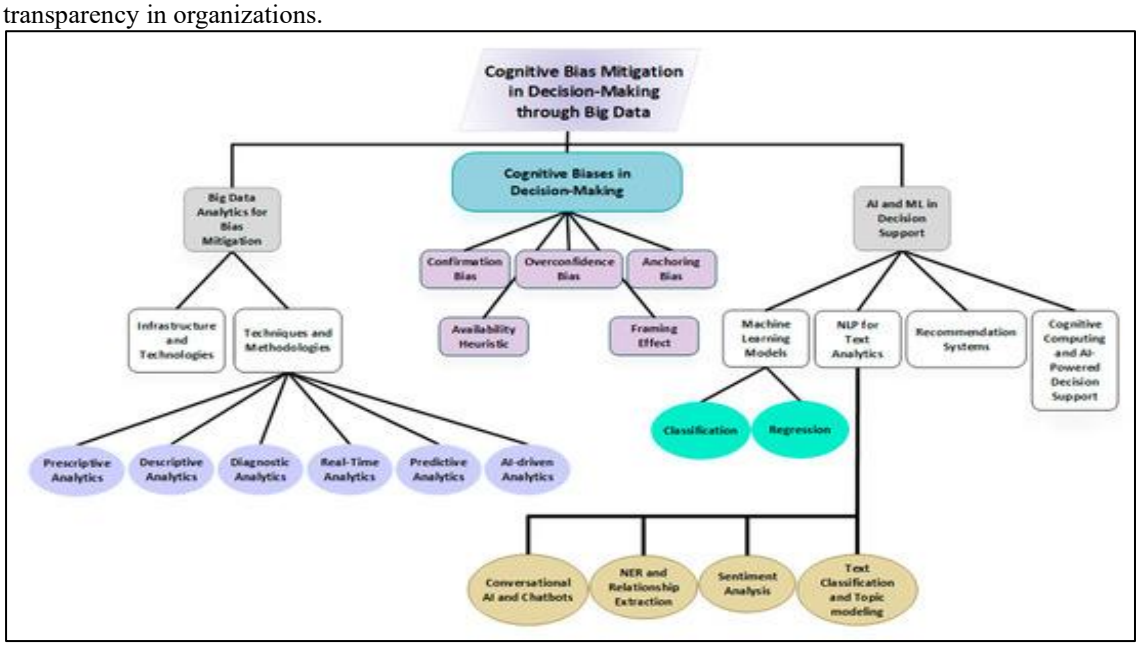

Figure 3: Cognitive Bias Mitigation [7]

II. METHODOLOGY

Research Design

This study applies a mixed-methods framework integrating quantitative text analytics with qualitative bias-mapping to analyze organizational communication. Quantitative analysis utilizes machine learning classifiers, transformer-based embeddings, and feature engineering to detect linguistic patterns associated with cognitive biases. Qualitative assessment complements the statistical models by interpreting contextual variations and mapping communication patterns to established psychological constructs [15]. Text datasets were processed using tokenization, lemmatization, dependency parsing, and contextual embedding extraction. The combined design supports both statistical correlation and interpretive depth, enabling a comprehensive understanding of how bias signals emerge and influence decisions [16]. A multi-class supervised learning model was developed to classify biases across five primary categories: confirmation bias, anchoring, optimism bias, overconfidence, and loss aversion.

Study Area and Data Sources

The study focuses on three organizational communication environments: executive emails, leadership meeting transcripts, and internal project reports. These data sources were selected based on their direct involvement in decision-making processes and their high linguistic richness. Data were anonymized and collected from multinational organizations spanning technology, finance, and consulting sectors between 2018 and 2024 [17].

Table 1: Dataset Characteristics

Data Source	Volume (Words)	Document Count	Primary Communication Function
Executive Emails	2.1M	18,450	Strategic directives, evaluations
Meeting Transcripts	1.7M	620	Consensus-building, risk assessment
Project Reports	2.9M	1,130	Forecasting, justification, performance analysis

(Source: Organizational Archives 2018–2024)

Variables and Feature Extraction

The dependent variable is Cognitive Bias, operationalized as a multi-class categorical variable. Independent variables include linguistic indicators such as sentiment polarity, certainty markers, lexical diversity, anchoring tokens, and selective reasoning scores. Contextual variables include role seniority and decision complexity.

