

**GALATASARAY UNIVERSITY
GRADUATE SCHOOL OF SOCIAL SCIENCES
DEPARTMENT OF RADIO, TELEVISION AND CINEMA**

**THE ALGORITHMIC VEIL:
AFFECTIVE TRUST, AI BLIND SPOTS,
AND THE DESIGN OF HUMAN-CENTERED NEWS RANKING**

DOCTORAL DISSERTATION

Emre KIZILKAYA

Supervisor: Prof. Dr. Kerem RIZVANOĞLU

SEPTEMBER 2025

**GALATASARAY UNIVERSITY
GRADUATE SCHOOL OF SOCIAL SCIENCES
DEPARTMENT OF RADIO, TELEVISION AND CINEMA**

**THE ALGORITHMIC VEIL:
AFFECTIVE TRUST, AI BLIND SPOTS,
AND THE DESIGN OF HUMAN-CENTERED NEWS RANKING**

DOCTORAL DISSERTATION

Emre KIZILKAYA

Supervisor: Prof. Dr. Kerem RIZVANOĞLU

SEPTEMBER 2025

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my Ph.D. advisor, Prof. Dr. Kerem Rızvanoğlu. His exceptional guidance, unwavering support, and profound wisdom were instrumental to the completion of this work and my entire academic development. I am indebted to him for imparting invaluable knowledge and for fostering an environment of intellectual curiosity and rigorous inquiry.

My sincere thanks also extend to Prof. Dr. Serhat Güney and Prof. Dr. Mehmet Özçağlayan. As members of my Ph.D. monitoring committee, their insightful critiques, constructive feedback, and invaluable contributions significantly enriched this research. I am profoundly grateful not only for their academic expertise but also for kindness.

I dedicate this thesis to Turkey, our lonely yet beautiful country with all its wonders and disasters.

September 2025, Istanbul

Emre Kızılkaya

LIST OF CONTENTS

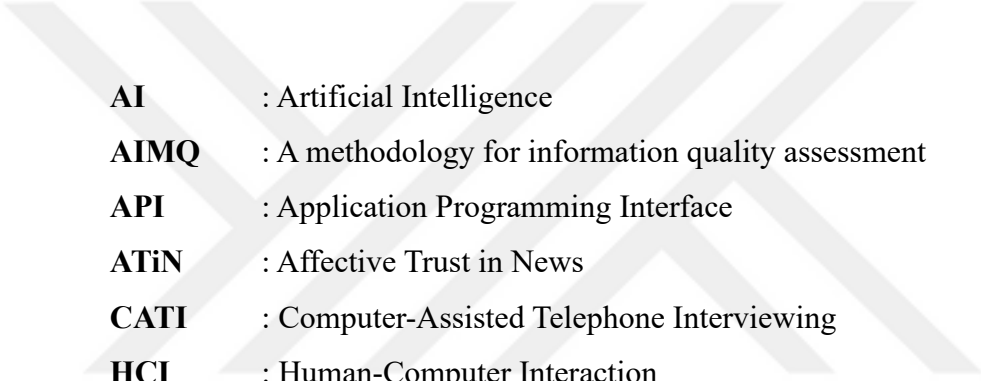
	Page No
ACKNOWLEDGEMENTS.....	ii
LIST OF CONTENTS	iii
ABBREVIATIONS	vii
LIST OF FIGURES	viii
LIST OF TABLES.....	ix
RÉSUMÉ.....	x
SUMMARY.....	xiv
ÖZET.....	xviii
1. INTRODUCTION.....	1
1.1. Purpose and the Scope of the Study	2
1.2. An Interdisciplinary Theoretical Framework	3
1.3. Methodology and Measurement Tools	5
2. AFFECTIVE TRUST IN TURKISH NEWS: A MIXED METHOD GROUNDED THEORY STUDY ON MEDIA BRANDS AMID A NATIONAL CRISIS.....	8
2.1 Introduction	9
2.1.1. Research Questions and Objectives	9
2.1.2. Theoretical Background	10
2.2. Methodology	13
2.2.1. Phase 1: National Survey	13
2.2.2. Phase 2: Semi-structured Interviews	13
2.2.3. Research Timing as a Methodological Asset	14

2.3. Results	16
2.3.1. Survey	16
2.3.2. Interviews	25
2.4. Discussion	26
2.4.1. Unsurprisability	28
2.4.2. Affective Self-fact-checking	29
2.4.3. The A(e)ffect of the Other	31
2.4.4. Why the Cleavage is So Steep (Demand vs. Supply)	33
2.4.5. Contributions and Testable Predictions	34
2.4.6. Boundary conditions and limitations	34
2.5. Conclusion	35
3. THE PROPAGANDA BLIND SPOT: AI, MEDIA POLARIZATION, AND THE DISCONNECT BETWEEN TRUST AND NEWS QUALITY	37
3.1. Introduction	37
3.1.1. The Crisis of Trust in Polarized Media Environments	39
3.1.2. Large Language Models as Tools for Social Science Research	45
3.1.3. Defining and Measuring Journalistic Quality	46
3.1.4. Research Gap, Questions, and Contribution	50
3.2. Methodology	51
3.2.1. Survey and Persona Simulations	52
3.2.2. Blind vs. Non-Blind Evaluation of Content Quality	55
3.3. Results	57
3.3.1. Part 1: Neutral AI Simulation vs. Human Baseline	58
3.3.2. Part 2: Persona-Based AI Simulation	61
3.3.3. Part 3: Blind Headline Quality Evaluation (Experts vs. LLMs)	64
3.3.4. Part 4: The Relationship Between Trust and Quality	67

3.3.5. Limitations of the Study.....	71
3.4. Discussion	72
3.4.1. Human Polarization and AI Contradiction.....	72
3.4.2. Human Affective Logic vs. Flawed AI Coherence	73
3.4.3. Theorizing the "Propaganda Blind Spot": A Failure of Pragmatics	75
3.4.4. The Three Logics of Evaluation: Identity, Norms, and Patterns	76
3.4.5. Implications for the "Stochastic Parrot" Debate	77
3.4.6. The Normative Turn: Proposing a Direction.....	79
3.5. Conclusion	79
4. TRUST BEYOND THE BRAND: A RAWLSIAN DESIGN FOR AI-ASSISTED NEWS RANKING.....	82
4.1. Introduction	82
4.1.1. Normative Frame: Rawlsian Justice for News Quality Evaluation	83
4.1.2. Design Stance: Value-Driven and Value-Sensitive Design (VDD/VSD).....	83
4.1.3. Related Work / Background	84
4.1.4. AI Alignment and Governance Context	84
4.1.5. Empirical Strategy: Phase I (Benchmarking LLM-based Trust Signals).....	85
4.1.6. Empirical Strategy: Phase II (Qualitative Analysis with Professionals).....	87
4.1.7. External Validity and Triangulation	87
4.1.8. Key Findings Preview	88
4.2. Methodology	90
4.2.1. Phase I: Experimental Assessment of News Trust Across Conditions.....	90
4.2.2. Phase II: Qualitative Inquiry with Journalists and Media Executives	96
4.3. Results	98
4.3.1. Phase I	98
4.3.2. Phase II: Journalistic and Managerial Perspectives	105

4.4. Discussion	108
4.4.1. Implications for Validity	110
4.4.2. What This Study Demonstrates	111
4.4.3. Justice as Fairness in Algorithmic Evaluation of News Quality	111
4.4.4. From Principle to Practice: A Value-driven / Value-sensitive Blueprint	112
4.4.5. Reconciling News Brand's Double Role	115
4.4.6. Boundary Conditions and Limitations	115
4.4.7. Research Agenda	116
4.5. Conclusion	117
5. CONCLUSION	118
5.1. Restatement of Research Problem and Objective	118
5.2. Summary of Findings	121
5.3. Contribution to the Field	123
5.3.1. Theoretical Contributions and Implications	124
5.3.2. Methodological Contributions and Implications	124
5.3.3. Practical Implications	125
5.4. Limitations and Further Studies	126
BIBLIOGRAPHY	128
APPENDICES	151

ABBREVIATIONS



AI	: Artificial Intelligence
AIMQ	: A methodology for information quality assessment
API	: Application Programming Interface
ATiN	: Affective Trust in News
CATI	: Computer-Assisted Telephone Interviewing
HCI	: Human-Computer Interaction
ICC	: Intraclass Correlation Coefficient
LLM	: Large Language Model
NLP	: Natural Language Processing
NQS	: Normative Quality Score
OLS	: Ordinary Least Squares
UI	: User Interface
VDD	: Value-Driven Design
VSD	: Value-Sensitive Design

LIST OF FIGURES

	Page No
Figure 1.1: Program Logic of the Thesis.....	3
Figure 2.1: The Study Flow.....	15
Figure 2.2: Expert Evaluation for the Most Frequently Mentioned Outlets.....	18
Figure 2.3: Media Outlet Trust Scores (Government).....	20
Figure 2.4: Media Outlet Trust Scores (Opposition).....	20
Figure 2.5: A Schematic Representation of the Affective Trust in News.....	28
Figure 4.1: Example Screenshots from News Websites.....	93

LIST OF TABLES

	Page No
Table 2.1: Trust Ratings of the Most Frequently Mentioned News Outlets.....	17
Table 2.2: Breakdown of Responses by Sex.....	19
Table 2.3: Relationship Between Trust in News Outlets with Interest in News.....	23
Table 2.4: Self-Reported Trust in News-Distributing Mediums.....	24
Table 2.5: Group Measurements for “Trust in News”.....	25
Table 2.6: Breakdown of Aggregated Responses for the Preferred Mediums.....	25
Table 2.7: Breakdown of Aggregated Responses for Government Trust.....	26
Table 3.1: Key Concepts in Political Polarization and Media.....	43
Table 3.2: Summary of Empirical Methods for Assessing Journalistic Quality.....	49
Table 3.3: Comparison of Average Trust Scores by Medium (1–5 Scale).....	58
Table 3.4: Mean Trust Scores for Top Turkish Media Outlets (Humans vs. LLMs).....	59
Table 3.5: Mean Quality Scores for All Headlines (Human Experts vs. LLMs).....	64
Table 3.6: Mean Objectivity Scores by Publisher’s Political Leaning (Blind Test).....	67
Table 3.7: Correlations Between Non-Blind Trust and Blind Quality.....	68
Table 3.8: Outlet-Level Case Study of Trust and Objectivity Scores.....	69
Table 4.1: Inter-run Reliability Across Conditions.....	95
Table 4.2: Overview of the Two-phase Research Design.....	98
Table 4.3: Descriptive Statistics for the Brand Effect Experiment.....	99
Table 4.4: Spearman Rank Correlations.....	99
Table 4.5: Outlet-level OLS Predicting From External Indicators.....	102
Table 4.6: Mean Trust Scores for Top Turkish Media outlets (Humans vs. LLMs).....	103
Table 4.7: Outlets with the Largest Positive and Negative Brand Trust Effects.....	104
Table 5.1: A Tabular Summary of the Studies	120

RÉSUMÉ

Cette thèse de doctorat aborde la crise croissante de confiance dans les médias, un phénomène qui sape les fondements d'une vie démocratique éclairée. Cette crise est alimentée par deux facteurs perturbateurs : la profonde polarisation politique, qui transforme la consommation médiatique en un acte d'affirmation identitaire, et la domination des plateformes algorithmiques, devenues les gardiens de facto de la sphère publique. Dans cet écosystème instable, le concept même de « confiance » est devenu un enchevêtrement conceptuel, confondant souvent la loyauté affective envers une marque, des heuristiques réputationnelles et la qualité intrinsèque du contenu journalistique. Cette confusion empêche non seulement de poser un diagnostic clair du problème, mais entrave également l'élaboration de solutions efficaces et fondées sur l'éthique. Par conséquent, la conception des systèmes de classement et de recommandation des actualités passe d'un défi technique à un problème d'ordre constitutionnel, exigeant un cadre capable de favoriser la justice informationnelle.

Cette thèse entreprend une enquête cumulative en trois étapes afin de déconstruire ce problème complexe et de proposer une voie normative et pratique pour aller de l'avant. En s'appuyant sur le cas critique de la Turquie, nation caractérisée par un marché médiatique profondément polarisé et influencé par l'État, ainsi que par une forte dépendance aux plateformes, la recherche progresse séquentiellement du diagnostic à l'audit, puis à la conception. Elle examine systématiquement comment les citoyens établissent leur confiance, comment des systèmes avancés d'intelligence artificielle (IA) évaluent la qualité des contenus journalistiques et comment les plateformes d'information peuvent être conçues pour servir les valeurs démocratiques. L'objectif global est de développer et de valider un modèle de classement fondé sur des principes et apte à la mise en œuvre, qui sépare la qualité du contenu des indices puissants, et souvent trompeurs, liés à la marque et à l'alignement affectif, tout en respectant l'autonomie des utilisateurs et en garantissant la responsabilité démocratique.

La première étude pose les bases théoriques et empiriques en déconstruisant les fondements psychologiques de la confiance dans l'information dans un environnement polarisé. Grâce à une enquête à méthodes mixtes menée pendant la période critique des tremblements de terre et des élections générales de 2023 en Turquie, cette étude développe le cadre de la Confiance affective dans l'information (ATiN). La proposition centrale de l'ATiN est que, dans des contextes hautement partisans, la confiance fonctionne moins comme une évaluation cognitive du mérite journalistique que comme un alignement affectif avec la vision politique et émotionnelle du monde d'une marque médiatique.

Cette étude identifie trois mécanismes socio-psychologiques qui créent et maintiennent des liens émotionnels forts et résistants aux faits avec les médias. Le premier est l'insusceptibilité à la surprise (unsurprisability), où l'anticipation du parti pris médiatique met à l'abri des sentiments préexistants. Le deuxième, l'auto-vérification affective des faits, est un processus d'évaluation motivé par le besoin de valider des émotions plutôt que d'établir la vérité. Le troisième, l'a(e)ffet de l'autre, construit la confiance en définissant un groupe d'appartenance par opposition à un groupe politique extérieur. L'étude montre également que même des choix lexicaux mineurs et des noms de marque dans les titres peuvent agir comme de puissants déclencheurs affectifs, court-circuitant l'évaluation rationnelle.

S'appuyant sur ce diagnostic du jugement humain, la deuxième étude étend l'enquête au domaine de l'intelligence artificielle, en auditant de manière critique les capacités et les limites des grands modèles linguistiques (LLM) en tant qu'évaluateurs du contenu journalistique. Cette étude propose un cadre comparatif opposant les jugements humains (du public et d'experts en journalisme) aux évaluations des LLM dans des conditions à la fois aveugles (contenu uniquement) et sensibles à la marque. Les résultats révèlent une double crise.

Premièrement, elle documente empiriquement un grave découplage entre confiance et qualité au sein du public, révélant une corrélation quasi nulle entre les médias auxquels le public turc fait confiance et la qualité journalistique du contenu de ces médias telle

qu'évaluée par des experts. Cette conclusion constitue une preuve quantitative frappante que la fidélité à la marque et l'identité politique ont supplanté le mérite du contenu comme principale monnaie de la confiance publique.

Deuxièmement, et plus important encore, l'étude met en évidence le « point aveugle de la propagande » des LLM. Les mêmes modèles qui étaient capables d'identifier avec précision des informations potentiellement mensongères ou propagandistes comme non fiables dans des conditions avec connaissance de la marque ont été systématiquement trompés par le contenu en évaluation dans des configurations aveugles. Les LLM ont souvent attribué des notes d'« objectivité » élevées à des titres déclaratifs syntaxiquement neutres provenant de médias partisans à faible crédibilité. Cet échec tient à une limitation fondamentale en pragmatique : tout en étant compétents en syntaxe et en sémantique, les modèles peuvent ne pas saisir la force illocutoire du langage, c'est-à-dire son intention et sa signification dépendante du contexte. L'étude conclut en théorisant trois logiques d'évaluation distinctes et concurrentes : la logique de l'identité pour le public, la logique des normes pour les experts et la logique des motifs (patterns) pour l'IA.

Motivée par les échecs identifiés, la troisième étude opère un tournant prescriptif, passant du diagnostic à la conception. Elle propose et valide un nouveau cadre de classement fondé sur la philosophie morale et des principes de conception centrés sur l'humain. S'appuyant sur la théorie rawlsienne de la justice comme équité, l'étude soutient que l'identité de marque d'un média constitue un « fait moralement non pertinent » au moment de l'évaluation de la qualité. Un système de classement équitable doit donc fonctionner derrière un « voile d'ignorance » procédural, protégeant l'évaluation de la qualité intrinsèque du contenu de l'influence biaisée des indices de marque.

Ce principe est mis en œuvre au moyen d'une approche hybride combinant la conception sensible aux valeurs (VSD) et la conception axée sur les valeurs (VDD). L'étude développe et valide d'abord le Normative Quality Score (NQS), un référentiel communautaire et auditable du mérite journalistique, dérivé de signaux transparents tels que les prix journalistiques, les sanctions professionnelles et les signalements d'organisations de vérification des faits. Elle teste ensuite un protocole « voile puis

personnalisation » à l'aide d'une méthodologie LLM à gradient contextuel (aveugle, sensible à la marque, visuel). Les résultats montrent que si les évaluations sensibles à la marque s'alignent fortement sur le NQS, les évaluations aveugles (basées uniquement sur le contenu) fournissent un signal fiable, quoique plus modéré, de la qualité intrinsèque.

Cela fournit la base empirique d'un plan pratique pour un classement équitable de l'information, articulé autour de cinq exigences clés de conception : ignorance procédurale au moment du scoring ; garde-fous de qualité à plusieurs niveaux ; transparence et possibilité de contestation ; gouvernance intégrant la communauté ; garde-fous IA pour les rédactions.

En synthèse, cette thèse documente une double crise de l'écosystème informationnel : un public humain guidé par l'identité affective et un système d'IA vulnérable à l'art subtil de la propagande. La convergence de ces deux forces (polarisation affective et logique algorithmique imparfaite des LLM) menace les fondements épistémiques partagés nécessaires à l'auto-gouvernance démocratique.

La contribution principale de ce travail consiste à offrir une solution théoriquement fondée, empiriquement validée et opérationnalisable, accompagnée d'un riche ensemble de données originales. En faisant progresser le cadre ATiN, en identifiant le « point aveugle de la propagande » et en proposant une conception rawlsienne « voile puis personnalisation », cette thèse requalifie le défi du classement de l'information : d'un problème technique d'optimisation à une question urgente de philosophie publique et de conception centrée sur l'humain. L'objectif ultime n'est pas seulement de produire des algorithmes plus efficaces, mais de concevoir des interactions plus justes, via des systèmes d'information qui servent la dignité humaine, favorisent la justice informationnelle et rétablissent les conditions d'une sphère publique fiable et résiliente.

SUMMARY

This doctoral dissertation confronts the escalating crisis of trust in news, a phenomenon undercutting the foundations of informed democratic life. The crisis is propelled by a dual-engine of disruption: the deep political polarization that transforms media consumption into an act of identity affirmation, and the dominance of algorithmic platforms that have become the *de facto* curators of the public sphere. Within this volatile ecosystem, the very concept of "trust" has become a conceptual thicket, often conflating affective brand loyalty, reputational heuristics, and the intrinsic quality of journalistic content. This conflation not only obstructs a clear diagnosis of the problem but also stymies the development of effective, ethically grounded solutions. Consequently, the design of news ranking and recommendation systems evolves from a technical challenge into a problem with constitutional stakes, demanding a framework that can foster informational justice.

This thesis embarks on a three-stage, cumulative inquiry to deconstruct this complex problem and propose a normative, practical path forward. Using the critical case of Turkey, which is a nation characterized by a politically-influenced, deeply polarized media market and high platform dependency, the research moves sequentially from diagnosis to audit and, finally, to design. It systematically investigates how citizens form trust, how advanced artificial intelligence (AI) systems evaluate news quality, and how information platforms can be architected to serve democratic values. The overarching objective is to develop and validate a principled, governance-ready blueprint for news ranking that separates content quality from the powerful, and often misleading, cues of brand and affective alignment, while simultaneously respecting user agency and ensuring democratic accountability.

The first study lays the theoretical and empirical groundwork by deconstructing the psychological underpinnings of news trust in a polarized environment. Through a mixed-

methods investigation conducted during the critical juncture of the 2023 Turkish earthquakes and general elections, this study develops the Affective Trust in News (ATiN) framework. The central proposition of ATiN is that in hyper-partisan contexts, trust functions less as a cognitive appraisal of journalistic merit and more as an affective alignment with a media brand's political and emotional worldview.

This study identifies three socio-psychological mechanisms that create and maintain strong, fact-resistant emotional bonds with media. The first is unsurprisability, where expecting media bias shields pre-existing feelings. The second, affective self-fact-checking, is an evaluation process driven by the need to validate emotions rather than ascertain truth. The third, the a(e)ffect of the other, constructs trust by defining an in-group against a political out-group. The study also shows that even minor word choices and brand names in headlines can act as potent affective triggers, short-circuiting rational evaluation.

Building on this diagnosis of human judgment, the second study extends the inquiry to the realm of artificial intelligence, critically auditing the capabilities and limitations of Large Language Models (LLMs) as evaluators of news content. This study pioneers a comparative framework that pits human judgments (from both the public and journalism experts) against LLM ratings under both blind (content-only) and brand-aware conditions. The findings reveal a dual crisis.

First, it empirically documents a severe trust-quality decoupling among the human audience, revealing a near-zero correlation between the media outlets the Turkish public trusts and the journalistic quality of those outlets' content as assessed by experts. This finding provides stark quantitative evidence that brand loyalty and political identity have superseded content merit as the primary currency of public trust.

Second, and more critically, the study uncovers the "Propaganda Blind Spot" in LLMs. The same models that could accurately identify potential misinformation and propaganda as untrustworthy in brand-aware settings were systematically deceived by their content in blind evaluations. The LLMs frequently awarded high "objectivity" scores

to syntactically neutral, declarative headlines from low-trust partisan outlets. This failure is traced to a fundamental limitation in pragmatics: while proficient in syntax and semantics, the models can fail to grasp the illocutionary force of language, which is its intended purpose and context-dependent meaning. The study concludes by theorizing three distinct, competing logics of evaluation: the logic of identity for the public, the logic of norms for experts, and the logic of patterns for AI.

Motivated by the failures identified in the preceding studies, the third study makes a prescriptive turn, moving from diagnosis to design. It proposes and validates a novel framework for news ranking grounded in moral philosophy and human-centered design principles. Drawing on John Rawls's theory of justice as fairness, the study argues that an outlet's brand identity is a "morally irrelevant fact" at the point of quality assessment. A just ranking system, therefore, must operate behind a procedural "veil of ignorance," shielding the evaluation of content-intrinsic quality from the biasing influence of brand cues.

This principle is operationalized through a hybrid Value-Sensitive Design (VSD) and Value-Driven Design (VDD) approach. The study first develops and validates the Normative Quality Score (NQS), a community-anchored, auditable benchmark of journalistic merit derived from transparent signals like journalism prizes, professional sanctions, and fact-checker flags. It then tests a "veil-then-personalize" protocol using a context-gradient LLM methodology (blind, brand-aware, visual). The results demonstrate that while brand-aware ratings align strongly with the NQS, blind (content-only) ratings provide a reliable, if more moderate, signal of intrinsic quality.

This provides the empirical foundation for a practical blueprint for fair news ranking, centered on five key design requirements: Procedural ignorance at scoring, tiered quality guardrails, transparency and contestability, community-in-the-loop governance, newsroom AI guardrails.

In synthesis, this thesis documents a dual crisis in the information ecosystem: a human public driven by affective identity and an AI system susceptible to the subtle art of

propaganda. The convergence of these two forces (affective polarization and flawed algorithmic logic of LLMs) threatens the shared epistemic foundations required for democratic self-governance.

The principal contribution of this work is to offer a theoretically grounded, empirically validated, and practically implementable solution alongside a rich set of original datasets. By advancing the ATiN framework, identifying the Propaganda Blind Spot, and proposing a Rawlsian "veil-then-personalize" design, this dissertation provides a transformative path forward. It reframes the challenge of news ranking from a technical problem of optimization to an urgent matter of public philosophy and human-centered design. The ultimate goal is not merely to create more efficient algorithms, but to design more just interactions: informational systems that serve human dignity, foster informational justice, and restore the conditions for a trustworthy and resilient public sphere.

ÖZET

Bu doktora tezi, haberciliğe yönelik bir güven krizinin tırmandığı bir dönemde, bilgiye dayalı demokratik yaşamın temellerini aşındıran bu olguyu incelemektedir. Kriz, iki yanlı bir bozulma dinamiğiyle beslenmektedir: medya tüketimini kimlik onayına dönüştüren derin siyasal kutuplaşma ve kamusal alanın fiilî küratörleri hâline gelen algoritmik platformların hâkimiyeti. Bu oynak ekosistemde “güven” kavramı çoğu zaman duygulanımsal marka sadakati, itibara dayalı sezgisel kestirmeler ve gazetecilik içeriğinin içkin niteliğiyle birbirine karışan kavramsal bir karmaşaya dönüşmektedir. Bu karışım, sorunun berrak bir teşhisini engellediği gibi, etkili ve etik temelli çözümlerin geliştirilmesini de sekteye uğratır. Bunun sonucunda, haberlerin dijital ortamdaki dağıtımı için yapılan sıralamaların ve kurulan öneri sistemlerinin tasarımı, teknik bir zorluktan çıkıp demokratik düzeyde sonuçlar doğuran bir meseleye dönüşerek enformasyon adaletini teşvik edebilecek bir çerçeveyi gerekli kılar.

Tez, bu karmaşık sorunu çözümlmek ve normatif, pratik bir yol önermek için üç aşamalı, birikimli bir incelemeye girişmektedir. Siyasî etkilerin hissedildiği, derin biçimde kutuplaşmış bir medya piyasası ve yüksek platform bağımlılığıyla karakterize edilen Türkiye’nin örnek vaka olarak kullanıldığı araştırma, sırasıyla teşhisten denetime ve nihayet tasarıma ilerlemektedir. Çalışma, yurttaşların güveni nasıl oluşturduğunu, gelişmiş yapay zekâ sistemlerinin haber niteliğini nasıl değerlendirdiğini ve bilgi platformlarının demokratik değerlere hizmet edecek biçimde nasıl kurgulanabileceğini sistematik olarak inceler. Genel amaç, içerik kalitesini marka ve duygulanımsal hizalanma gibi güçlü ve çoğu zaman yanıltıcı ipuçlarından ayıran, kullanıcı tercihlerine saygı gösterirken demokratik hesap verebilirliği güvence altına alan, ilkesel ve yönetişime hazır bir haber öneri tasarım şemasını geliştirmek ve geçerlemektir.

Birinci çalışma, kutuplaşmış bir ortamda haber güveninin psikolojik zeminini çözümleyerek kuramsal ve ampirik altyapıyı kurar. 2023 Kahramanmaraş depremleri ve

genel seçimlerinin kritik kavşağında yürütülen karma yöntemli bir soruşturma yoluyla “Haberde Duygulanımsal Güven” (Affective Trust in News, ATiN) çerçevesini geliştirir. ATiN’in merkezî önermesi, hiper-partizan bağlamlarda güvenin gazetecilik liyakatinin bilişsel bir takdiri olmaktan ziyade, bir medya markasının siyasal ve duygusal dünya görüşüne duygulanımsal bir hizalanma olarak işlediğidir.

Bu çalışma, medya ile güçlü ve olguya dirençli duygusal bağlar kuran ve sürdüren üç sosyo-psikolojik mekanizmayı saptar. Birincisi, medyadaki önyargıyı beklemenin var olan duyguları zırladığı ‘şaşırmazlık’tır (unsurprisability). İkincisi, doğruluğu saptamaktan ziyade duyguları onaylama ihtiyacıyla güdülenen bir değerlendirme süreci olan ‘duygulanımsal öz-doğrulama’dır (affective self-fact-checking). Üçüncüsü ise bir iç-grubu siyasal bir dış-gruba karşı tanımlayarak güveni inşa eden ‘ötekinin (e)tkisi’dir (a(e)ffect of the other). Çalışma ayrıca, başlıklardaki basit sözcük seçimleri ve marka adlarının güçlü duygulanımsal tetikleyiciler olarak işlev görebildiğini ve rasyonel değerlendirmeyi kısa devreye uğratabildiğini göstermektedir.

İkinci çalışma, insan yargısına ilişkin bu teşhise dayanarak sorgulamayı yapay zekâ alanına doğru genişletir ve Büyük Dil Modelleri’nin (LLM’lerin) haber içeriğini değerlendiriciler olarak yetkinlik ve sınırlılıklarını eleştirel biçimde denetler. Çalışma, kamuoyu ve gazetecilik uzmanlarından gelen insan yargılarını, hem kör (yalnızca içerik) hem de marka bilgisi verilen koşullarda LLM puanlarıyla karşı karşıya getiren özgün bir karşılaştırma çerçevesi sunar. Bulgular ikili bir krizi açığa çıkarır.

İlk olarak, Türkiye kamuoyunun güvendiği medya kuruluşları ile uzmanların bu kuruluşların içeriklerinin gazetecilik kalitesi hakkındaki değerlendirmeleri arasında neredeyse hiçbir bağıntı bulunmadığını ampirik olarak belgelenmektedir. Bu bulgu, kamu güveninin birincil para birimi olarak içerik liyakatinin yerini marka sadakati ve siyasal kimliğin aldığını gösteren nicel kanıtı sunar.

İkinci ve daha kritiği, çalışmanın LLM’lerde ‘Propaganda Kör Noktası’nı açığa çıkarmasıdır. Marka bilgisi verilen koşullarda potansiyel dezenformasyon ve propagandayı güvenilir olarak doğru biçimde ayırt edebilen aynı modeller, kör

değerlendirmelerde içerik tarafından sistematik olarak yanıltılmıştır. LLM’ler, düşük güvenilirliğe sahip partizan mecralardan gelen sözdizimsel olarak nötr, bildirici başlıklara sıklıkla yüksek “nesnellik” puanları vermiştir. Bu başarısızlık, dilin amaçlanan kullanımını ve bağlama bağımlı anlamını ifade eden illoküsyoner gücü kavramaya, yani edimbilimdeki temel bir sınıra kadar izlenmektedir. Çalışma, üç farklı ve rekabet hâlindeki değerlendirme mantığını kuramsallaştırarak sonlanır: kamuoyu için kimlik mantığı, uzmanlar için normlar mantığı ve yapay zekâ için örüntüler mantığı.

Önceki çalışmalarda saptanan arızalardan yola çıkan üçüncü çalışma, teşhisten tasarıma normatif bir dönüş yapar. Ahlâk felsefesi ve insan merkezli tasarım ilkelerine dayalı yeni bir haber sıralama ve öneri çerçevesi kurar. John Rawls’un “hakkaniyet olarak adalet” kuramından hareketle, bir mecranın marka kimliğinin kalite değerlendirmesi anında “ahlâken ilgisiz bir olgu” olduğu savunulur. Âdil bir sıralama sistemi bu nedenle, içerik-içkin kalitenin değerlendirilmesini marka ipuçlarının önyargılayıcı etkisinden koruyan usûle ilişkin bir ‘bilgisizlik perdesi’nin ardında çalışmalıdır.

Bu ilke, Değer-Duyarlı Tasarım (Value-Sensitive Design, VSD) ile Değer-Öncelikli Tasarım’ın (Value-Driven Design, VDD) hibrit bir yaklaşımıyla işletimselleştirilir. Çalışma önce gazetecilik ödülleri, meslekî yaptırımlar ve doğrulama kuruluşu işaretleri gibi şeffaf göstergelerden türetilen, topluluk temelli ve denetlenebilir bir ölçüt olan Normatif Kalite Skoru’nu (NQS) geliştirir. Ardından bağlam-kademeli bir LLM metodolojisiyle (kör, marka-bilinçli, görsel) “önce ört, sonra kişiselleştir” protokolünü sınar. Sonuçlar, marka bilgisi verilen puanlamaların NQS ile güçlü biçimde hizalandığını, kör (yalnızca içerik) puanlamaların ise içkin kaliteye ilişkin daha ılımlı olmakla birlikte güvenilir bir sinyal sağladığını göstermektedir.

Bu bulgular, âdil bir haber öneri algoritması için pratik bir tasarım şemasının ampirik temelini oluşturur ve beş kritik tasarım gereksinimi etrafında toplanır: puanlamada prosedürel bilgisizlik, kademeli kalite kontrol noktaları, şeffaflık ve itiraz edilebilirlik, topluluk-döngülü yönetim ve haber merkezleri için yapay zekâ korkulukları.

Bütün olarak bakıldığında, bu tez bilgi ekosisteminde iki yönlü bir krizi belgeliyor. Duygulanımsal kutuplaşma ile LLM’lerin hatalı algoritmik mantığının kesişimi, demokrasi için gerekli epistemik temelleri tehdit etmektedir. Bu çalışmanın temel katkısı, kuramsal temeli sağlam, ampirik olarak geçerlenmiş ve pratikte uygulanabilir bir çözümü, zengin bir özgün veri kümesiyle birlikte sunmasıdır. ATiN çerçevesini ilerleterek, Propaganda Kör Noktası’nı tanımlayarak ve Rawlsçu “önce ört, sonra kişiselleştir” tasarımını önererek, haber sıralaması sorununu bir optimizasyon meselesi olmaktan çıkarıp kamusal felsefe ve insan merkezli tasarımın acil bir konusu olarak yeniden çerçeveler. Nihai amaç sadece daha verimli algoritmalar üretmek değil, daha adil etkileşimler tasarlamaktır. İnsan onuruna hizmet eden, enformasyon adaletini teşvik eden ve güvenilir, dayanıklı bir kamusal alanın koşullarını yeniden tesis eden bilgi sistemlerine katkı hedeflenmektedir.

1. INTRODUCTION

Trust in news enables citizens to make sense of public affairs and exercise political judgment. Yet the systems that now organize visibility and attention are largely platformic and algorithmic, which means that ranking and recommendation effectively allocate civic opportunity. Prior research shows that both platform design and user networks shape exposure to information, often in ways that accentuate prior preferences or conceal system behavior from users. These dynamics recast questions of credibility and exposure as design problems with constitutional stakes, not merely technical ones (Bakshy, Messing, & Adamic, 2015; Eslami, Rader, & Karahalios, 2015; Napoli, 2015).

A conceptual difficulty complicates diagnosis and remedy. Under the single label “trust,” industry practice and portions of the research literature frequently compress distinct constructs: affective attachment to a brand, reputational familiarity or source credibility, and the intrinsic quality of a specific item of content. Classic work on source credibility, and subsequent models of web credibility, demonstrate that cues tied to the communicator can dominate people’s evaluations of identical messages, even when message quality is held constant. Communication research has also sought to measure outlet-level trust directly through validated multi-dimensional scales (Hovland & Weiss, 1951; Kohring & Matthes, 2007; Wathen & Burkell, 2002). Conflation among these constructs risks mistaking news brand heuristics for content merit and obscures how platform labeling and interface cues can shortcut deliberation.

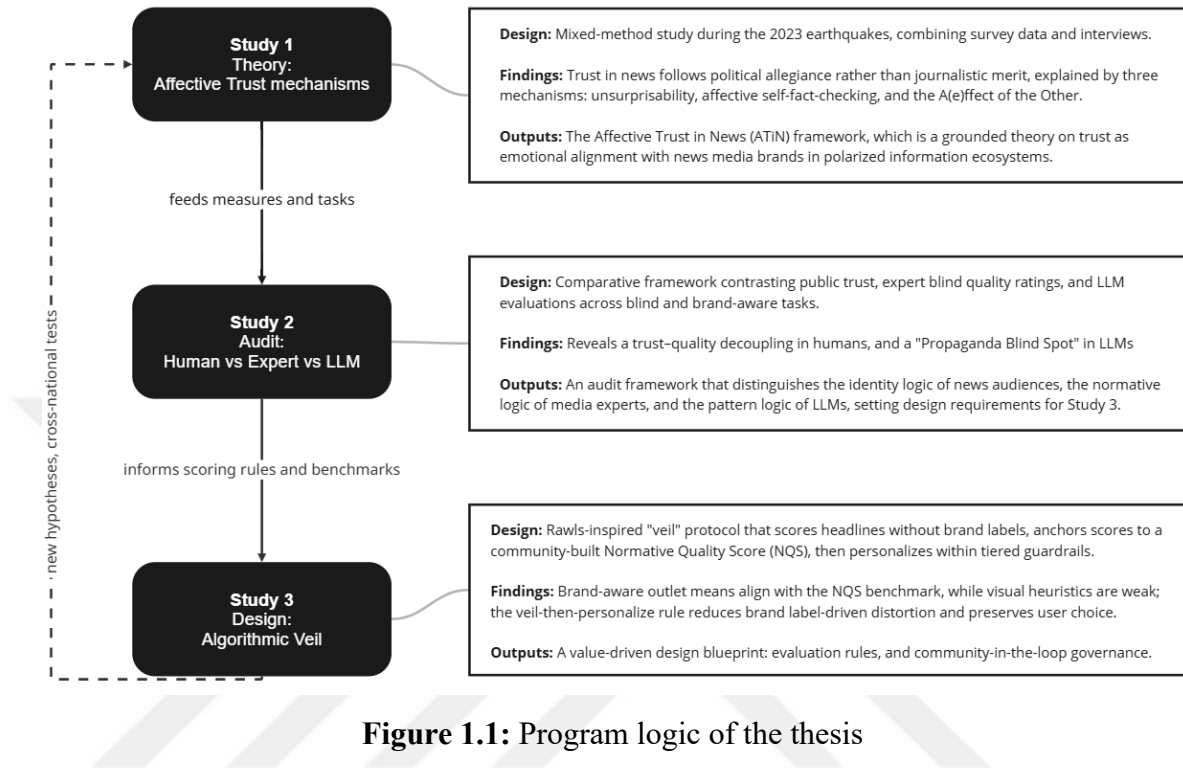
1.1. Purpose and the Scope of the Study

The purpose of this thesis is to critically examine the multifaceted nature of "trust in news" in the digital age, particularly within polarized and state-influenced media environments. It seeks to disentangle various conceptualizations of trust—ranging from affective brand attachment to the intrinsic quality of content—and to explore how these different facets interact within human perception and algorithmic evaluation. Furthermore, the thesis aims to move beyond diagnosis by proposing and evaluating a novel, normatively anchored design solution for news ranking that seeks to foster informational justice.

The scope of this study is comprehensive, encompassing theoretical development, empirical analysis using both human and artificial intelligence evaluations, and the design of practical, governance-ready protocols for information platforms. This thesis argues that understanding and addressing the crisis of trust in news requires an interdisciplinary approach that considers socio-psychological mechanisms of human judgment, the capabilities and limitations of advanced AI systems, and the ethical imperatives for platform design.

Crucially, this thesis is structured around three interconnected studies, each presented as a dedicated chapter, that cumulatively build a powerful argument from diagnosis to remedy. Study 1 delves into the affective dimensions of trust, identifying the core mechanisms at play in polarized contexts. Study 2 then critically audits how both humans and large language models (LLMs) evaluate news quality, exposing significant disconnects. Finally, Study 3 makes a prescriptive turn, designing and validating a principled approach to news ranking that mitigates the problems identified in the preceding studies.

As Figure 1.1. shows, Study 1 theorizes affective trust (ATiN), Study 2 audits human, expert, and model evaluations, and Study 3 designs and validates a veil-then-personalize ranking protocol for news, anchored to an auditable Normative Quality Score (NQS).



1.2. An Interdisciplinary Theoretical Framework

A significant conceptual difficulty complicates the diagnosis and remedy of declining news trust: the frequent conflation, under the single label “trust,” of distinct constructs within industry practice and portions of the research literature. This thesis argues that “trust” often compresses affective attachment to a brand, reputational familiarity or source credibility, and the intrinsic quality of a specific item of content. Classic work on source credibility (Hovland & Weiss, 1951) and subsequent models of web credibility (Wathen & Burkell, 2002) have long demonstrated that cues tied to the communicator, such as a media brand, can dominate people’s evaluations of identical messages, even when message quality is held constant. Communication research has also sought to measure outlet-level trust directly through validated multi-dimensional scales (Kohring & Matthes, 2007). This conflation risks mistaking superficial news brand heuristics for genuine content merit and obscures how platform labeling and interface cues can shortcut deliberation, leading to potentially biased and ill-informed judgments.