Table 2: Feature Indicators and Measurement Methods

Variable	Description	Measurement Source	Expected Influence
Bias-Related Terms	Lexicon of cognitive bias keywords	Custom cognitive lexicon	+
Certainty Score	Degree of assertive language	LIWC + custom parser	+
Anchoring Presence	Initial value anchoring terms	Numeric dependency parser	+
Emotional Loading	Affective intensity	NRC Lexicon	+
Semantic Coherence	Topic drift and alignment	BERT embeddings	-

Analytical Framework

Bias detection employed a transformer-based classification pipeline combining BERT embeddings with a multi-layer perceptron classifier. Feature importance was computed using SHAP values, and model robustness was evaluated via cross-validation. Multicollinearity was assessed using VIF, and statistical significance was tested at the 95% confidence level [18].

Qualitative and Thematic Analysis

A comprehensive thematic coding strategy was applied to analyze narrative patterns embedded within organizational communication. The qualitative phase served as a critical complement to the computational model by offering interpretive depth that pure statistical outputs cannot capture. The coding framework was developed using an iterative process that combined inductive category formation with deductive mapping based on established psychological constructs. Initially, all texts were segmented into analytical units such as argument structures, justification patterns, evaluative statements, and risk narratives. These units were then coded for linguistic markers associated with core cognitive biases including confirmation bias, anchoring, optimism bias, overconfidence, and loss aversion. Special attention was given to rhetorical devices such as selective evidence presentation, early fixation on initial values, dismissive treatment of alternatives, exaggerated certainty, emotionally charged reasoning, and future-positive exaggerations. Each coded segment was subsequently mapped to validated bias constructs drawn from behavioral decision theory, enabling interpretive confirmation of patterns detected by the NLP classifiers.

The analysis also examined contextual variables such as decision urgency, power hierarchy, and interaction dynamics, which often amplify or suppress bias expression. For example, communication from senior leadership displayed higher narrative rigidity, stronger anchoring on initial assumptions, and limited openness to counterevidence. Conversely, cross-functional team discussions showed greater variability in language patterns, with bias expressions emerging mainly during consensus formation or conflict negotiation. Thematic clustering further revealed recurring discourse motifs such as defensive justification, risk minimization rhetoric, and forward-looking optimism, all of which aligned closely with computationally identified bias categories. This triangulation between machine-detected signals and qualitative interpretation enhanced both reliability and construct validity. Ultimately, the thematic analysis illuminated how cognitive distortions manifest linguistically across different organizational settings, providing a nuanced understanding of the interplay between communication behavior, cognitive heuristics, and strategic judgment [19].

Ethical Considerations

Only anonymized communication datasets were used, with all personally identifiable information removed to ensure full compliance with ethical data-handling standards. Organizational consent protocols were strictly followed, including the formal approval of data use agreements and adherence to internal privacy guidelines. All documents, transcripts, and email sets were stripped of names, job titles, email IDs, and location markers prior to analysis. Data processing was conducted within secure, access-controlled environments to prevent unauthorized exposure or misuse of sensitive organizational information. Since the study relied exclusively on previously generated corporate data and involved no direct interaction with employees, no human subjects were engaged, and therefore institutional ethical clearance was not required under standard research guidelines. The study also ensured that computational models did not produce outputs that could be reverse-engineered to identify individuals or reveal confidential strategic content. These safeguards collectively ensured that the research maintained high standards of privacy, confidentiality, and ethical responsibility throughout all stages of data handling and analysis [20].