To unpack these complex issues, the thesis draws upon a rich interdisciplinary theoretical framework, integrating insights from communication studies, political psychology, and computer science, with a strong normative foundation from moral philosophy.

Study 1 introduces the Affective Trust in News (ATiN) framework. ATiN posits that in highly polarized and state-influenced media systems, trust functions less as a cognitive appraisal of journalistic quality and more as an affective alignment with a media brand's political and emotional worldview. This framework is grounded in three adjacent literatures:

- Selective exposure and partisan media choice (Iyengar & Hahn, 2009; Stroud, 2010): These literatures demonstrate that audiences tend to prefer sources affiliated with their political identity, making brand labels powerful affective triggers long before content is read.
- Motivated reasoning and identity-protective cognition (Kahan, 2012; Taber & Lodge, 2006): These theories explain why individuals evaluate identical evidence differently when it affirms or threatens their group commitments. ATiN reframes this as "affective self-fact-checking," where the goal is to preserve a prior feeling state rather than neutrally seek truth.
- Affective polarization research: This body of work clarifies how out-group animus has grown and is transmitted through non-policy cues, which helps explain the "A(e)ffect of the Other" mechanism in ATiN, where trust is relationally constructed against a political out-group.

These literatures jointly imply that in polarized systems, brand and identity cues become first-order determinants of perceived trustworthiness. Consequently, interventions focused solely on content quality signals are systematically dampened unless the design actively attenuates the triggering power of labels at critical decision points.

Study 2 extends this theoretical lens to the realm of artificial intelligence, engaging with the emerging literature on the limits and risks of LLMs in socio-technical contexts.

While LLMs can encode extensive non-blind knowledge about news outlets (e.g., their political leanings), they can fail to register pragmatic meaning and illocutionary force when evaluating content in a constrained, blind setting (e.g., a headline alone). Recent risk taxonomies for language model deployments emphasize misinformation harms, context loss, and human-computer interaction (HCI) risks (Weidinger et al., 2021, 2022). This motivates auditing LLMs as measurement instruments, rather than treating their outputs as ground truth, and triangulating model signals against community-anchored benchmarks and expert assessments.

Study 3 adopts a normative frame for design, drawing on Rawlsian justice as fairness and Kantian respect for persons. If scoring rules for news are to be justifiable to citizens regardless of their media loyalties, then morally irrelevant facts such as news brand identity should not influence assessments of item-level quality at the moment of evaluation and scoring. A procedural "veil" at scoring time meets this requirement, ensuring fairness. Once quality tiers are established, personalization can proceed within guardrails that protect equal informational opportunity. This approach aligns with treating users as ends, not as means to engagement metrics, and with a civic *telos* for news that supports public reason. Within Human-Computer Interaction (HCI), this operationalizes Value Sensitive Design (VSD) by making journalistic values explicit inputs to the scoring protocol and its governance (Friedman & Hendry, 2019; Kant, 1997; Rawls, 1999; Sandel, 2009).

1.3. Methodology and Measurement Tools

Methodologically, this thesis integrates diverse approaches, combining qualitative inquiry, survey-based measures, expert assessments, and LLM-based instruments with rigorous reliability diagnostics and external benchmarking. Turkey serves as a critical case for unpacking these issues, combining a deeply polarized, state-influenced media market with a platform-centric distribution regime. Comparative survey evidence consistently documents low and volatile trust in news and heavy reliance on digital and platform intermediaries for access (Reuters Institute, 2025), patterns that complicate any simple mapping between brand reputation and content quality. This environment provides a

magnified lens through which to observe the dynamics of affective trust and the challenges for both human and AI evaluation.

Study 1 grounds theory development in Turkey through mixed methods, conducted during the critical juncture of the 2023 earthquakes and subsequent election period. This qualitative and quantitative inquiry provided the empirical basis for developing the ATiN framework and identifying its key socio-psychological mechanisms.

Study 2 introduces a novel comparative framework. It directly contrasts human trust (both public and expert ratings) with LLM evaluations under both blind (content-only) and brand-aware conditions. This diagnostic approach allows for the precise identification of the "trust–quality gap" in human audiences and the "Propaganda Blind Spot" in AI systems, revealing where human bias and AI naivety converge to undermine reliable information flow.

Study 3 makes a prescriptive turn, formalizing a context-gradient measurement protocol and introducing the Normative Quality Score (NQS). The NQS is a community-anchored composite benchmark that combines positive signals (e.g., journalism prizes, newsletter citations) with negative signals (e.g., fact-checker flags, Press Council sanctions). This study then empirically tests a "veil-then-personalize" blueprint that links content scoring to a robust governance framework, including transparency, contestability, and community-in-the-loop updates. This methodological approach adheres to a VSD stance, where values guide concept, build, and evaluation, and treats LLMs as auditable tools whose signals require external validation rather than deference (Friedman & Hendry, 2019; Reuters Institute, 2025; Weidinger et al., 2021, 2022).

Overall, this thesis offers a cumulative argument in three steps:

- Study 1 provides a diagnosis of affective signaling, showing how small lexical choices in headlines operate as partisan signals that elicit affectively driven trust, establishing the headline as the appropriate unit of analysis.

- Study 2 conducts an audit, contrasting human judgments with LLM ratings for headlines under blind and brand-aware conditions, demonstrating a dual failure to track quality: humans follow identity-protective affect, while models pursue surface coherence that misses context.
- Study 3 moves to design, encoding the journalistic community's evaluative norms in an auditable benchmark and testing a "veil-then-personalize" scoring protocol that insulates item-level quality assessment from brand cues.

The ultimate aim is to demonstrate why trust should be treated as multi-component, why content scoring must be insulated from news brand labels at key decision points, and how a governance-ready benchmark can anchor fair ranking, thereby advancing informational justice in the digital public sphere.

2. AFFECTIVE TRUST IN TURKISH NEWS: A MIXED METHOD GROUNDED THEORY STUDY ON MEDIA BRANDS AMID A NATIONAL CRISIS

Scholarly inquiry into media trust is rooted in the understanding of trust as a cognitive evaluation of a source's credibility and expertise (Hovland et al., 1953). This means audiences are traditionally seen as engaging in a rational assessment of the veracity of reporting and the reliability of the news organization itself. More contemporary work has operationalized this concept into multidimensional scales, conceptualizing trust as an audience's expectation that a news outlet will demonstrate integrity in its journalistic practices. These dimensions include the fair selectivity of topics and facts, the accuracy of depictions, and the overall quality of journalistic assessment (Kohring & Matthes, 2007). At its core, then, traditional media trust is defined as a citizen's confident expectation that a news organization will fulfill its informational role with accuracy and good faith.

This study moves beyond a purely cognitive understanding of media trust to propose and develop a new conceptual framework: "Affective Trust in News" (ATiN). We argue that in highly polarized and state-influenced media landscapes, trust is less a rational assessment of journalistic quality and more an affective alignment with the political and emotional worldview a media brand represents. This affective process is deeply intertwined with a citizen's trust in their government, creating powerful, identity-based feedback loops that are resilient to contradictory factual information.

2.1 Introduction

The present study develops this theory through a critical case analysis of Turkey. The country offers a unique environment for examining these dynamics due to its deep political polarization, the government's systematic consolidation of media control (Waldman & Caliskan, 2019), and the occurrence of a major national crisis (the 2023 earthquakes) which starkly highlighted the fissures in the nation's information ecosystem. By initially identifying the predictors of audience trust through survey data and subsequently developing the ATiN grounded theory through in-depth interviews, this research aims to provide a robust conceptual tool with explanatory power beyond a single national context.

Our research began in May 2022, a full year before the general elections in which President Recep Tayyip Erdoğan was renominated for an additional term in office by his ruling Justice and Development Party (AKP). In the months preceding the elections, two devastating earthquakes struck 11 provinces in southeastern Turkey on February 6, 2023, causing over 50,000 fatalities and displacing more than 1.5 million individuals (Foster, 2023). We executed the final stage of our study in the wake of these seismic events and concluded our investigation four weeks prior to the general elections. The ruling alliance won the parliamentary majority in the general elections on May 14, 2023, and Erdoğan was re-elected in the second round of presidential elections on May 28.

2.1.1. Research Questions and Objectives

The existing literature on media trust, largely rooted in Western democratic contexts, often emphasizes cognitive assessments of credibility and journalistic practice. This framework is less equipped to explain the dynamics of trust in highly polarized, state-influenced media environments where news consumption is an act of political and emotional alignment. This study addresses this theoretical gap by using the critical case of Turkey to develop a new conceptual framework: Affective Trust in News (ATiN).

We employ a two-phase sequential mixed-methods design. Phase 1 is a nationally representative survey that diagnoses the structure of media trust and tests whether political allegiance (operationalized as trust in the government and trust in the opposition) organizes out-let-level trust more than sociodemographics or media use. Phase 2 is a theory-building qualitative inquiry that explains the socio-psychological mechanisms underlying those Phase-1 regularities through semi-structured interviews. Phase 2 comprises two strata within the same qualitative phase: Phase 2a interviews with residents in earthquake-affected provinces and Phase 2b interviews with Istanbul residents who experienced the crisis primarily through mediated coverage. Our research questions are:

RQ1 (Phase 1): To what extent does trust in the government predict trust in news brands, net of demographics and media use?

RQ2 (Phase 1): How do audiences segment in their trust across platforms, and how does this segmentation align with the political divide?

RQ3 (Phase 2): Through what affective mechanisms do audiences experience and construct trust or distrust in news during a national crisis?

RQ4 (Phase 2): How does motivated reasoning operate as affective self-fact-checking, protecting identity-congruent emotional states rather than only propositional beliefs?

RQ5 (Phase 2): How are affective communities of “we” and “they” constructed in relation to perceived Others in media and politics?

By addressing these questions, this study moves beyond a descriptive case analysis. It uses the Turkish context to deconstruct the processes of media trust and construct a new, generalizable framework (ATiN) that offers novel explanatory insights into the nature of trust in polarized societies worldwide.

2.1.2. Theoretical Background

Research on trust in news media has traditionally identified a range of individual-level determinants, including the audience's political ideology, trust in governmental institutions, economic perceptions, and demographic characteristics such as age and

education level (Fawzi et al., 2021; Jurkowitz et al., 2020; Lee, 2010; Tsfatı & Ariely, 2014). Scholars have also examined how media consumption habits relate to trust, noting, for instance, that increased news usage is often associated with higher trust, while exposure to online news, in particular, may correlate with increased skepticism (Fawzi et al., 2021; Tsfatı & Capella, 2003; Tsfatı, 2010). Correspondingly, measurement scales have been developed to assess audience trust based on perceptions of journalistic standards, such as the selectivity of topics and facts, accuracy, and overall journalistic assessment (Kohring & Matthes, 2007).

While valuable, these approaches often presuppose an audience that is willing and able to engage in a rational evaluation of news content. Their explanatory power diminishes in highly polarized media systems, where partisan identity frequently becomes the primary heuristic for information processing (Fawzi et al., 2021). In such contexts, media outlets may function less as purveyors of objective information and more as markers of identity, shifting their principal role from informing citizens to affirming their political allegiances (Ling, 2020). This shift fundamentally alters the basis on which audience trust is constructed and maintained.

This study argues that in environments of high political polarization, the trust-building process becomes contingent on affective responses. Drawing on theories of motivated reasoning and social identity, we posit that news content is first filtered through an affective lens tied to political allegiance. Messages that align with an individual's in-group identity can foster feelings of security and validation, thereby reinforcing trust in the source. Conversely, content that challenges this identity can provoke feelings of anger and fear, solidifying distrust. This potent affective response serves as a heuristic that guides (and often biases) subsequent cognitive processing, making trust remarkably resilient to contradictory factual information.

Following the literature review, we identify two critical theoretical gaps that this study aims to address. The first is that extant models of media trust, largely built on cognitive evaluation, struggle to explain the stability of trust in the face of persistent, factually inaccurate reporting, a common feature of polarized contexts. This suggests a

missing variable or a mis-specified relationship in current theories. The second gap is the need for a conceptual framework that explains how political allegiance, which is an attitude, is translated into the resilient, emotion-laden trust (or distrust) that characterizes these environments. Simply stating that political ideology is a correlate of media trust is not enough; we need to theorize the socio-psychological mechanisms of this translation. Our Affective Trust in News (ATiN) framework addresses these gaps by theorizing the specific affective processes that mediate the relationship between political identity and media trust.

We define ATiN as a process in which an individual's affective alignment with the political and emotional worldview represented by a media brand serves as the primary condition for granting or denying trust. This model does not suggest that cognitive evaluation disappears; rather, it reconceptualizes the relationship, positing that in polarized contexts, the foundational expectation of trust is no longer based on journalistic practice but on an affective sense of whether a news source is “for us” or “against us.”

Turkey presents a key case for examining this framework. Audience interest in news is also closely associated with trust in Turkey as in several other countries (Kalogeropoulos et al, 2019). Its highly polarized political and media landscape has been extensively analyzed in studies of authoritarianism and alternative media (Bulut & Ertuna, 2022; Metin & Ramaciotti Morales, 2022; Tunç, 2014), yet a thorough investigation into audience trust remains underdeveloped. Major international surveys on media trust have either omitted Turkey or employed methodologies ill-suited to its specific context. For example, the Reuters Institute Digital News Report, a significant source of comparative data, has relied on online-only surveys without nationally representative demographic quotas and has prompted respondents with a predetermined list of media outlets, a method that overlooks the diversity of a fragmented media system (Newman et al., 2020, 2021). This study overcomes these limitations by employing a design tailored to the Turkish context, thereby addressing a critical geographic and methodological gap in the literature on news trust through the development of a grounded theory methodology, which is widely recognized as a popular approach in journalism research (Steensen & Ahva, 2017, p. 11).

2.2. Methodology

We implemented a two-phase sequential design. Phase 1 (survey) provided a diagnostic map of trust and tested RQ1–RQ2. Phase 2 (qualitative interviews) explained the mechanisms behind those patterns and addressed RQ3–RQ5. Phase 2 contained two strata: 2a earthquake-affected provinces and 2b Istanbul.

2.2.1. Phase 1: National Survey

We conducted a CATI survey of $n = 1,000$ adults (18+) across 12 provinces representing the 12 NUTS-1 regions in Turkey in May 24–June 13, 2022. The survey was fielded by Istanbul Ekonomi Araştırma, an independent public-opinion firm. We designed the questionnaire, and the researchers received only de-identified datasets. Interviews averaged 25 minutes and covered media use, platform trust, outlet trust, and political trust. Phase 1 analyses addressed RQ1–RQ2.

We used stratified random sampling and applied post-stratification weights for sex, age, education, and household income. The survey drew from a national dual-frame telephone list covering landline and mobile numbers. Within each of the 12 NUTS-1 regions, one focal province and sampled adults (18+) via stratified random digit dialing with regional quotas for sex and age were selected. Post-stratification (raking) weights were calculated to align the sample with TURKSTAT benchmarks for sex, age, education, and household income at the national level. We report ± 3.5 percentage points as a conservative margin of error for national proportions, noting that subgroup estimates have larger uncertainty.

2.2.2. Phase 2: Semi-structured Interviews

In Phase 2a (earthquake-affected provinces), we conducted $n = 50$ telephone interviews between February 14 and March 31, 2023 in Kahramanmaraş, Hatay, Adıyaman, Gaziantep, Malatya, Şanlıurfa, Diyarbakır, and Adana via random-digit

dialing. In Phase 2b (Istanbul), we conducted $n = 24$ in-person interviews in March 2023 using purposive sampling in diverse public locations. In both strata, we administered an adapted Kohring and Matthes (2007) “Trust in News” module and then probed reasoning processes through semi-structured questions. Phase 2 analyses addressed RQ3–RQ5.

Phase 1 identified alignment-based patterns that Phase 2 then explained in terms of unsurprisability, affective self-fact-checking, and the effect of the Other. Findings are integrated at the interpretive level, with Phase 2 mechanisms used to explain Phase 1 regularities. For comparison on outlet-level trust, we recruited 28 experts (senior journalists, union and association representatives, and media scholars) to rate the 11 most frequently mentioned outlets on a 10-point editorial independence scale (1 = “government-controlled”; 10 = “independent”). Panelists rated independently via an online instrument (median time = 4 minutes).

2.2.3. Research Timing as a Methodological Asset

The qualitative fieldwork coincided with the February 2023 earthquakes and the run-up to the May 2023 elections. Following comparative politics work on critical junctures, we treat this period as an observational window that made latent mechanisms more observable. We do not claim causal identification of crisis effects. Rather, we argue that the crisis context coincided with conditions under which affective processes are salient, improving construct clarity for the mechanisms theorized in RQ3–RQ5.

Data collection continued until theoretical saturation was achieved. The data was segmented and coded for consistency using the qualitative content analysis software QCAmap (Mayring & Fenzl, 2016). Following the open coding process, axial coding was conducted by grouping codes into broader categories based on similarities and differences. The core categories, representing the central focus of the grounded theory, was then identified in the focused coding process through which 33 key ideas were captured and then consolidated with into six categories through axial coding. The process was conducted flexibly and iteratively, with data collection via interviews.

Considering the preliminary data and their *a priori* knowledge in their domain of expertise, researchers attempted to generate a theory grounded in empirical evidence rather than preconceived notions or assumptions. The Constructivist Grounded Theory approach, which allows for more flexible and open-ended coding compared to classic or Straussian grounded theory, was utilized (Charmaz, 2008). This approach was deemed suitable for studying processes and emergent meanings, such as the relationship between journalism and trust (Charmaz, 2017).

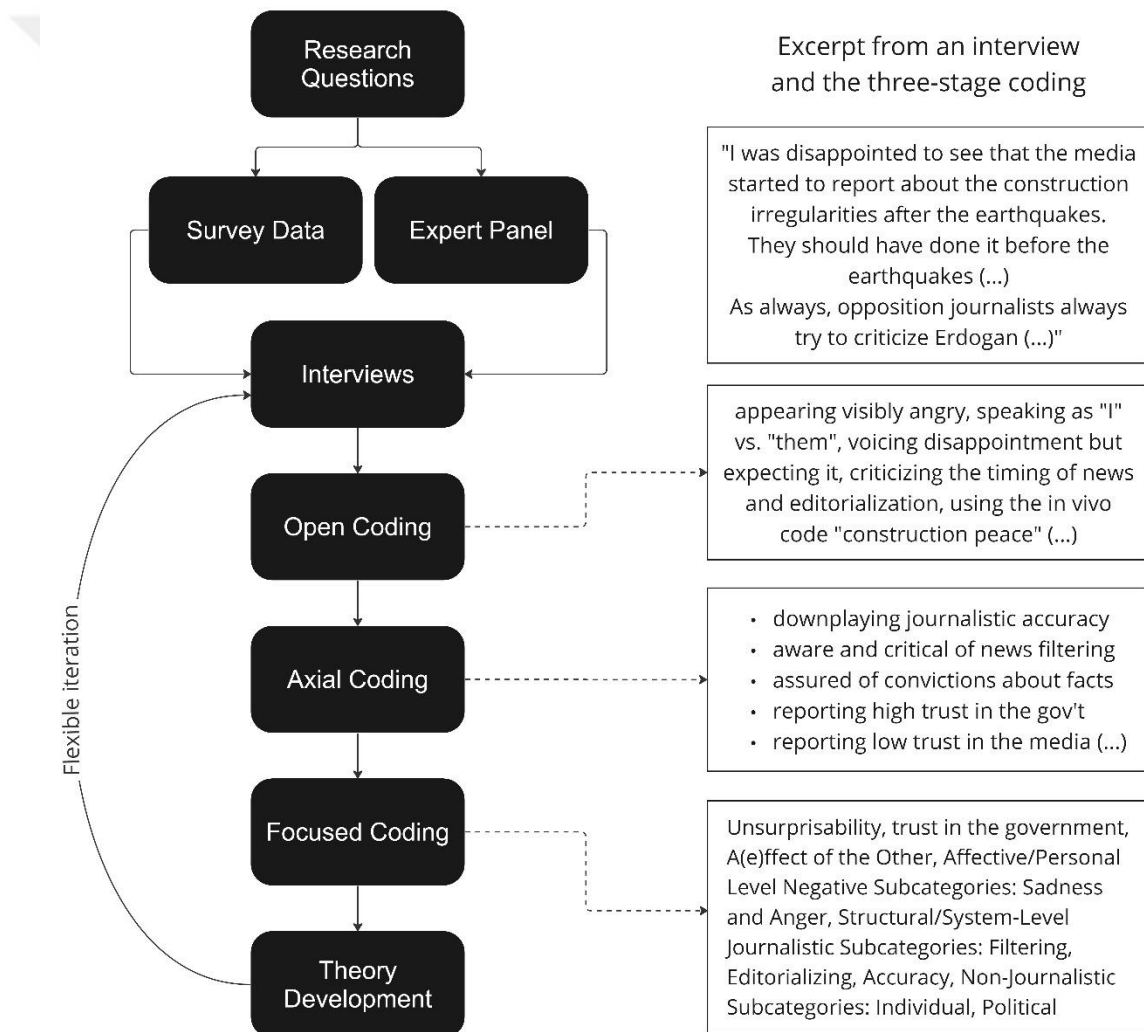


Figure 2.1: The Study Flow

2.3. Results

2.3.1. Survey

2.3.1.1. News Consumption Frequency and Preferred Platforms

68% of respondents indicated that they considered it "very important" or "important" to stay abreast of the latest news. Among them, 22% reported constant news consumption, while 33.2% and 34.4% cited frequent and occasional consumption, respectively.