Limitations

Bias language may vary across industries, cultures, and managerial hierarchies, which limits the generalizability of the detection models across diverse organizational settings. Contextual ambiguity in natural language can also produce false positives, particularly when emotionally neutral statements resemble bias-related linguistic structures. Variations in communication norms such as directness in executive correspondence or collaborative phrasing in team discussions further complicate interpretation, requiring multi-layer analytical approaches. Additionally, the study relies on textual data alone, excluding non-verbal cues such as tone, hesitation, or conversational dynamics that often influence cognitive bias expression. Domain adaptation challenges may also arise when applying the model to sectors with specialized terminology, affecting overall accuracy. Finally, organizational communication evolves over time, meaning that models require periodic retraining to remain contextually relevant and sensitive to emerging linguistic patterns [21].

RESULTS AND ANALYSIS

Overview of Bias Detection Patterns

The model showed strong predictive patterns across all three datasets. Confirmation bias appeared most frequently in strategic planning documents, while anchoring dominated financial forecasting and negotiation transcripts. Optimism bias and overconfidence were common in executive communication, especially during high-stakes decision cycles. Semantic coherence analysis showed that bias-heavy texts exhibited sharper topic rigidity and reduced flexibility in considering alternative viewpoints.

Table 3: Cognitive Bias Prevalence Across Data Sources

Data Source	Confirmation Bias (%)	Anchoring (%)	Overconfidence (%)	Optimism Bias (%)	Loss Aversion (%)
Executive Emails	37.8	23.4	31.1	26.8	11.9
Meeting Transcripts	42.6	34.9	28.5	22.4	14.3
Project Reports	33.2	40.8	18.6	21.3	17.5

The results suggest that organizational communication is particularly vulnerable to biases during collaborative decision forums, where selective reasoning and early-value anchoring strongly shape group dynamics.

Model Performance and Feature Importance

The classification model achieved high accuracy, with BERT-based embeddings outperforming classical representations. Key features influencing predictions included certainty markers, anchoring tokens, polarity shifts, and reduced lexical diversity.

Table 4: Model Performance Metrics

Metric	Value
Accuracy	0.89
Precision	0.87
Recall	0.85
F1 Score	0.86

Visual analysis of feature contributions indicated that bias-heavy texts are characterized by strong assertive phrasing, narrow viewpoint framing, and anchored numeric references.

Demographic and Contextual Insights

Senior executives displayed higher levels of overconfidence and optimism bias, while middle managers showed more anchoring and confirmation bias. High-stakes decisions exhibited denser bias patterns, particularly in forecasting and negotiation settings.

Organizational Impact Assessment

Bias-heavy communication correlated with misaligned forecasts, delayed project timelines, and risk miscalculations. Teams with lower bias scores demonstrated more adaptive reasoning and higher decision accuracy. Key obstacles include contextual ambiguity, inconsistent communication structures, and reluctance to adopt computational monitoring tools. The results affirm that NLP can effectively identify cognitive distortions that undermine decision accuracy. Bias presence is highest where uncertainty, hierarchy, and rapid decision cycles converge.

CONCLUSION

This study demonstrates that NLP-driven cognitive bias detection offers a powerful mechanism for improving organizational decision-making. By analyzing linguistic markers across executive emails, meeting transcripts, and project reports, the research reveals strong correlations between biased communication and suboptimal decision outcomes. Transformer-based models effectively capture contextual, semantic, and emotional signals associated with heuristics such as confirmation bias, anchoring, and overconfidence. The framework shows that integrating computational detection into organizational workflows enhances transparency, strengthens decision governance, and mitigates cognitive risks. However, limitations related to context variability, adaptation requirements, and ethical constraints highlight the need for balanced implementation. NLP-based bias detection is not simply a diagnostic tool it is a strategic capability capable of transforming organizational judgment by promoting rational, data-driven decision practices.

VI. Future Work

Future research should expand datasets to cross-cultural and multilingual environments, incorporate real-time communication analytics, and integrate multimodal signals such as voice tone or meeting behaviour. Incorporating psychological profiling, ethical safeguards, and domain-specific fine-tuning will further strengthen model reliability. Researchers should also explore integration into decision intelligence systems that combine predictive modelling, behavioural monitoring, and cognitive simulation for comprehensive organizational governance.