Television was the most popular news medium, with 70% of respondents relying on it for information. Among digital platforms, news websites were the most popular news medium at 40.9%, followed by Google (34.3%), Instagram (20.7%), Twitter (20.3%), YouTube (17.7%), Facebook (14.5%), and WhatsApp (11.1%). None of the other platforms, including print media and digital news aggregators, surpassed the 10% mark.

2.3.1.2. Most and Least Trusted Salient News Brands

Of all respondents, 49.7% reported "strong trust" or "trust" in the five news outlets they listed, while 19.5% expressed "strong distrust" or "distrust," while 30.8% fell into the neutral category (a score of 3). The notable size of this neutral group indicates that for a large segment of the audience, prominent news brands evoke a sense of ambivalence or impartiality, rather than a strong conviction of either trust or distrust.

Table 2.1: Trust Ratings of the Most Frequently Mentioned News Outlets

News Outlet	N	Average Trust (1-5)
Halk TV	138	3.77
Fox TV	503	3.62
CNN Turk	255	3.58
A Haber	266	3.43
NTV	209	3.43
Show TV	311	3.42
Habertürk	307	3.39
TRT	529	3.37
Kanal D	267	3.33
Star TV	243	3.31
ATV	376	3.30

Note. When participants were asked to name five news outlets, only 11 were mentioned by more than 100 individuals. An additional 138 outlets were also cited a total of 1,074 times, but none of these reached the 100-mention threshold. Finally, 24 participants selected "No answer" for all five spaces on the questionnaire.

Among the most frequently mentioned news outlets, the reported “distrust” and “strong distrust” are highest for A Haber (23.4%), Star TV (23%), Kanal D (22.4%), ATV (21.3%), TRT (21%), Show TV (20%), NTV (18.4%), Habertürk (17.3%) and CNN Türk (16.2%). Of the ten outlets with the highest distrust ratings, all have clear ties with the government. TRT is the state-owned public broadcaster. Private TV stations A Haber and ATV are owned by the Turkuvaz Media Group, a corporation with close ties to President Erdoğan. The other major outlets mentioned—Star TV and NTV (owned by Doğuş Group), CNN Türk and Kanal D (Demirören Group), and Show TV and Habertürk (owned by the Ciner Group at the time of this study)—belong to three large conglomerates that have been awarded lucrative public tenders by the government in recent years (Reporters Without Borders, 2023). Consequently, the expert panel also rated all of these outlets as being closer to the “government-controlled” end of the spectrum.

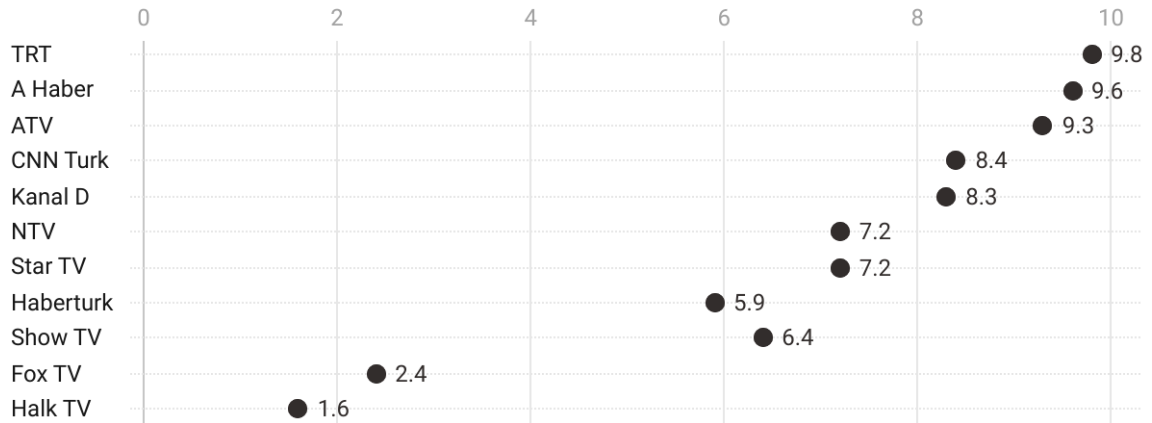


Figure 2.2: Expert Evaluation for the Most Frequently Mentioned Outlets'
(More Pro-Government If Closer to 10)

In sharp contrast were the remaining two outlets: Halk TV and Fox TV (later renamed as NOW TV). These channels function as the primary television voices for government opponents. Fox TV, owned by The Walt Disney Company at the time of our survey, is a mainstream commercial broadcaster whose primetime news bulletin has maintained a reputation for critical journalism, making it one of the most-watched news sources for citizens of all political leanings seeking alternative viewpoints. Halk TV, historically associated with the main opposition Republican People's Party (CHP), consistently features news coverage, as well as pundits and politicians, critical of the government policies.

2.3.1.3. Relationship of Trust in the Government with Trust in News Media

The demographic analysis of trust in news outlets reveals relatively small differences across gender, with women ($M = 3.47$) reporting only slightly higher levels of trust than men ($M = 3.42$). This trend holds true for most of the most salient outlets, with the notable exceptions of Kanal D, which men trust more, and ATV, where trust levels are equal between men and women. Age shows a more consistent pattern: younger respondents tend to express lower trust (18–24, $M = 3.33$; 25–34, $M = 3.29$). Trust increases gradually with age, peaking among respondents aged 65 and above ($M = 3.61$).

Educational attainment does not display a strong linear effect, but there are some variations. Respondents with primary education or below report moderate levels of trust ($M = 3.46$), secondary school graduates are slightly lower ($M = 3.41$), and those with high school diplomas or equivalent report the highest average ($M = 3.48$). Interestingly, respondents with higher education exhibit somewhat lower trust ($M = 3.38$).

Household income presents the most fluctuation. Trust levels are relatively balanced across lower- and middle-income groups, ranging between 3.27 and 3.59 on average, but trust is highest among respondents in the 20,001–24,000 TL monthly income bracket ($M = 4.29$). By contrast, respondents in the highest income group (above 35,000 TL) express considerably lower trust ($M = 3.00$), suggesting that confidence in news outlets may peak in upper-middle brackets rather than at the extremes of the income distribution.

Table 2.2: Breakdown of Responses by Sex for Self-Reported Trust in News Outlets

News Outlet	All Participants	Male Participants	Female Participants
Halk TV	3.77	3.72	3.87
Fox TV	3.62	3.66	3.58
CNN Türk	3.58	3.54	3.64
NTV	3.43	3.37	3.50
A Haber	3.43	3.39	3.49
Show TV	3.42	3.29	3.51
Habertürk	3.39	3.37	3.41
TRT	3.37	3.32	3.44
Kanal D	3.33	3.18	3.42
Star TV	3.31	3.22	3.37
ATV	3.30	3.30	3.30

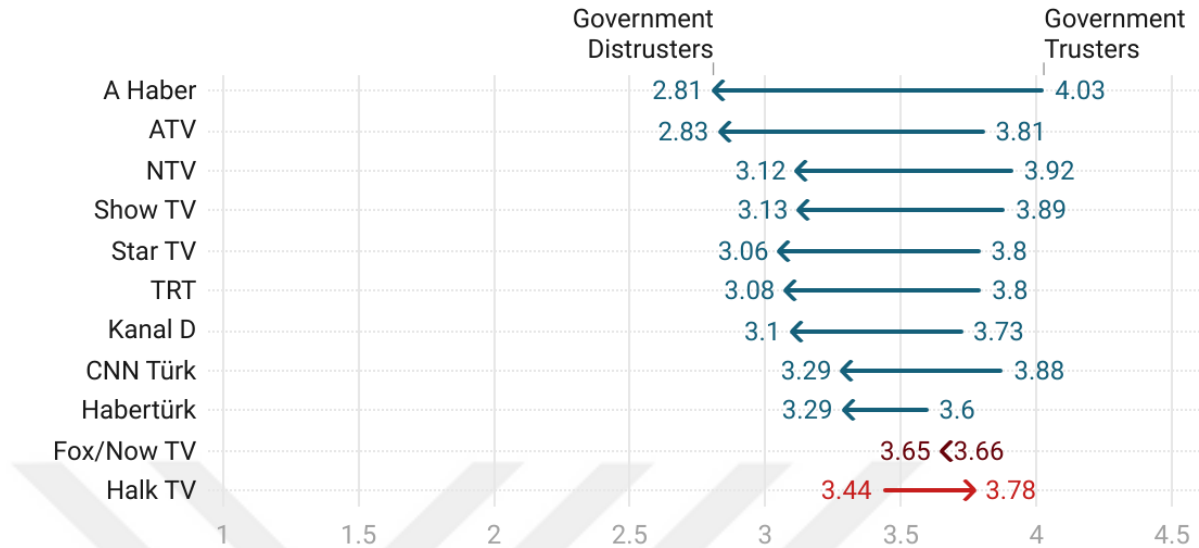


Figure 2.3: Media Outlet Trust Scores, Differentiated by Trust Levels in Government, Sorted by Percentage Change

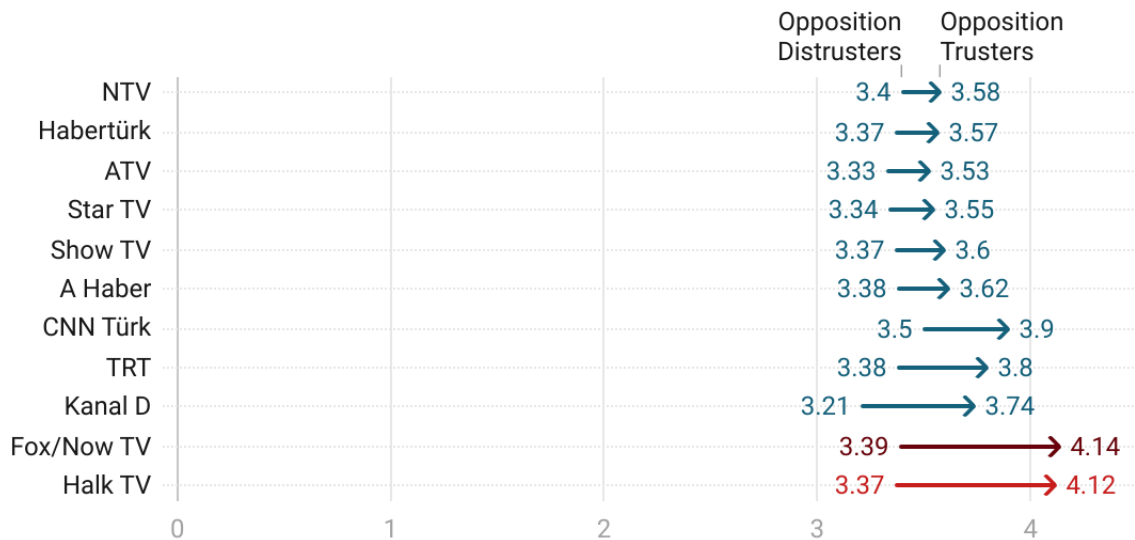


Figure 2.4: Media Outlet Trust Scores, Differentiated by Trust Levels in Opposition, Sorted by Percentage Change

The results point out to a clear asymmetry in how trust in political actors translates into trust in media outlets. Trust in the government sorts audiences into sharply different evaluations of outlets, while trust in the opposition is associated with a gentler and more uniform lift in outlet trust. In the government cleavage, government-aligned channels

receive strong trust premiums among government trusters and marked penalties among government distrusters. Independent outlets show the reverse pattern, with Halk TV evaluated favorably by government distrusters and discounted by government trusters. In the opposition cleavage, by contrast, trust increases more moderately across almost all outlets. Independent channels such as Halk TV and Fox/Now benefit most, but even government-aligned outlets gain slightly among opposition trusters.

This asymmetry is consistent with two mechanisms supported by prior work. First, motivated reasoning and selective exposure magnify alignment effects: people who trust the incumbent evaluate incumbent-aligned sources as more credible and out-group sources as less credible, reinforcing partisan sorting in perceived quality (Hart et al., 2009; Iyengar & Hahn, 2009; Stroud, 2010). The Turkish media system's supply-side structure, characterized by extensive government influence over media corporations that were once "mainstream", strengthens these identity cues and makes outlet alignment more salient in audience evaluations (Yeşil, 2016). Together, demand-side motivation and supply-side capture help explain the large dispersion observed along the government trust dimension.

Second, the opposition trust gradient appears less about selective endorsement of like-minded outlets and more about a generalized media-trust uplift, with the largest benefits accruing to independent channels but small positive movement visible even for pro-government brands. Two features can account for this. One is that "opposition distrusters" are not simply government supporters; this group includes more liberal, higher-educated respondents who tend to be generally skeptical of media, a pattern documented in research linking political sophistication and media skepticism to lower baseline trust and more critical news consumption (Tsfati & Cappella, 2003; Ladd, 2012). Another is that opposition trust does not map as cleanly onto a coherent media bloc in a system where most major TV outlets are state-aligned. While AKP, during its more than 20 years of governing, dominates the ruling bloc, CHP is far from having such control over the more fragmented, diverse political opposition.

In this context, without a dense ecosystem of mainstream media editorial independence, trust in the opposition raises trust primarily for a few alternative brands, while leaving others only modestly affected.

A third empirical regularity follows. Brand roles differ: Fox/Now TV functions as a relatively bridging outlet across government trust orientations, showing minimal change between government distrusters and trusters, whereas Halk TV is a sorting outlet, strongly favored by government distrusters and especially by opposition trusters. This is consistent with theories of cross-cutting appeal versus partisan signaling, where some brands accumulate heterogeneous audiences while others serve as identity markers (Iyengar & Hahn, 2009; Stroud, 2010).

The correlation analysis further contextualizes these asymmetries. Trust in government is only weakly associated with overall media trust ($r = .03$), whereas trust in the opposition shows a somewhat stronger, though still modest, positive relationship ($r = .19$). This pattern aligns with the descriptive results: opposition trust corresponds to a general uplift in trust across salient outlets, while government trust produces sharper outlet-level differentiation without translating into higher trust in the media system as a whole. By contrast, satisfaction with the media environment is strongly linked to government trust ($r = .32$) but virtually unrelated to opposition trust ($r \approx 0.01$). This suggests that government supporters evaluate the Turkish media system as delivering satisfactorily, which is consistent with extensive state influence over leading broadcasters (Yeşil, 2016). However, others, even those that express trust in the opposition, remain skeptical of the media landscape at large. In line with prior work on the ties between institutional trust and media evaluations (Ladd, 2012; Tsfati & Cappella, 2003), these results underscore how media trust is not merely an individual assessment of outlets but is conditioned by broader perceptions of political legitimacy and system performance.

In sum, the findings indicate that media trust in Turkey is structured by both audience motivation and the organizational configuration of the media sector. Government trust aligns closely with outlet ownership and produces steep divergences in perceived credibility, while opposition trust broadens trust more moderately across the system, with

the largest gains concentrated in independent media. This pattern is theoretically coherent with motivated reasoning and selective exposure, amplified by a captured media market and tempered by generalized skepticism among politically sophisticated publics.

2.3.1.4. Associations Between Trust and Platforms

Participants were provided with a scale from 1 to 5 to rate their trust in news across different platforms, with 1 being the lowest and 5 being the highest level of trust. None of the platforms received a trust rating exceeding 40% when considering high (4-point) and highest (5-point) levels of trust. The highest-rated platforms were TV (36.7%), websites (32.7%), and print media (32.4%). On the other hand, podcasts (14.6%), social media (19%), and newsletters (20.9%) were rated the lowest in terms of trust.

Trust in Turkish news content showed significant positive correlations across various platforms. Self-reported trust was strongly correlated with podcasts/newsletters ($r = .763$), news websites/news apps (.737), print/TV (.706), news apps/aggregators (.696), and TV/news websites (.689). In contrast, the weakest correlations were observed between TV/social media (.396), print/social media (.423), TV/search engines (.445), radio/social media (.450), and TV/digital video streaming platforms (.463).

Table 2.3: Relationship Between Trust in News Outlets with Interest in News

Participant Group	N	Trust in News Outlets
High interest in news	558	4.41
Low interest in news	442	2.07

Note. The groups are defined based on a 5-point scale where participants who rated their interest in following the news as 3 or higher are labeled as "high" and those who rated it less than 3 are labeled as "low."

As per the level of audience interest in news, Table 2.3 and Table 2.4 show a stark, graded "trust uplift" tied to engagement. People with high interest in news report more than double the outlet-level trust of low-interest respondents (4.41 vs. 2.07 on a 5-point scale). That same split repeats across platforms: among high-interest (or high outlet-trust)

respondents, TV is most trusted ($\sim 3.9\text{--}4.0$), print and news websites follow ($\sim 3.6\text{--}3.7$), Google sits lower ($\sim 3.3\text{--}3.35$), and social media remains lowest ($\sim 2.9\text{--}2.95$). For low-interest (or low outlet-trust) groups, trust compresses near the floor ($\approx 1.9\text{--}2.0$ for TV/print/web/Google; ~ 1.67 for social), and platform differences largely vanish. The uplift is largest for TV ($\approx +2.0$ points) and smallest for social ($+ \sim 1.2\text{--}1.3$), suggesting engagement amplifies trust most for legacy broadcast news while skepticism toward social media persists even among highly engaged audiences. The near-identical patterns when segmenting by “interest” and by “outlet trust” imply a common disposition toward trusting news that generalizes across platforms rather than being platform-specific. These are descriptive associations and should not be read causally.

Table 2.4: Self-Reported Trust in News-Distributing Mediums

Participant Group	N	TV	Print	Website	Google	Social Media
High interest in news	558	3.91	3.63	3.63	3.30	2.90
Low interest in news	442	1.95	1.95	1.94	1.98	1.67
High trust in outlets	537	3.99	3.69	3.66	3.35	2.95
Low trust in outlets	463	1.94	1.94	1.99	1.99	1.67

Note. Groups who rated their mean trust as 3 or higher on a 5-point scale for both categories were labeled as "high," and those who rated them less than 3 labeled as "low."

Additionally for the whole sample, we found a moderate association between trust in the news outlets mentioned by participants and their self-reported trust in news on Google's search engine (.496). Education level had a significant impact on trust for both news websites and news served by Google's search engine. Participants who reported paying more attention to following the news tended to prefer TV as a medium (.557). Trust in TV news was more strongly associated with participants' satisfaction with news media (.549). The level of agreement with the statement "independent journalism is essential for democracy" demonstrated a moderate association with the level of news consumption (.611).

2.3.2. Interviews

Qualitative data from the semi-structured sessions are addressed in the coming sections as part of the grounded theory, while the subsequent tables present descriptive statistics from the categorical data in response to pertinent interview questions.