REFERENCES

[1] Kahneman, D., Thinking, Fast and Slow. New York: Farrar, Straus and Giroux, 2011.
[2] Tversky, A., and Kahneman, D., “Judgment under uncertainty: Heuristics and biases,” Science, vol. 185, no. 4157, pp. 1124–1131, 1974.
[3] Arnott, D., “Cognitive biases and decision support systems development: A design science approach,” Information Systems Journal, vol. 16, no. 1, pp. 55–78, 2006.
[4] Muslu, V., Rebello, M., and Xu, Z., “Detecting managerial biases through textual analysis of corporate disclosures,” Journal of Accounting Research, vol. 57, no. 4, pp. 1009–1050, 2019.
[5] Li, F., “The information content of forward-looking statements in corporate disclosures: Evidence from textual analysis,” The Accounting Review, vol. 85, no. 4, pp. 1167–1197, 2010.
[6] Saffari, M. et al., “Contextual bias identification using transformer-based models,” in Proceedings of the ACL, 2020.
[7] Huang, A. H., Teoh, S. H., and Zhang, Y., “Tone management in earnings announcements: A machine learning approach,” The Review of Financial Studies, vol. 31, no. 9, pp. 3669–3709, 2018.
[8] Loughran, T., and McDonald, B., “Textual analysis in accounting and finance: A survey,” Journal of Accounting Research, vol. 58, no. 2, pp. 299–356, 2020.
[9] Bussman, K., and Papenbrock, J., “Explainable AI for NLP-based decision-making systems,” Journal of Financial Transformation, vol. 53, pp. 102–110, 2021.
[10] Reimers, N., and Gurevych, I., “Sentence-BERT: Sentence embeddings using Siamese BERT networks,” in EMNLP Conference Proceedings, 2019.

[11] Matz, S., and Netzer, O., "Using big data to study psychological traits in organizations," Current Opinion in Behavioral Sciences, vol. 18, pp. 7–12, 2017.
[12] Jaidka, K., et al., "Text-based inference of leadership communication patterns," Journal of Personality and Social Psychology, vol. 119, no. 2, pp. 409–431, 2020.
[13] Milkman, K. L., Chugh, D., and Bazerman, M., "How can decision-making be improved?," Perspectives on Psychological Science, vol. 4, no. 4, pp. 379–383, 2009.
[14] Srivastava, S., and Sahami, M., "Modeling linguistic markers of cognitive bias in group decision-making," in Proceedings of the AAAI Conference, 2021.
[15] Pennebaker, J. W., Booth, R. J., and Francis, M. E., Linguistic Inquiry and Word Count (LIWC): Technical Manual, 2015.
[16] Thagard, P., Mind-Society: How Thinking Emerges from Social Interactions. Oxford University Press, 2020.
[17] Zhang, Y., and Yang, Q., "An overview of multi-task learning in deep neural networks," National Science Review, vol. 5, no. 1, pp. 30–43, 2018.
[18] Bansal, G., et al., "Beyond accuracy: Evaluating NLP models for organizational decision support," Transactions of the ACL, vol. 9, pp. 356–372, 2021.
[19] Braun, V., and Clarke, V., "Using thematic analysis in psychology," Qualitative Research in Psychology, vol. 3, no. 2, pp. 77–101, 2006.
[20] EU GDPR, "General Data Protection Regulation," Official Journal of the European Union, 2018.
[21] Eisenhardt, K. M., and Zbaracki, M. J., "Strategic decision making," Strategic Management Journal, vol. 13, pp. 17–37, 1992.
[22] Peterson, R. S., and Hicks, M. D., Leader Mistakes and Cognitive Biases in Organizations. Wiley, 2016.
[23] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K., "BERT: Pre-training of deep bidirectional transformers," in Proceedings of NAACL,
2019.
[24] Tetlock, P. E., and Gardner, D., Superforecasting: The Art and Science of Prediction. New York: Crown Publishing, 2015.
[25] Gigerenzer, G., "Heuristics in decision making," Annual Review of Psychology, vol. 62, pp. 451–482, 2011.