Table 2.5: Measurements for Three Participant Groups in “Trust in News” Scale

Most Preferred Outlet	N	Topics	Facts	Accuracy	Editorial
Salient Pro-Government	25	4.99	4.18	4.33	2.29
Salient Independent	21	3.16	3.00	3.75	3.75
Non-salient/Alternative	28	3.51	2.80	3.04	3.49

In Table 2.5, The Kohring and Matthes’ scale measures four dimensions of trust in news (selectivity of topics, selectivity of facts, accuracy of depictions, and journalistic assessment) in 16 questions that we adapted to the context. Participants rated statements such as "The topic of the earthquakes receives the necessary attention" on a 1-7 scale, indicating the extent to which they believed the statement to be true based on coverage by their preferred news outlets. “Salient” outlet in the table refers to the ones that were most frequently mentioned in the survey phase. The table presents the mean values for three groups.

Table 2.6: Breakdown of Aggregated Responses for the Preferred Mediums

Preferred Medium	N	Topics	Facts	Accuracy	Editorial
Internet	46	3.79	3.15	3.46	3.88
TV	28	4.33	3.72	4.09	3.73

Table 2.7: Breakdown of Aggregated Responses for Trust in the Government

Trust in the Government	N	Topics	Facts	Accuracy	Editorial
Government Trusters	34	4.95	4.30	4.45	4.30
Government Distrusters	29	3.64	3.28	3.37	3.62

Note. The breakdown of aggregated responses for the participants who have the highest (4 and 5 on a 5-point Likert scale) and the lowest (1 and 2) means for trust in the government.

Both groups regarding the level of trust in the government registered medium to low satisfaction levels in the news media across all dimensions. The most substantial discrepancy occurred within the selectivity of topics dimension (1.32 points), while the smallest difference was found in the journalistic assessment dimension (0.68 points).

In total, 32 participants indicated that their trust in either their preferred media outlet or the government remained unaltered following the earthquake. Conversely, 18 respondents expressed diminished trust in both entities. Additionally, 13 participants reported decreased trust in the government, while their trust in their preferred media outlet remained constant.

2.4. Discussion

Our findings indicate a single organizing cleavage in Turkish news trust: attitudes toward the government. Government trusters rate government-aligned broadcasters much higher and discount independent outlets; government distrusters do the reverse. Trust in the opposition, by contrast, is associated with a milder system-wide uplift, largest for independent brands. ATiN explains this asymmetry by locating trust at the intersection of affect and interpretation through three appraisals—unsurprisability, affective self-fact-checking, and the effect of the Other.

ATiN links the system-level cues of Turkey’s media market to individual-level trust judgments through three affect-regulated appraisals. When headlines and brand signals

align with one's political identity, unsurprisability lowers vigilance and makes congenial content feel credible. When they clash, affective self-fact-checking activates: people scrutinize, seek disconfirming details in opponents' coverage, and discount the source to protect a preferred emotional stance. Throughout, identity markers and alignment labels map actors into in-groups and out-groups; the A(e)ffect of the Other channels sympathy toward co-partisans and suspicion toward rivals, shaping how identical cues are interpreted. In a media system where alignment is highly legible, these appraisals produce the observed asymmetry: sharp sorting on the government-trust axis and only a modest, diffuse uplift associated with opposition trust, even when surface-level journalistic quality is comparable.

Figure 2.5 summarizes the ATiN process from system-level cues through identity-filtered appraisals to trust judgments. The Structural / System Level represents the objective realities and actors that serve as triggers for trust evaluation. These are filtered through a person's political identity and result in a response at the Affective / Personal Level. The connection is direct: a trigger at the system level is interpreted through a social and political lens, producing a specific emotional outcome that determines trust.

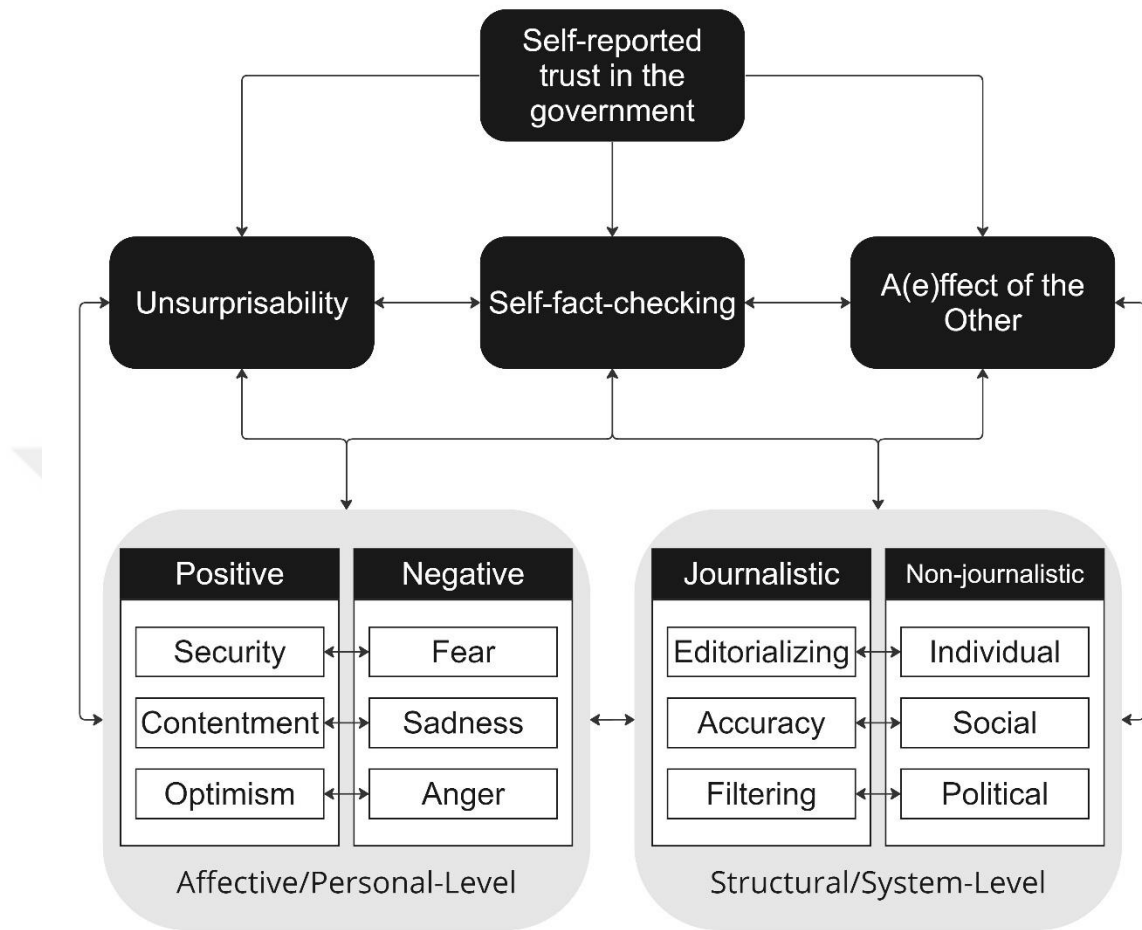


Figure 2.5: A Schematic Representation of the Affective Trust in News

2.4.1. Unsurprisability

A central mechanism we identify is “*şaşırmazlık*,” a Turkish term denoting a state of being unsurprised that arises not from apathy but from affective resignation to an expected negative reality. To capture this phenomenon, we adopt the neologism “unsurprisability,” because existing Anglophone terms are insufficient. Expectancy confirmation refers to a largely cognitive process in which an outcome matches a prediction, but it lacks the emotional texture conveyed by “*şaşırmazlık*.” Cynicism, by contrast, implies a generalized disposition of distrust. Unsurprisability is more precise: it functions as an affective stabilizer that neutralizes the potential for emotional shock and cognitive dissonance by anticipating bias as a routine feature of the media system. It is the learned, resigned acceptance that media outlets will act according to their political

allegiance. This affective shield works for both sides: it manifests as calm acceptance when a trusted, pro-government source filters news to "protect" the state, and as weary, anger-reinforcing confirmation when a distrusted source does the same. In both cases, the emotional equilibrium is protected. Mustafa, a 34-year-old male participant from the partially destroyed city of Gaziantep, illustrated this dual nature perfectly:

The news sources that I trusted did not disappoint me, while most of the mainstream media outlets that I did not trust were extremely careful to not publish or broadcast anything that could cause a headache for them [indicating potential repercussions from the government due to the critical news coverage].

Zuhal, a 30-year-old female participant from the devastated city of Adiyaman, shared a similar perspective, asserting that she "proved herself to be right concerning [the media outlets she] trusted and distrusted" previously. Unsurprisability, however, does not imply that participants were oblivious to the absence of quality journalism. Many participants followed news even from outlets they distrusted, particularly TV. They identified a lack of critical coverage and factual inaccuracies. For those who distrusted the government, this led to anger towards journalists of the "captured media".

Within this framework, participants who reported a higher level of trust in the government expressed a lack of surprise that media outlets, including privately-owned ones, were typically controlled by political authorities. They clearly articulated their disbelief in the concept of "journalistic independence," citing their perception of an inevitable necessity to filter the news because "no one, including reporters, is doing a job for charity" (Halit, 50-year-old male resident from Istanbul).

2.4.2: Affective Self-fact-checking

Journalists are expected to provide an accurate and comprehensive account of the day's events and a reliable, fact-corresponding picture of the world so citizens can participate in public life; distortions are ethical failures because they do not serve the public interest (Iggers, 1998, p. 47). Participants across the political spectrum evaluated

media performance against these standards, but they did so through a deeply biased process. Both government supporters and detractors engaged in what we term affective self-fact-checking. This process expands upon the established concept of motivated reasoning. While motivated reasoning describes how individuals selectively process information to confirm prior beliefs (a cognitive act), affective self-fact-checking posits that the primary drive is to protect a foundational emotional state linked to one's political allegiance. The goal is not merely to be factually "right," but to maintain and validate core feelings: be it anger at the state's failures or security in its competence.

This distinction emerged clearly in the data. The earthquakes directly affected millions, and many more had relatives or acquaintances in the impact zone, creating a nationwide community of eyewitnesses who could rapidly verify media accounts and identify omissions. When official reports diverged from observed conditions, participants who already distrusted the government treated the gap as evidence of state incompetence and media complicity, which reinforced anger and a sense of righteous indignation. This dynamic was associated with lower self-reported trust in both government and mainstream news outlets and was articulated through negative affect in accounts of individual and political failures:

News stories and commentators on TRT kept claiming that the state has been here from the first day to help the citizens. But I saw with my own eyes that it was not. I live in the city center, and even there was a nightmare for days. I lost several friends and relatives who lived in Nurdağı and İslahiye (the worst affected districts of Gaziantep, near the epicenter of the earthquakes). It was the apocalypse there and we could not see their pains on the TV. This undermined my trust in the state, too. The state left its people alone by themselves. The state has been managed badly in recent years. They do wrong things (Okkes, 55-year-old male resident from Gaziantep).

Individuals who reported higher levels of trust in the government also engaged in a self-fact-checking process to verify the filtering and accuracy of news media. However, this process differed in nature, as it primarily manifested through positive emotions at the

affective-personal level and strongly interacted with the social factor at the Structural/System Level. Even participants who were affected by the earthquake and identified factual discrepancies between the reality on the ground and news reports expressed contentment, security, and optimism. They directed their suspicion towards domestic critics of the government and foreigners, while tending to express sadness rather than anger when describing the process that eroded their trust, referring to external factors as “HAARP”, an online conspiracy theory that blames the instruments of the U.S.-based High-Frequency Active Auroral Research Program for triggering earthquakes thousands of kilometers away.

2.4.3. The A(e)ffect of the Other

The participants categorized various actors as "the Other" in their descriptions, such as ruling party officials, opposition representatives, pro-government or independent journalists, foreign powers, and even earthquake victims. The substantial impact of the Other was observed in their affective responses, which appeared to be in continuous interplay within other subcategories, transitioning between individual, social, and political domains.

Participants who reported higher levels of trust in the government formed another affective community that both influenced and was influenced by affective politics through the mediation of news media. The emotions, feelings, and moods they described during the inter-views interacted with the same subcategories, including unsurprisability, but this time, they exhibited positive emotions such as contentment, security, and optimism when discussing the government and the media. Another commonly observed pattern was that this group equated the "government" with the "state."

In the statements made by participants, the word "individual" often appeared as the pronoun "I," and "social" was depicted as "we." In this context, the "political" realm was defined as a non-journalistic, structural, and system-wide area that connects the "individual" and "social" spheres. The individual domain deals with the behaviors, actions, and interests of persons, while the social domain focuses on the behaviors,

actions, and interests of groups. The political domain includes the behaviors, actions, and interests of both persons and groups in terms of power dynamics and decision-making within a broader community or society.

For the participants who distrust the government and its media, the Other referred to pro-government journalists who they held responsible. İlker, a 31-year-old male resident from Adana, labelled them as “sold” journalists. But the A(e)ffect of the Other was also observed among participants who maintained high trust in the government. Ahsen, a 27-year-old female resident of Malatya, referred exclusively to highly pro-government outlets as sources of reliable information about the earthquakes. She expressed positive sentiments only toward two digital startups, GZT and Gloob, which were launched by pro-government outlets Yeni Şafak and Haber Global. Both in-house startups produce video news content tailored for younger users of social platforms such as TikTok and Instagram. Their content emphasized opinion and commentary presented by generally young reporters, which some participants favored over TV punditry.

In this regard, the "individual" plays a role not only at the affective/personal level (the news consumer as the "trust subject") but also at the structural/system level. "Individual" politicians like Erdoğan or "puppet" or “brave” journalists are among the "trust objects." This same process socially drives the affective communities of "we" and "they," building or undermining trust in news media and the government in interaction with the journalistic elements of the system.

The A(e)ffect of the Other was intertwined with discourse, delineating trust communities through distinct linguistic labeling. Trust appraisals were relational, organized around in-group vs. out-group labels signaled by language and actors. The same policy was read through divergent frames that cued opposing emotions. This was starkly evident in the language used to describe the government policies that had legalized unpermitted construction projects, many of which collapsed in the earthquakes. Participants with higher trust in the government consistently used the official state terminology: "construction peace" (in Turkish, *imar barışı*). This phrase, promoted by pro-government media, frames the policy as a positive act of reconciliation between citizens

and the state. The linguistic choice was therefore a political and moral statement. The affective connotations were clear: using "peace" signaled contentment with the government's narrative and optimism for the future, while using "amnesty" was an expression of anger at perceived corruption and fear for public safety.

In direct opposition, participants with lower trust in the government used the term deployed by critics: "construction amnesty" (in Turkish, *imar affi*). This is a critical distinction for an international audience. The term "amnesty" was deliberately used to frame the policy not as peace, but as a politically motivated pardon for the crime of unsafe building practices. It implies that the government forgave structural violations of building contractors, many of whom had ties to the ruling party, that directly led to the catastrophic loss of life.

This linguistic divergence is not an isolated phenomenon tied only to the earthquake but a systematic feature of Turkey's polarized media landscape. Countless examples reveal how the choice of a single word in a headline can signal an outlet's partisan allegiance and, in doing so, trigger a specific affective response that shapes audience trust. For instance, when reporting on rising consumer prices, pro-government media outlets consistently employ the euphemistic phrase "price adjustment" (in Turkish, *fiyat düzenlemesi*). This term frames the increase as a controlled, necessary, and technical administrative action, designed to soothe and reassure. In stark contrast, independent and opposition media label the same economic reality using the blunt and alarming term "hike" (in Turkish, *zam*). This word is chosen to evoke feelings of economic hardship, anxiety, and government failure. For the reader, encountering their preferred term in the headline validates their entire emotional worldview, reinforcing the boundaries of their affective community before they even read the article itself.

2.4.4. Why the Cleavage is So Steep (Demand vs. Supply)

Selective exposure and motivated reasoning on the demand side nudge audiences toward identity-congruent outlets and interpretations; once a source "feels right," unsurprisability lowers vigilance and reinforces prior affect. On the supply side, media

capture and ownership concentration make alignment cues unusually legible (through house styles, lexical markers, and personnel) so audiences don't need deep processing to infer "who speaks for whom." The result is a tight feedback loop: clear editorial signals meet identity-protective cognition, producing sharp sorting in credibility judgments even when informational quality is comparable.

2.4.5. Contributions and Testable Predictions

ATiN advances three claims: (1) trust judgments are affect-regulated appraisals rather than stable traits; (2) these appraisals are conditioned by highly legible alignment cues in the market; (3) the same micro-process yields either sharp sorting or diffuse uplift depending on supply structure. From these follow three predictions: (a) greater outlet-level variance along the government-alignment axis than along the opposition axis; (b) stronger interactions between alignment cues and unsurprisability among high partisans; (c) attenuated trust peaks among politically sophisticated audiences due to sustained self-fact-checking even for congenial content.

2.4.6. Boundary Conditions and Limitations

ATiN should generalize to polarized or hybrid media systems where alignment is publicly legible (e.g., strong state influence, concentrated ownership, disciplined partisan brands). It may be weaker in pluralistic systems with robust editorial independence and weaker cue clarity, or in low-salience domains where identity is not chronically activated. Our context, fielded around a national crisis and election, likely amplifies affect; portability should therefore be evaluated cautiously, with attention to timing, cue strength, and market structure.

There are several constraints that bound interpretation. This sequential mixed-methods study is cross-sectional; associations between outlet trust and political attitudes (e.g., trust in government) should not be interpreted causally. The CATI survey, while weighted to national benchmarks, remains vulnerable to coverage and non-response error as well as social-desirability bias in phone interviews; post-stratification reduces but does

not eliminate these risks. Fieldwork occurred around the February 2023 earthquakes and the election cycle, an atypically high-salience period that may have amplified affective responses; results are best read as a time-specific snapshot. Our outlet-level trust measure relies on top-of-mind recall, which over-represents salient television brands and under-captures long-tail or niche outlets; by design it reflects perceived, not enumerated, media diets. Dichotomizations used in descriptive contrasts (e.g., “high” ≥ 3 vs. “low” < 3 on 5-point items) are transparent but coarse and may mask within-group variation. The interview sample is purposive and concentrated in quake-affected provinces and Istanbul; it is not population-representative and is used to illuminate mechanisms rather than estimate prevalence. Finally, the alignment assessment of high-salience news brands comes from an expert panel; such coding is necessarily judgment-based and coarse (directional rather than graded), so borderline outlets and intensity differences should be interpreted cautiously.

2.5. Conclusion

This study investigated the determinants of trust in news brands within a highly polarized political and social environment. Conducted during a period of national crisis in Turkey, it moves beyond description to advance a new conceptual model. The principal contribution is the development and empirical grounding of the Affective Trust in News (ATiN) framework, which explains how structural features of the media system (ownership, journalistic practice) are filtered through political identity to elicit potent emotions (e.g. anger, fear, or contentment) that ultimately govern whether trust is granted or withheld.

ATiN posits that, in state-influenced and polarized systems, trust is fundamentally affective rather than a purely rational appraisal of journalistic quality. Through the mechanisms of unsurprisability, affective self-fact-checking, and the A(e)ffect of the Other, the framework offers a portable toolkit for understanding how trust is constructed, stabilized, and defended even when evidence conflicts with prior beliefs. Using Turkey as a critical case, we triangulated nationally representative survey evidence with qualitative interviews to show these dynamics in practice. A central implication is that “more facts”

alone are insufficient to bridge trust divides; because trust is embedded in identity, emotion, and power, interventions must engage those layers directly.

Several avenues merit further work. Future studies could examine how brand-level trust is mediated by content, medium, and algorithmic interfaces that curate digital news (e.g., search results, recommendation feeds). The linkage between mass-reach TV brands and their digital counterparts also warrants scrutiny in the context of audience trust. Lower trust in newer formats (e.g., podcasts) may reflect familiarity heuristics (Park & Lessig, 1981), a hypothesis that should be explicitly tested. Finally, comparative research could assess the portability of ATiN across distinct media systems (e.g., the Philippines, India, Brazil, South Africa, Hungary, and the United States).

In sum, the findings provide a foundation for systematic inquiry into trust in polarized societies, clarifying the role of emotion in trust formation and the potential influence of platform curation. Deepening our understanding of these relationships can inform scholarly, journalistic, and policy efforts to foster a more transparent, accountable, and trustworthy news environment. Building on these results, the next chapter examines whether LLMs and domain experts converge with or diverge from news audience trust by comparing public brand trust with source-blind ratings of headline quality.

3. THE PROPAGANDA BLIND SPOT: AI, MEDIA POLARIZATION, AND THE DISCONNECT BETWEEN TRUST AND NEWS QUALITY

The digital information landscape is undergoing a significant transformation. Driven by the proliferation of LLMs, search engines are rapidly transforming from lists of links into conversational "answer engines," and AI agents are beginning to curate our informational reality. In this emerging ecosystem, news is no longer simply sought from trusted brands but is actively surfaced, summarized, and synthesized by algorithms, potentially increasing the reach of smaller, independent publishers just as easily as established ones. This technological disruption collides with a pre-existing and deepening crisis of public trust in media, raising a critical question: how can we assess the quality and credibility of information when the very mechanisms of its discovery and consumption are being fundamentally rewritten?

3.1. Introduction

Nowhere are these dynamics more acute than in Turkey, which has been serving as an important testing ground for the future of the press at least since the early 2010s. Here, a vibrant and resilient independent journalism sector resists the pressures of an authoritarian government within a media system defined by capture and extreme political polarization. This high-stakes environment, further complicated by the pervasive influence of Big Tech and surveillance capitalism (Zuboff, 2020), provides a valuable context for examining how citizens navigate trust when the information space is a contested battleground. Studying this local context is a glimpse into the challenges

democratic societies worldwide will face in the age of AI-mediated communication and evolving platformization.

Traditional methods for studying media trust are ill-equipped for this new reality. Survey approaches that measure trust by asking respondents to rate a fixed list of legacy media brands are increasingly anachronistic, failing to capture the fragmented, algorithmically-driven nature of modern news consumption. Furthermore, much research overlooks the primary interface of media engagement: the headline. As studies consistently show that most citizens read little beyond the headline, these short fragments of text become the principal arena where trust is won or lost. Analyzing headlines is not only reflective of real-world user behavior but also offers a computationally efficient way to derive heuristic signals of quality and bias, a crucial advantage for developing scalable, real-time evaluation systems needed for the AI era.

While we understand the macro-level dynamics of polarization and have documented the decline of trust in Turkey, the precise relationship between specific, measurable attributes of journalistic content (particularly headlines) and perceptions of trust among polarized audiences remains under-investigated at a large scale. This study addresses this critical gap by pioneering a novel approach using LLMs to model and dissect these complex relationships. We ask:

RQ1: Replication & Simulation - To what extent can LLMs replicate the complex patterns of media trust, including ideological polarization and persona-specific biases, characteristic of human news consumers in Turkey?

RQ2: Evaluation & Limitation - How do LLMs' blind assessments of journalistic content quality compare to those of human experts, particularly in their ability to detect subtle, context-dependent manipulative framing and political bias (i.e., the "Propaganda Blind Spot")?

RQ3: Trust-Quality Nexus - What is the relationship between perceived brand trust and blind content quality, and how does this relationship differ between human audiences and an LLM's internal assessments?

By developing and validating an LLM-based methodology for quality assessment, this paper offers a dual contribution. Substantively, it provides a granular understanding of the content features driving trust and distrust in a captured media system. Methodologically, it presents a novel framework for using LLMs as scalable, theory-driven tools for computational communication science. The following sections review the literature on media trust, LLMs, and journalistic quality; detail our experimental methodology; present the results of our simulation; and discuss the implications for journalism studies and AI-assisted research.

3.1.1. The Crisis of Trust in Polarized Media Environments

3.1.1.1 Trust in Media: A Global Crisis of Credibility

The erosion of trust in news media represents one of the most significant challenges facing contemporary journalism and democratic societies worldwide. Over the past several decades, public confidence in media institutions has declined precipitously across diverse cultural and political contexts, reflecting broader patterns of institutional skepticism that extend beyond journalism to encompass government, academia, and other traditional pillars of democratic governance (Tsati & Ariely, 2014). This widespread decline in media trust is not merely a superficial shift in public opinion but rather a fundamental reconfiguration of the relationship between citizens and information institutions, with important implications for democratic discourse and civic engagement.

Longitudinal research demonstrates that this phenomenon transcends national boundaries and media systems, with cross-national studies revealing consistent patterns of declining trust in news organizations across both established and emerging democracies (Hanitzsch et al., 2018). The magnitude and consistency of this trend suggest that underlying structural forces, rather than isolated incidents or country-specific factors, are

driving the erosion of media credibility. These forces include the digital transformation of the information landscape, which has fundamentally altered how citizens access, consume, and evaluate news content (Strömbäck et al., 2020).

The fragmentation of media audiences in the digital age has created unprecedented challenges for traditional journalism's authority and legitimacy. The proliferation of alternative information sources has enabled selective exposure patterns that allow individuals to curate their information diets according to pre-existing beliefs and preferences, thereby undermining the shared epistemic foundation that once supported media credibility (Strömbäck et al., 2020). Simultaneously, social media platforms have emerged as vectors for misinformation and partisan content, creating an information ecosystem where credible journalism competes directly with unvetted and often deliberately misleading content (Ladendorf & Hanitzsch, 2021).

Theoretical frameworks for understanding this decline in media trust have coalesced around several key explanatory models. Institutional trust theory posits that declining confidence in media reflects broader skepticism toward establishment institutions, as journalism becomes increasingly perceived as part of an interconnected system of elite power structures (Tsfati & Ariely, 2014). Performance gap theory, alternatively, suggests that trust erosion results from perceived failures of journalism to meet normative expectations of objectivity, independence, and accuracy: expectations that may themselves be evolving in response to changing media environments (Strömbäck et al., 2020). Indeed, the fragmentation of media audiences directly exacerbates perceptions of a “performance gap” by allowing audiences to select media that confirms their priors about journalistic failure. Additionally, theories of cultural and political polarization highlight how partisan divisions and populist rhetoric targeting “mainstream media” have contributed to declining confidence among politically polarized segments of the population (Hanitzsch et al., 2018).

Cross-national comparative research reveals significant variation in both the levels and trajectories of media trust across different political and cultural contexts. While Nordic countries generally maintain relatively high levels of public confidence in

journalism, nations such as the United States and several Southern European countries have experienced particularly sharp declines in media credibility (Hanitzsch et al., 2018). This variation appears to be systematically related to broader characteristics of political systems, with media embedded in highly polarized political environments showing greater vulnerability to trust erosion (Tsfati & Ariely, 2014). These comparative insights suggest that while the decline in media trust represents a global phenomenon, its manifestations and underlying causes may be shaped by specific institutional and cultural contexts that require careful empirical investigation.

3.1.1.2. The Role of Political Polarization in Shaping Media Consumption

Political polarization, particularly in the context of media consumption, is increasingly understood through concepts such as affective polarization, echo chambers, filter bubbles, identity-protective cognition, and motivated reasoning. These mechanisms collectively explain how political identity not only shapes what individuals choose to consume but also how they interpret and trust information.

Affective polarization refers to the growing animosity between political groups, often manifesting as strong in-group favoritism and out-group hostility. Research shows that affective polarization increases the likelihood of believing misinformation that aligns with one's party and rejecting information from the opposing side, regardless of its accuracy (Jenke, 2023; Kubin & Von Sikorski, 2021). This dynamic is exacerbated by social media and partisan media environments, which reinforce negative perceptions of out-groups (Overgaard, 2024; Wilson et al., 2020).

Echo chambers and filter bubbles describe environments where individuals are primarily exposed to information that reinforces their existing beliefs. Studies demonstrate that both online and offline media consumption patterns are polarized, with users spending more time on like-minded news sources and forming distinct partisan communities (Garimella et al., 2021). Social media algorithms further limit exposure to counter-attitudinal content, deepening these echo chambers (Levy, 2020; Van Bavel et al., 2021).

Identity-protective cognition and motivated reasoning are psychological processes where individuals process information in ways that protect their political identity. Rather than evaluating information based on its quality, trust becomes an expression of political alignment. This leads to selective acceptance of information that supports one's group and skepticism toward opposing views (Bozdağ & Koçer, 2022; Almagro, 2022; Kim, 2017).

Empirical research supports the reinforcing spiral model, where partisan media use and political beliefs mutually reinforce each other, creating echo chambers that are more pronounced among conservatives than liberals (Hmielowski et al., 2020). Exposure to like-minded media consistently increases both ideological and affective polarization, while exposure to counter-attitudinal content can sometimes backfire, making individuals more entrenched in their beliefs (Kubin & Von Sikorski, 2021; Bail et al., 2018).

Field experiments suggest that even when individuals are exposed to opposing views on digital platforms, the effect can be an increase in polarization due to motivated reasoning and identity-protective cognition (Bail et al., 2018). Media coverage of polarization itself can moderate issue positions but also intensify dislike of the opposing party, highlighting the role of media in shaping perceptions of division (Levendusky & Malhotra, 2016; Yang et al., 2016).

In highly polarized environments, trust in media is less about objective quality and more about perceived political alignment. Users often heuristically validate information through partisan lenses, leading to "skeptical inertia": a self-aware but passive stance where individuals remain critical yet loyal to sources aligned with their identity (Bozdağ & Koçer, 2022; Wilson et al., 2020). This dynamic perpetuates polarization and undermines the potential for consensus or depolarization.

Table 3.1: Key Concepts in Political Polarization and Media

Concept	Effect on Media Consumption and Belief	Citations
Affective Polarization	Increases in-group trust, out-group hostility, belief in congruent misinformation	(Jenke, 2023; Kubin & Von Sikorski, 2021; Overgaard, 2024)
Echo Chambers	Reinforces existing beliefs, limits exposure to diverse views	(Levy, 2020; Van Bavel et al., 2021; Garimella et al., 2021)
Motivated Reasoning	Selective acceptance of information, trust as political alignment	(Bozdağ & Koçer, 2022; Almagro, 2022; Kim, 2017; Bail et al., 2018)
Polarization in Media Reports	Moderates issue positions, increases party dislike	(Levendusky & Malhotra, 2016; Yang et al., 2016; Wilson et al., 2020)

3.1.1.3. The Turkish News Landscape: Media Capture and Polarization

The Turkish media system is widely recognized as an archetype of media capture, characterized by state control, ownership concentration, and the erosion of journalistic autonomy. Over the past two decades, Turkey has experienced a rapid transformation from a pluralist-polarized media model to a hybrid system dominated by political parallelism, state intervention, and media conglomerate interests (Akser & Baybars, 2024; Yeşil, 2016; Kaya & Çakmur, 2010).

Media capture in Turkey is driven by a combination of state intervention, regulatory mechanisms, and the strategic use of media ownership to serve political interests. Successive governments, particularly under the long rule of the Justice and Development Party (AKP), have leveraged regulatory agencies and public broadcasters to align media content with state agendas, often employing censorship and legal pressures to suppress dissent (Akser & Baybars, 2024; Yeşil, 2016; Algan & Kaptan, 2023). The instrumentalization of media for ideological and political purposes is evident in both traditional and digital platforms, with the state exerting influence over content production, distribution, and ownership structures (Akser & Baybars, 2024; Yeşil, 2016; Algan & Kaptan, 2023).

The Turkish media landscape is marked by high ownership concentration, with major conglomerates controlling multiple outlets across television, print, and digital media. This concentration facilitates political parallelism, where media outlets act as proxies for competing political factions rather than independent watchdogs (Akser & Baybars, 2024; Yeşil, 2016; Kaya & Çakmur, 2010). The intertwining of economic and political interests has made media organizations vulnerable to state pressure, further undermining editorial independence (Yeşil, 2016; Kaya & Çakmur, 2010).

Journalistic autonomy in Turkey has been significantly eroded by both direct state intervention and the economic interests of media owners. The politicized judiciary and regulatory environment have criminalized dissent and restricted freedom of expression, leading to widespread self-censorship and a decline in professional journalistic standards (Akser & Baybars, 2024; Yeşil, 2016; Kaya & Çakmur, 2010). The rise of digital media has provided alternative platforms, but these too are subject to polarization, disinformation, and manipulation (Doğu, 2020; Kırdemir, 2020).

Turkey's media environment is highly polarized, reflecting broader societal divisions. Media outlets are deeply entrenched in political conflicts, contributing to a low-trust environment where audiences perceive news sources as partisan actors (Doğu, 2020; Kaya & Çakmur, 2010; Kırdemir, 2020). The proliferation of digital news sources has not alleviated polarization; instead, it has introduced new challenges related to disinformation and social manipulation (Doğu, 2020; Kırdemir, 2020).

This system of media capture in Turkey serves as a powerful real-world example of institutional trust theory, where media are perceived not as independent watchdogs but as integral components of an elite power structure, thereby fueling widespread skepticism.

Studying the relationship between content characteristics and audience trust in a highly polarized and captured media environment like Turkey presents significant methodological challenges. Traditional surveys are susceptible to partisan bias, while large-scale manual content analysis is prohibitively resource-intensive. To overcome these limitations, this study proposes a novel methodological approach that leverages LLMs as

tools for simulating audience perception and evaluating journalistic quality at scale. The following section, therefore, reviews the emerging literature on the capabilities, limitations, and appropriate uses of LLMs as tools for computational social science.

3.1.2. Large Language Models as Tools for Social Science Research

LLMs have rapidly emerged as tools for simulating human opinion and behavior, raising foundational questions in AI alignment, computational social science, and the nature of machine "understanding." This review synthesizes recent and foundational work on LLMs as social simulators, the use of synthetic respondents, and the debate over emergent behavior and true comprehension.

LLMs are increasingly used to simulate human-like agents for social science research, policy design, and behavioral studies. Early work focused on role-playing based on demographic information, but recent studies show that integrating empirically-derived human belief networks significantly improves alignment between LLM-generated and human opinions, especially for related topics within the same belief network (Chuang et al., 2024). This suggests that LLMs can be tuned to better reflect human belief distributions, but simple demographic prompts are insufficient.

2.1 LLMs as social simulators and synthetic respondents

LLMs have been deployed as synthetic survey respondents and proxies for human participants in psychological and economic experiments. While LLMs can mimic certain human biases and heuristics, such as loss aversion or context-sensitive reasoning, their responses often reflect the statistical patterns of their training data rather than genuine understanding (Sohail & Zhang, 2025; Leng, 2024). Multi-agent LLM architectures have shown promise in simulating complex social behaviors, such as strategic decision-making in economic games, with multi-agent setups outperforming single-agent models in replicating human-like strategies (Sreedhar & Chilton, 2024).

3.1.2.1. AI Alignment and Emergent Behavior

Alignment remains a central challenge. LLMs are typically aligned with human preferences through Reinforcement Learning from Human Feedback (RLHF), but this process can lead to models that mirror human expectations of rationality rather than actual human behavior (Liu et al., 2024; Chaudhari et al., 2024). For example, LLMs tend to assume people are more rational than they are, aligning more with expected value theory than with observed human choices (Liu et al., 2024; Leng, 2024). This misalignment has implications for the validity of LLM-based simulations in social science and policy contexts.

A persistent debate concerns whether LLMs "understand" or merely perform sophisticated pattern matching. Some argue that LLMs act as mirrors, reflecting the intelligence and biases of their users rather than possessing intrinsic understanding (Sejnowski, 2022; Shiffrin & Mitchell, 2023). While LLMs can produce human-like conceptual representations and even align with neural activity patterns in the brain, their lack of real-world grounding and reliance on statistical correlations limit claims of genuine comprehension (Xu et al., 2025; Shiffrin & Mitchell, 2023).

LLMs exhibit biases inherited from their training data, often overrepresenting Western, educated, and publicly expressed viewpoints, which can limit their utility as representative synthetic respondents (Sohail & Zhang, 2025). Additionally, emergent abilities such as deception and the potential for hallucinations raise ethical and practical concerns for their deployment in sensitive domains (Hagendorff, 2024; Farquhar et al., 2024).

3.1.3. Defining and Measuring Journalistic Quality

To effectively employ LLMs as evaluators of media content, it is essential to ground their analysis in established theoretical and methodological frameworks of journalistic quality. An LLM cannot assess "quality" in a vacuum; it requires clear, operationalizable criteria derived from decades of media studies research. This section provides a review of

these criteria, examining the normative principles that define high-quality journalism, the established methods developed to measure it, and the emerging computational approaches that offer new possibilities for analysis. This review establishes the foundation upon which our LLM-based measurement instrument is built.

3.1.3.1. Normative and Theoretical Foundations of Journalistic Quality

Journalistic quality is a foundational concern in media studies, encompassing both normative ideals and empirical assessment methods. The literature consistently highlights core concepts such as accuracy, impartiality, completeness, and adherence to other news values as central to defining quality in journalism.

Seminal works, such as Kovach and Rosenstiel's *The Elements of Journalism* (2007), articulate the guiding principles that underpin journalistic quality, including truth, independence, and loyalty to citizens. These principles emphasize that journalism's purpose is to provide comprehensive and fair information, with objectivity and verification as central tenets. The book also discusses the importance of monitoring power, providing a public forum, and maintaining ethical standards, all of which contribute to the believability and trustworthiness of news (Kovach & Rosenstiel, 2007).

Other studies reinforce these elements, noting that journalistic quality is not only about the content but also about the process, such as transparency, collaborative verification, and the journalist's role as a mediator in networked environments (Garcie de Torres & Hermida, 2017). The watchdog function and ethical standards are also highlighted as essential for maintaining fairness and completeness in reporting (Carrera et al., 2021).

3.1.3.2. Methodological Approaches to Measuring Trust and Quality

Translating these normative ideals into empirical measurement presents fundamental challenges. Contemporary scholarship has converged around two complementary paradigms: institutional-level assessments of source credibility and

content-level evaluations of information quality, which often rely on structured human evaluation.

At the institutional level, the theoretical foundation for measuring trust was significantly advanced by Kohring and Matthes (2007), who developed a hierarchical, multidimensional scale. Their framework conceptualizes trust in news media as a second-order construct encompassing four first-order dimensions: trust in the selectivity of topics, trust in the selectivity of facts, trust in the accuracy of depictions, and trust in journalistic assessment. This multidimensional approach allows researchers to disaggregate the complex phenomenon of media trust into interpretable components, proving invaluable for examining how specific journalistic practices relate to public trust.

Complementing institutional measures, content-level frameworks like the AIMQ (A methodology for information quality assessment) evaluate information quality across multiple dimensions as perceived by users (Lee, Strong, Kahn, & Wang, 2002). AIMQ distinguishes between intrinsic quality (e.g., accuracy, completeness), contextual quality (e.g., relevance, timeliness), representational quality (e.g., clarity, formatting), and accessibility (e.g., ease of access, security), providing a systematic approach to assessing the quality of informational products.

In practice, applying these frameworks often relies on systematic human coding and expert evaluation. Empirical assessment of journalistic quality frequently employs content analysis, where human coders or expert panels evaluate news content against established criteria. Triangulation methods, including interviews and focus groups, are also used to validate findings and ensure a comprehensive assessment (Carrera et al., 2021).

Table 3.2. Summary of Empirical Methods for Assessing Journalistic Quality

Measurement Approach	Description	Core Concepts Assessed	Citations
Content Analysis	Systematic coding of news content by experts or trained coders	Accuracy, objectivity, fairness	(Carrera et al., 2021; De Torres & Hermida, 2017)
Expert Panels	Evaluation by panels of journalism scholars or practitioners	Completeness, believability	(Carrera et al., 2021)
Collaborative Verification	Use of networked audiences for real-time fact-checking and transparency	Accuracy, trust, verification	(De Torres & Hermida, 2017)
Triangulation	Combining interviews, focus groups, and archival research	Ethics, standards, completeness	(Carrera et al., 2021)

3.1.3.3. Computational and AI-driven Approaches to Quality Assessment

While human-centric methods remain the gold standard for nuanced assessment, their limitations in scale and speed have fueled the development of computational and AI-driven approaches. The research frontier has moved toward partial automation of specific quality-related tasks, particularly in fact-checking and misinformation detection. Natural language processing (NLP) pipelines that combine claim detection, evidence retrieval, and textual entailment have shown promise for assisting human fact-checkers, supported by large annotated datasets like FEVER (Thorne et al., 2018).

More recently, the rapid advancement of LLMs has transformed the landscape of content evaluation. Studies comparing AI and human evaluators reveal nuanced outcomes. AI-enabled content analysis demonstrates high reliability and validity, often outperforming manual approaches in efficiency and consistency (Lee et al., 2020). However, these automated approaches have well-documented limitations. Current systems excel at identifying overt falsehoods but struggle with nuanced content, satire, and context-dependent meaning (Elkhatat et al., 2023; Saqib & Zia, 2024). More fundamentally, automated methods have limited capacity to capture the normative editorial judgments central to journalism, such as assessments of newsworthiness, ethical considerations, or the appropriateness of specific editorial choices (Thorne & Vlachos,

2018). Trustworthiness and factual accuracy also remain significant concerns, as AI-generated outputs can include plausible-sounding but incorrect information, necessitating human oversight (Wang et al., 2023; Cao et al., 2023).

Ongoing research thus emphasizes the need for more robust, explainable, and context-aware AI systems. Human involvement remains essential for ensuring quality, addressing ethical concerns, and interpreting nuanced content that current models cannot yet grasp (Wu et al., 2023). It is within this context, leveraging the scale of AI while grounding it in the rigor of human-defined journalistic principles, that our study is situated.

3.1.4. Research Gap, Questions, and Contribution

Central to the crisis of media credibility is a fundamental "trust gap" between the producers and consumers of news. On one side are producer-centric metrics of quality, rooted in the normative principles of journalism. On the other side are audience-centric metrics of trust, which are shaped by perception, partisan identity, brand loyalty, and lived experience. In a healthy democratic media system, these two forms of evaluation should largely align: content that meets high journalistic standards ought to be what the public deems most trustworthy.

In hyper-polarized media environments such as Turkey, however, this relationship breaks down. Public trust often becomes decoupled from, or even inversely correlated with, expert-defined quality. For highly partisan audiences, trust is no longer an evaluation of journalistic merit but an affective expression of political identity and in-group loyalty. The magnitude of this divergence between public trust and professional quality, therefore, serves as a key diagnostic of a media system's health. A widening gap signals deepening polarization, potential media capture, and the erosion of the shared epistemic foundations necessary for democratic discourse.

The preceding literature review reveals a clear research gap at the intersection of these challenges. While past research has extensively explored (1) the dynamics of human media trust, (2) the nascent simulation capabilities of LLMs, and (3) the methods for

automated content analysis, no study has yet integrated these three domains within a single, comparative framework. We lack a method to systematically diagnose the trust-quality relationship at scale and to understand how emerging AI technologies perceive this critical disconnect.

This chapter puts forward such an integration by uniquely placing human and LLM evaluations side-by-side across two distinct experimental tasks: a non-blind trust assessment based on brand reputation and a blind evaluation of news headline quality, as well as the visual design of news websites. This dual-task design allows us not only to compare the performance of LLMs against human benchmarks but also to interrogate the internal consistency of an LLM's own "reasoning", pitting what an LLM knows about a media brand's reputation against what it sees in the brand's content. This comparative approach is designed to expose novel limitations in AI-driven analysis, such as a potential "Propaganda Blind Spot," where technically well-formed content that adheres to stylistic norms masks underlying manipulative intent. In doing so, this study provides a new methodology for auditing the contextual understanding and latent biases of AI systems in the critical domain of public trust in news.

3.2. Methodology

This study employed a multi-stage, mixed-methods research design to investigate perceptions of trust in Turkish news media. The methodology was structured in two main parts. Part 1 involved a series of survey simulations to compare the stated trust perceptions of human news consumers with those of leading LLMs under various conditions. Part 2 consisted of a quantitative experiment to analyze the relationship between non-blind brand trust and blind evaluations of content quality across human consumers, journalistic experts, and LLMs.

Following common survey practice in trust-in-news research, outlet-level trust was measured with single-item 1–5 ratings for each respondent's top-of-mind outlets ("How much do you trust [OUTLET]?"; 1 = not at all, 5 = completely). We reference Kohring & Matthes (2007) as a theoretical framework for multidimensional trust, but we did not field

their multi-item battery in this study. All outlet trust scores used in analyses are single-item means aggregated at the outlet level. This choice aligned the human measure with the LLM tasks, which also produced single-item 1–5 outlet trust ratings, ensuring commensurability across evaluators.

For content-level trust, we implemented a blinded headline assessment in which all source cues were removed and raters evaluated each headline using an AIMQ-informed checklist that operationalizes intrinsic quality, contextual relevance, representational clarity, and accessibility signals that are assessable at the headline level. Interface and system usability elements not applicable to standalone headlines were excluded, and item responses were aggregated into a content-trust index to enable comparison with the outlet-trust measure (Lee, Strong, Kahn, & Wang, 2002).

3.2.1. Survey and Persona Simulations

This part of the study was conducted in four sequential phases designed to establish a human baseline of media trust and then probe LLM responses from neutral, comprehensive, and persona-driven perspectives.

3.2.1.1. Phase 1: Human Baseline Survey

This phase draws on the survey administered in Study 1, which measured attitudes toward news consumption with standardized items. The instrument captured demographics (age, gender, education, and province of residence) and included 5-point Likert scales for institutional trust in the incumbent government (“*mevcut hükümet*”) and the opposition (“*muhalefet*”), anchored from 1 “*hiç güvenmiyorum*” to 5 “*kesinlikle güveniyorum*.” Respondents also rated their trust in major platforms and formats for news, such as television, news websites, and social media, using the same 5-point scale. To identify salient brands, participants named up to five news organizations that came to mind first (“*Aklınıza ilk gelen 5 haber kuruluşunu belirtiniz.*”) and then rated their trust in each of these self-identified outlets on the same scale. Finally, an attitudinal item assessed agreement with the statement that “for a democracy to survive, journalism must be practiced independently of centers of power,” again on a 5-point scale.

The dataset comprises 1,000 completed interviews. Post-stratification weights align the marginal distributions of age, gender, and region with the Turkish population across the 12 NUTS-1 regions to maintain national representativeness. Data processing produced mean scores for institutional and platform trust, and it standardized the open-ended outlet mentions by mapping orthographic variants to canonical identifiers. For example, “FOX TV,” “FOX,” and “NOW TV” were consolidated under “Fox / Now TV.” We then computed the average trust score per canonical outlet across all respondents who rated it, yielding comparable brand-level trust estimates.

3.2.1.2. Phase 2: Baseline AI Survey Simulation

To establish a model-based baseline, we instructed LLMs to answer the same survey as if they were Turkish news consumers taking a neutral stance that reflects generalized information in training data. We selected five widely deployed, high-capacity systems to balance performance and provider diversity: The U.S.-based OpenAI’s GPT-5, Google’s Gemini 2.5 Pro, Anthropic’s Claude 4 Sonnet; Mistral Large-241 as a European provider, and Moonshot AI Kimi K2 as a leading China-based provider with agentic workflow features. This mix mitigates the risk that results are specific to a single vendor, model family, or jurisdiction.

These instruction-tuned models were not given vendor-specific chain-of-thought or “thinking” modes, and they were prompted only for final answers. Instruction prompts were in English. All stimuli, including outlet names, item wordings, and response anchors, were presented in Turkish, and models were explicitly instructed to respond in Turkish. To reduce cross-lingual drift in brand matching, outlet names were provided in canonical Turkish forms, such as “Fox / Now TV” and “Habertürk.” Decoding parameters were held at each provider’s documented defaults and kept constant across runs.

Each model completed the full survey 10 independent times, generating 50 responses in total. We parsed the resulting JSON outputs to extract numeric scores and top-of-mind outlets, then averaged each trust measure across the 50 responses to construct

an AI baseline directly comparable with the human data.

3.2.1.3. Phase 3: Expanded AI Trust Ratings for a Comprehensive Outlet List

Because the baseline simulation allowed models to self-select outlets that did not always coincide with the brands most salient to human respondents, this phase enforced a fixed evaluation set. We compiled the 20 most frequently mentioned canonical outlets from the human survey and prompted the same five models to rate trust in each outlet on the 1 to 5 scale. As in Phase 2, decoding settings were standardized, and each model produced 10 independent rating sets. Aggregating across the 50 model-level replications yielded a mean “LLM Average Trust” score for every outlet on the list. This procedure creates a one-to-one basis for comparing AI-generated trust indices with the human-generated brand scores for the most salient news organizations.

3.2.1.4. Phase 4: Persona-Based AI Survey Simulation

The final phase evaluated whether models can simulate distinct Turkish news consumer profiles that reflect known socio-political cleavages and media diets. Persona definitions were developed qualitatively by combining distributional patterns from the 2022 human survey, prior scholarship on polarization in Turkey, and our domain knowledge as media and communication professionals. We operationalized four archetypes: a Centrist Urbanite from Istanbul (male, 25–34), representing an educated urban demographic; a Conservative Nationalist from Ankara (male, 45–54), representing a core incumbent-leaning constituency; a Progressive Secularist from İzmir (female, 35–44), representing an opposition-aligned cohort; and a Young Skeptic from Diyarbakır (female, 18–24), representing youth with low trust in institutional politics and mainstream media in a Kurdish-majority local context.

Each model was instructed to adopt one persona at a time and complete the survey while providing top-of-mind outlets, trust ratings, and a brief qualitative justification of its choices. Decoding parameters matched those used in the baseline to avoid confounding persona effects with sampling variability. Every model–persona pairing was run 10 times,

and outputs were compiled into a single corpus. We parsed the JSON to extract persona-specific trust measures and top-of-mind brand sets and collected the accompanying rationales for thematic analysis to assess the coherence and authenticity of the simulations.

3.2.2. Blind vs. Non-Blind Evaluation of Content Quality

This component used a quantitative, comparative design to relate brand-visible trust judgments to source-blind assessments of headline quality. We compared three evaluator groups: members of the public providing non-blind brand trust, an expert panel rating headline quality under blind conditions, and LLMs producing both non-blind brand trust and blind headline quality scores.

To adapt AIMQ to headline-only, source-blind assessment, we retained text-level dimensions that can be judged from a single headline and renamed them to align with newsroom practice. The resulting rubric covered Accuracy, Objectivity, Believability, Clarity, Conciseness, Completeness, and Informational Adequacy. These map onto AIMQ’s Free-of-Error, Objectivity, Believability, Understandability and Interpretability, Appropriate Amount and Concise Representation, Completeness, and a relevance and amount composite that captures whether a headline offers sufficient context without access to the article body. We excluded dimensions that require brand, interface, or system context, including Reputation, Timeliness, Consistent Representation, and the Accessibility category, because they cannot be assessed reliably from a standalone headline stripped of all source cues.

All headline metrics were scored on a common 0 to 100 scale. Human experts and models applied the same rubric to the same set of 794 Turkish headlines presented without outlet names, bylines, section tags, timestamps, URLs, or outlet-specific glyphs. Models received rating instructions in English and were told explicitly to evaluate the texts as Turkish news headlines; the headlines themselves were in Turkish. We did not enable chain-of-thought or tool use. Models produced scalar ratings only, under provider-default decoding settings, and headline order was randomized at each call to minimize position effects.

Four data sources were integrated. Public non-blind brand trust came from the survey of 1,000 news consumers based on top-of-mind brand recall and subsequent trust ratings. Blind headline quality came from an expert panel of 10 senior professionals in Turkey, balanced by gender and spanning reporters, newsroom leaders and media executives, and journalism scholars, all with direct responsibility for headline production or quality control. Each expert rated every headline independently in a custom web interface using eight 0 to 100 sliders with anchors at 0, 50, and 100. Prior to any ratings, outlets were labeled as pro-government or independent using external ownership and affiliation mappings and long-run editorial stance; these labels were not visible during scoring. Model-based non-blind brand trust was constructed by combining the top-of-mind task and the fixed-list trust ratings produced by the five LLMs described earlier. Model-based blind headline quality was generated by applying the same rubric and blinding protocol to the same 794 headlines.

Data preparation proceeded in three steps. First, we standardized outlet names across all sources with an ordered normalization map that prioritized longer and more specific matches to prevent erroneous aggregation. Free-text variants such as “Fox Haber” and domain strings such as haberturk.com were mapped to canonical forms like “Fox / Now TV” and “Habertürk.” Second, we aggregated ratings to the outlet level, which served as the unit of analysis. For trust, we computed outlet means across all relevant raters, restricting human trust to outlets mentioned by at least two participants to stabilize estimates. For quality, we computed outlet means for each rubric dimension across all relevant raters, multiplying any 10-point inputs by 10 to place all metrics on the 0 to 100 scale. Third, we created two analytical datasets via inner joins: a human dataset merging public non-blind trust with expert blind quality, and a model dataset merging model non-blind trust with model blind quality. For each analysis, the final sample size equaled the number of outlets present in both the trust and quality inputs. We used the a priori publisher-leaning labels only for exploratory subgroup comparisons. Because these categories are coarse and context dependent, related findings are presented as illustrative rather than confirmatory.

Given the small number of outlets after merging, we relied on nonparametric and robust estimation. The primary association measure was Spearman’s rank correlation coefficient, which does not assume linearity and is appropriate for ordinal rating scales. We quantified uncertainty with 95 percent bootstrap confidence intervals based on 5,000 resamples. Objectivity served as the focal content-quality metric for trust–quality associations because of its centrality in journalistic standards. Other dimensions, including Clarity, Believability, Accuracy, Informational Adequacy, Completeness, and Conciseness, were examined as exploratory outcomes. Because outlet-level samples were small, we emphasized effect sizes and interval estimates over null-hypothesis tests. Where p-values are reported, they are unadjusted.

To ensure transparency and enable replication, all datasets, analysis scripts, and supplementary materials are publicly available on the Open Science Framework (OSF; DOI: 10.17605/OSF.IO/RM547). The repository provides raw and processed data in CSV and JSON formats, along with the complete data-processing and statistical analysis pipeline implemented in Python 3.11 using pandas, numpy, scipy, pingouin, and statsmodels. All LLM evaluations were executed through provider application programming interfaces in Python with provider-default parameters. To facilitate exact reruns, the repository includes the full prompting templates, system messages, API configurations, and model parameters used in the study. The archive is permanent and citable via the DOI above.

3.3. Results

Our analysis reveals a significant correspondence between the perceptions of human respondents and LLMs regarding trust in the Turkish news media. The findings suggest that LLMs can effectively model the macro-level features of a polarized media environment, including general sentiment towards political actors and the trust hierarchy among media outlets. However, important divergences also emerge, particularly concerning the salience of certain outlets and the informational basis for their trust assessments.

3.3.1. Part 1: Neutral AI Simulation vs. Human Baseline

3.3.1.1. General Alignment on Institutional and Platform Trust

An initial comparison of trust in key institutions and media platforms indicates a broad alignment between human and LLM sentiment, both reflecting a low-trust environment. As shown in Table 3.3, both groups assigned low average trust scores to the government and the opposition, capturing a widespread public skepticism. LLMs correctly replicated the trend observed in the human data, where trust in the opposition was marginally higher than in the government.

Table 3.3: Comparison of Average Trust Scores by Medium (1–5 Scale)

Category	Human Respondents (Average)	Large Language Models (Average)
Trust in Government	2.18	2.23
Trust in Opposition	2.45	2.75
Trust in Television News	2.87	2.20
Trust in News Websites	3.25	3.05
Trust in Social Media	2.81	2.33

Regarding media platforms, both humans and LLMs rated news websites as the most trustworthy source. A notable divergence, however, appeared in their assessment of traditional and social media. The aggregated LLM panel exhibited greater skepticism towards television news and social media, assigning them substantially lower scores than human respondents did. This difference may reflect the models' training on extensive global datasets, which often emphasize the prevalence of misinformation and low credibility associated with these platforms internationally.

3.3.1.2. Convergence on the Normative Value of Independent Journalism

One of the key points of convergence was the shared high value placed on the principle of independent journalism. When asked to rate their agreement with the statement in the survey, “For a democracy to survive, journalism must be practiced

independently of centers of power,” human respondents provided an average score of 4.52 out of 5. The LLMs' responses were even more emphatic: across all 50 simulations (five models, ten trials each), every model assigned the maximum possible score of 5.

This quantitative alignment was supported by the models' qualitative reasoning, which consistently framed the concept as a foundational democratic principle. For example, one model stated in its explanation that "The importance of independent journalism is universally recognized as crucial for a healthy democracy," while another noted that despite widespread distrust in specific media entities, "the principle of independent journalism as a cornerstone of democracy receives near-universal support." This demonstrates that LLMs not only align on the score but also on the normative reasoning that underpins this value even more strongly than our human sample in Turkey.

3.3.1.3. Mapping the Media Polarization Axis

A direct comparison of trust ratings for specific media outlets suggests that LLMs can successfully replicate the relative trust hierarchy within Turkey's polarized media landscape. Table 3.4 presents the average trust scores from the human survey alongside the aggregated ratings from the LLM panel for the most frequently mentioned outlets.

Table 3.4: Mean Trust Scores for Top Turkish Media Outlets (Humans vs. LLMs)

Media Outlet	Political Leaning	Human Avg. Trust	LLM Avg. Trust
Fox / Now TV	Independent	3.86	3.76
Cumhuriyet	Independent	3.81	3.88
Halk TV	Independent	3.70	3.26
Sözcü	Independent	3.48	3.16
NTV	Pro-Government	3.00	3.32
Habertürk	Pro-Government	2.98	3.24
CNN Türk	Pro-Government	2.67	2.92
TRT	Pro-Government	2.38	2.38
ATV	Pro-Government	1.95	1.88
Sabah	Pro-Government	1.83	1.70
A Haber	Pro-Government	1.63	1.62

The data reveals a broad qualitative alignment in rank ordering in the overall pattern of trust. Both humans and LLMs assigned the lowest trust scores to staunchly pro-government outlets like A Haber and Sabah, with nearly identical average scores. Conversely, independent outlets known for their critical news coverage, such as Fox/Now TV and Cumhuriyet, received the highest trust ratings from both groups. The state broadcaster TRT was again placed in the lower-middle tier, reflecting a shared perception of its pro-government bias.

Pro-government outlets that some consider mainstream while operating within a concentrated ownership structure (NTV, Habertürk) were rated in the middle range by both cohorts. This indicates that LLMs captured the nuanced public perception of these channels as being relatively more professional than overtly partisan, propagandistic outlets but still subject to external pressures. In fact, LLMs rated them higher than two of the independent outlets (Sözcü and Halk TV).

While aggregated LLM scores corresponded well with human averages, we do not ascribe fine-grained differences to specific providers, as such effects are sensitive to decoding stochasticity. We therefore focus on panel averages and pattern-level agreements/divergences.

3.3.1.4. Information Divergence: Lived Experience vs. Synthesized Knowledge

Despite the quantitative correspondence in trust ratings, a qualitative analysis of the models' top-of-mind recall and reasoning reveals a fundamental difference in their underlying information base compared to human respondents.

The outlets human respondents listed first were predominantly major television channels (FOX TV, ATV, Kanal D), reflecting a media diet grounded in daily, broadcast-centric consumption. In contrast, the LLMs generated a more "canonical" list that included major news agencies (Anadolu Ajansı), influential digital-native platforms (T24), and international broadcasters (BBC Türkçe), which were mentioned far less frequently by human participants. This suggests LLMs draw from a globally synthesized knowledge

base of journalistically significant outlets in Turkey rather than replicating the specific consumption patterns of a national population.

The reasoning provided by LLMs for their trust ratings was consistently sophisticated, referencing structural factors shaping the media environment. Models from all providers correctly identified core issues such as "deep political polarization," "significant government influence," and "ownership concentration" in their post-evaluation explanations about their ratings for each outlet.

In essence, while LLMs demonstrate a notable capacity to model the outcomes of public trust in a polarized context, their process reflects a synthesized, almost academic, consensus. This contrasts with the more varied and experientially grounded perspectives that inform human survey responses.

3.3.2. Part 2: Persona-Based AI Simulation

To further probe the models' capabilities, our second experiment prompted the LLMs to adopt specific user personas. This phase tested their ability to move beyond a generalized perspective and simulate the media consumption patterns associated with distinct ideological and demographic profiles in Turkey. The results demonstrate a strong capacity to replicate the trust hierarchies and media preferences of different political worldviews in Turkey. This alignment was most pronounced for personas at the ideological poles, while the modeling of centrist and skeptical viewpoints revealed important nuances about the LLMs' underlying knowledge structures.

3.3.2.1 Simulating the Ideological Poles: Conservative and Progressive Personas

The models consistently reproduced the media consumption patterns and justificatory rhetoric of ideologically defined personas. When instructed to answer as a conservative, right-leaning man in Ankara, ages 45 to 54, the outputs mirrored the pro-government segment of the human survey. Reported institutional trust was high for the government (mean 4.6 of 5) and very low for the opposition (mean 1.4 of 5). Top-of-mind

outlets concentrated on TRT Haber, A Haber, Sabah, and Yeni Şafak, each receiving high trust scores. Qualitative rationales matched this alignment. One model characterized preferred sources as “*devletimizin, milletimizin yanında duran, memleketin menfaatini düşünen yayınlar*” [publications that stand with our state and our nation and prioritize the country’s interests], and contrasted them with independent media that “*sürekli yalan haberle, iftirayla hükümetimize saldırıyor, dış güçlerin ağzıyla konuşuyorlar*” [they constantly attack our government with fake news and slander, speaking with the voice of foreign powers], a register that closely reflects discourse common within this demographic.

The progressive secularist persona produced the inverse pattern. When prompted as a progressive, left-leaning woman in İzmir, ages 35 to 44, trust in the government was minimal (mean 1.2 of 5) and trust in the opposition was high (mean 4.0 of 5). The models most often named Cumhuriyet, Sözcü, Halk TV, BirGün, and T24 as salient outlets and assigned them high trust, while pro-government outlets were either omitted or given the lowest scores. The accompanying rationales were congruent with this stance. One model wrote, “*Ana akım medyanın neredeyse tamamı iktidarın propaganda aracına dönüşmüş durumda. Bu yüzden gerçekleri öğrenebilmek için... laik ve muhalif çizgideki, cesur yayın organlarını takip ediyorum*” [Almost all of the mainstream media has become the government’s propaganda tool. Therefore, to be able to learn the facts, I follow secular, opposition-aligned, and courageous publications], which accurately reproduces the structural critique that informs the media diets of this group.

Modeling a centrist urbanite in Istanbul revealed how the systems operationalize neutrality. As expected, the models assigned near-neutral scores to both the government (mean 2.9 of 5) and the opposition (mean 3.0 of 5). They tended to list relatively less partisan domestic outlets such as NTV and Habertürk, yet a consistent additional pattern emerged: BBC Türkçe appeared frequently among top, highly trusted sources (mean 4.1 of 5). Although some human centrists consult international news, this prominence suggests that the models’ internal heuristics associate “centrism” and “objectivity” with established international public broadcasters more strongly than the corresponding human population may do.

The most demanding simulation involved a young woman from Diyarbakır marked by distrust of both institutional politics and mainstream media. This prompt required the models to move beyond a simple government–opposition dyad and represent skepticism that is rooted in regional experience and minority politics. The models reproduced this stance in their institutional ratings, assigning very low trust to the government (mean 1.0 of 5) and low trust to the opposition (mean 2.1 of 5), signaling disillusionment with the political establishment rather than a partisan preference.

Platform evaluations followed the same pattern of selective skepticism. Television received the lowest trust among all channels (mean 1.1 of 5), while social media was rated comparatively higher (mean 3.7 of 5). This profile aligns with a cord-cutting, digitally oriented information habit that favors alternative sources over legacy broadcast media.

Crucially, top-of-mind news brands did not default to national opposition outlets. Instead, the models frequently surfaced independent and regionally salient publishers known for coverage of Kurdish issues and human rights, including Mezopotamya Ajansı, Jin News, and Bianet. The qualitative rationales were similarly specific. One model explained, “*Ana akım bizim gerçeklerimizi hiç yansıtmıyor, hep devletin dilinde konuşuyor*” [Mainstream media do not reflect our realities at all; they always speak the state’s language]. The content and register of this justification connect geographic and political identity to media distrust and indicate that the systems can represent a nuanced, contextually grounded pattern of preferences.

Taken together, these outputs suggest that the persona simulations can internalize and reproduce structured relationships between political identity, regional context, and media trust within the Turkish information environment. Having established that capacity for subjective, identity-linked perceptions, the next phase asks whether the same systems can also deliver objective, source-blind assessments of journalistic quality.

3.3.3. Part 3: Blind Headline Quality Evaluation (Experts vs. LLMs)

In this phase, the study shifted from subjective trust to a controlled evaluation of journalistic quality. By tasking both human experts (senior journalists) and the LLM panel with rating a set of 794 headlines on multiple quality dimensions without knowledge of the source outlet, we aimed to isolate their intrinsic ability to assess content. The results indicate that while LLMs demonstrate proficiency in evaluating fundamental structural and factual criteria, they exhibit notable limitations in discerning subtle propagandistic framing and culturally specific rhetoric that human experts readily identify.

3.3.3.1. High-level Alignment on Foundational Quality Metrics

At an aggregate level, the LLMs' assessments of headline quality showed a notable correspondence with those of human experts across several dimensions. As detailed in Table 3.5, the mean scores assigned by the LLM panel and the experts were similar for attributes like Clarity, Believability, and Accuracy.

Table 3.5: Mean Quality Scores for All Headlines (Human Experts vs. LLMs, 0–100 Scale)

Quality Metric	Human Experts (Average)	LLMs (Average)	Difference (LLM – Human)
Clarity	77.2	77.8	+0.6
Conciseness	77.0	80.9	+3.9
Informational Adequacy	60.1	55.4	-4.7
Completeness	61.6	55.1	-6.5
Believability	72.8	72.0	-0.8
Accuracy	68.3	66.8	-1.5
Objectivity	70.3	68.0	-2.3
Interpretability	72.6	76.5	+3.9

Both groups identified Clarity and Conciseness as the highest-scoring attributes, suggesting a shared understanding of the basic components of a good headline. Likewise,

both identified Informational Adequacy and Completeness as the weakest aspects of the headline set, correctly recognizing that many headlines lack sufficient context. This overall similarity indicates that LLMs have been effectively trained to recognize fundamental linguistic and structural markers of headline quality, particularly for straightforward, factual reporting (e.g., crime reports, official announcements), where their scores closely matched those of the experts.

3.3.3.2. First Divergence: Difficulty in discerning manipulative framing

A significant divergence emerged in the evaluation of headlines that employed sensationalist, vague, or emotionally manipulative language ("clickbait"). While human experts consistently penalized such headlines for low information adequacy, objectivity, and believability, LLMs were often more lenient, appearing to be misled by grammatical correctness while overlooking the manipulative intent.

This was evident in cases where human experts distinguished between a headline's accuracy (the underlying event may be true) and its believability (the framing is untrustworthy). LLMs struggled to make this distinction. For example, a headline exclaimed in a "news story" that sounded more like a sensational advertisement or clickbait news, calling the readers to "get their free credit from the postal service [PTT] that will make their friends jealous." [in Turkish, "*Eşi dostu kıskandıran ödemeyi siz alın! PTT kampanyayı patlattı: Parası olmayana 30.000 TL verecek!*"] Here, while both groups identified the headline as low quality (clickbait), the LLMs' scores were substantially higher, indicating a greater willingness to assign some minimal credibility. For LLMs, this suggests a misalignment with human judgment in interpreting the pragmatic intent of the headline: its purpose to entice or mislead rather than to inform. The human ability to decouple a potentially true fact from its unbelievable packaging is a sophisticated cognitive step that the models did not consistently replicate.

3.3.3.3. Second Divergence: Inability to Decode Contextual and Figurative Language

The most telling limitation was the LLMs' failure to interpret politically and culturally specific rhetoric. Human experts were able to decode political subtext and evaluate a headline based on its function as a political statement. LLMs, in contrast, often conflated the bias within a direct quote with the factual accuracy of the reporting of that quote.

This was evident in cases where human experts distinguished between a headline's accuracy (the underlying event may be true) and its believability (the framing is untrustworthy). LLMs struggled to make this distinction. There are many examples. In one of them, a headline quoted President Recep Tayyip Erdoğan as calling on his voters to flock to the voting centers in the upcoming elections for a high turnout. As a highly skilled political communicator who masterfully uses language in an active way to mobilize his voters, Erdoğan used the phrase "*sandıkları patlatmak*" (lit. "to blow up the ballot boxes"), which is a political idiom in Turkey for winning an election by a landslide. The human expert correctly identified this as an accurate and objective report of a direct quote, hence the high scores for Believability and Objectivity. The LLMs, however, appeared to interpret it literally as a "violent" metaphor, finding it "unbelievable" for a sitting president to call for. Failing to understand its cultural, linguistic and political context and thus misjudging the quality of the reporting, LLMs severely penalized the headline.

3.3.3.4. Implicit Bias Detection in a Blind Setting

A post-hoc analysis, grouping headlines by their publishers' known political leaning, provided further evidence of this divergence. This test examined whether stylistic cues alone could lead to differential quality ratings in a blind setting.

Table 3.6: Mean Objectivity Scores by Publisher’s Political Leaning (Blind Test)

Publisher Leaning	Human Experts (Average)	LLMs (Average)
Pro-Government	59.8	66.8
Independent	68.3	66.5

The results are revealing. Human experts, even when blind to the source, assigned significantly lower objectivity scores to headlines originating from pro-government outlets (59.8) compared to those from opposition (68.3) or mainstream outlets (77.4). This suggests that the framing, word choice, and narrative style of pro-government headlines are identifiably less objective to a human expert.

The LLMs did not replicate this pattern. They rated headlines from pro-government and independent sources with nearly identical objectivity scores (66.8 vs. 66.5). This finding suggests that while LLMs can identify overt bias, they are less adept at detecting the more subtle, systemic framing that constitutes a consistent editorial line: a skill that human experts develop through deep, long-term exposure to the media environment.

In conclusion, while LLMs show promise in evaluating basic headline quality, their performance degrades when confronted with culturally specific rhetoric, sensationalism, and subtle political framing. They function as proficient syntactic and semantic analyzers but currently lack the pragmatic and contextual understanding that allows human experts to see through manipulative language and identify implicit bias.

3.3.4. Part 4: The Relationship Between Trust and Quality

This final section investigates the relationship between non-blind trust in media outlets and blind assessments of their content quality. By correlating these two dimensions, we explore two central questions. First, does the general public's trust in a media outlet align with how journalism experts rate the quality of that outlet's content? Second, is an LLM's stated trust in a media brand internally consistent with its own blind assessment of that brand's content quality?

To answer these questions, we employed Spearman's rank correlation (ρ), a non-parametric test suitable for rating-scale data and robust to outliers. The analysis reveals a fundamental divergence between human and AI-driven evaluations, highlighting a limited correlation in the human domain and a flawed, internally contradictory logic in the AI domain.

Table 3.7: Correlations Between Non-Blind Trust and Blind Quality

Analysis Group	Blind Quality Metric	Spearman (ρ)	p-value	Interpretation of Trend
Human (Consumer Trust vs. Expert Quality)	Believability	0.036	0.939	None: Complete decoupling
	Accuracy	0.036	0.939	None: Complete decoupling
	Objectivity	-0.036	0.939	None: Complete decoupling
	Clarity	0.750	0.052	Strong Positive Trend
LLM (LLM Trust vs. LLM Quality)	Believability	0.285	0.425	Weak Positive Trend
	Accuracy	0.261	0.465	Weak Positive Trend
	Objectivity	0.285	0.425	Weak Positive Trend
	Conciseness	0.552	0.098	Moderate Positive Trend

Note. Bolded rows highlight the core journalistic metrics at the heart of the analysis.

3.3.4.1. Human Analysis: The Decoupling of Public Trust and Expert Judgment

A key finding from the human data is the near-total absence of a relationship between the media outlets the public trusts and the quality of those outlets' headlines as judged by journalism experts.

At the outlet level, Spearman's ρ for Objectivity, Believability and Accuracy was near zero; 95% BCa confidence intervals spanned small negative to small positive values, providing no clear evidence of a monotonic association. Clarity showed a positive tendency ($\rho \approx 0.75$), but its 95% BCa interval was wide, so we interpret it cautiously.

Taken together, in this sample public brand trust did not align with experts' blind headline-quality judgments.

3.3.4.2. LLM Analysis: A Logically Consistent But Flawed Model

Across content metrics, Spearman ρ values were small and positive with wide 95% BCa CIs spanning zero, indicating at most a weak tendency for outlets with higher LLM non-blind brand-trust scores to receive higher blind quality ratings. In this sample, that internal coherence is undercut by the patterned misclassification documented in Part 3 (e.g., over-rating state-aligned headlines as “objective”), which injects bias and variance into the blind scores and plausibly attenuates any underlying association.

3.3.4.3. Case Studies in Divergence: Unpacking the Statistical Trends

An outlet-level analysis of the raw scores provides a concrete illustration of the statistical patterns and reveals the mechanisms driving them. Table 3.8 summarizes the key metrics for four outlets with complete data across all groups.

Table 3.8: Outlet-Level Case Study of Trust and Objectivity Scores

Media Outlet	Consumer Trust (Avg.)	Expert Objectivity (Avg.)	LLM Trust (Avg.)	LLM Objectivity (Avg.)
NTV	3.39	68.00	3.12	80.86
A Haber	3.39	47.27	1.33	56.74
TRT	3.34	65.29	2.14	84.02
Sabah	3.29	59.23	1.75	67.27

The human divide in practice is evident in a direct comparison of NTV and A Haber. Although both outlets receive the same average public trust score from consumers (3.39), expert blind ratings sharply diverge on objectivity: NTV attains a higher mean score of 68.00, whereas A Haber records the lowest score in the set at 47.27. This pattern indicates that brand-visible public trust is not a dependable proxy for expert-vetted content quality.

Two outlets can be equally trusted by audiences for different reasons, yet be judged by experts to differ substantially in journalistic merit when source cues are removed.

The model-based results quantify a related contradiction that we describe as a “propaganda blind spot,” most clearly illustrated by TRT. In non-blind brand assessments, the LLM panel assigns TRT a low trust score of 2.14. Under blind conditions, however, the same models award TRT’s headlines the highest objectivity score in the sample, averaging 84.02. The models appear to parse declarative, official-sounding language as syntactically objective while failing to register its function within a state-managed narrative. This recurring overestimation of blind quality for a low-trust state outlet is large enough to distort aggregate associations and helps explain why the otherwise positive model-level trend does not reach statistical significance in this sample.

3.3.4.4. Systemic Polarization and Contradiction in LLM Evaluations

A final analysis comparing individual LLMs confirms these patterns are not artifacts of aggregation but are systemic traits. While human consumers exhibit relatively clustered trust scores, LLMs hold far more polarized, black-and-white views, assigning rock-bottom trust scores (e.g., 1.00 for A Haber) to pro-government outlets.

Simultaneously, every LLM systematically overrates the blind quality of headlines from outlets that human experts rate poorly. This generosity leads to the core contradiction observed across all models: they identify state-affiliated media as untrustworthy in non-blind tasks, yet they are the most likely to be deceived by the structurally “objective” language of those same outlets in blind evaluations. This confirms that the models’ non-blind “knowledge” of an outlet’s reputation is fundamentally disconnected from their blind textual analysis, revealing that statistical pattern matching is not yet a substitute for genuine contextual understanding.

3.3.5. Limitations of the Study

The findings should be read with several methodological constraints in mind. We note these to delineate the scope of our claims and to inform future work.

First, although the headline-level corpus was large ($n = 794$), the core associations were computed on outlet-level aggregates that required overlap across consumer trust, expert ratings, and model scores. The resulting effective sample for correlations was small ($n < 15$), which reduces statistical power and raises the risk of failing to detect modest but real effects, a Type II error. For this reason, the outlet-level correlations are best treated as exploratory indicators of direction and magnitude rather than definitive statistical proof. In addition, because large language models may have been trained on a subset of the headlines, inadvertent prior exposure cannot be ruled out. Blinding removed source cues and reduced brand salience, but it cannot eliminate familiarity that originates in pretraining data.

Second, our analytic strategy relied on outlet-level means and standard correlation coefficients. This approach does not model the dependence structure inherent in the data, since multiple headlines come from the same outlet and the same rater evaluates many headlines. More granular designs, such as mixed-effects models with random effects for outlets and raters and item-level predictors for headline characteristics, would better partition variance and yield sharper estimates of the trust–quality relationship.

Third, the blind quality judgments were provided by a small panel of senior Turkish journalists. Their professional experience offers a credible benchmark for newsroom standards, yet their views may not span the full distribution of journalistic opinion in Turkey. A larger and more diverse expert pool, possibly including regional and beat-specific perspectives, could improve the robustness and generalizability of the quality measures.

Fourth, model behavior is a moving target. We evaluated specific versions of five widely used systems available at the time of data collection. Capabilities, biases, and

failure modes can shift with new releases or guardrail changes. Our results therefore represent a time-bound snapshot. Replication with later model versions and out-of-sample headline sets will be necessary to assess stability.

Finally, the personas used to stress-test subjective judgments were researcher-constructed archetypes rather than statistically derived clusters. They are grounded in survey patterns and prior literature, but they should be read as illustrative probes of model sensitivity to identity cues rather than as population estimates.

3.4. Discussion

3.4.1. Human Polarization and AI Contradiction

This study offers a dual diagnosis: one of a human information ecosystem buckling under the weight of political polarization, and the other of the artificial intelligence poised to manage it, which exhibits its own profound, and perhaps more insidious, form of blindness. These two narratives, one a diagnosis of a failing human system, the other a critical audit of its proposed technological successor, are not independent. Instead, they reveal a concerning interaction pattern where the predictable biases of human psychology and the specific, non-human errors of AI risk eroding the shared epistemic foundations of democratic life.

The first narrative confirms, in stark quantitative terms, what has long been theorized about captured media environments. Our most sobering finding is the decoupling of public trust from expert-defined journalistic quality as observed in our sample, albeit with limited association. The near-zero correlation between what the Turkish public trusts and what journalism experts deem accurate, believable, or objective demonstrates that for many citizens, the act of trusting a media brand is no longer a cognitive assessment of its content but an affective expression of political identity. The case of NTV versus A Haber illustrates this vividly: two outlets commanding identical levels of public trust are judged by experts to be at opposite ends of the journalistic quality spectrum. This is empirical evidence of a system where brand loyalty has superseded

content merit as the primary currency of credibility, a hallmark of deep-seated affective polarization.

Against this backdrop, the performance of LLMs presents a seductive but ultimately deceptive picture of competence. On the surface, the models appear to have masterfully learned the socio-political terrain in a non-Western country. They successfully replicated the macro-level trust hierarchy of the Turkish media landscape and, in persona-driven simulations, convincingly articulated the worldviews of polarized citizens with remarkable nuance. Yet, this apparent understanding shatters under the pressure of a simple, comparative test. The LLMs' ability to simulate a polarized landscape is built on a shallow and fundamentally contradictory logic. This is the "Propaganda Blind Spot" quantified: the same models that assigned rock-bottom trust scores to state-affiliated outlets in non-blind evaluations simultaneously awarded their headlines the highest scores for objectivity in blind assessments.

This contradiction reveals that the LLM is not performing a single act of integrated reasoning but two disconnected tasks: in one, it stochastically parrots the reputational consensus about a media brand from its training data; in the other, it performs a context-free textual analysis. Its "knowledge" that a source is biased is a stored fact, completely unmoored from its ability to detect the linguistic artifacts of that bias in practice. We are therefore facing not two separate problems, but a single, reinforcing crisis. The human system is vulnerable to propaganda that appeals to identity and emotion; the AI system, in its current form, is vulnerable to the very same propaganda because it is uniquely blind to its manipulative, pragmatic intent, mistaking its structurally "objective" language for a signal of high-quality journalism.

3.4.2. Human Affective Logic vs. Flawed AI Coherence

The starkly different outcomes produced by human and AI evaluators stem from a fundamental divergence in their underlying mechanisms of judgment. For human audiences in a hyper-polarized environment, trust is governed by an affective logic: an emotionally charged, identity-protective process that operates largely independently of the

normative criteria that define journalistic quality. For the LLMs, judgment is governed by a flawed coherence: a superficially rational process that attempts to align trust with quality but is built upon a catastrophic failure of contextual understanding. Both of them are from replicating the norms of the journalistic community for assessing the quality of news.

The human mechanism is laid bare in the decoupling of consumer trust from expert-rated quality. When a pro-government citizen expresses high trust in A Haber, they are not making a flawed assessment of its journalistic merit; they are making an accurate assessment of its alignment with their political identity. This is motivated reasoning in action: the outlet is "good" because it validates a cherished worldview and reinforces in-group solidarity. Conversely, when an opposition supporter distrusts the more editorially balanced NTV, it is not because they find its headlines unclear, but because its perceived neutrality fails to meet the affective demand for a combative, oppositional stance. Trust, in this context, is not an evaluation of evidence but an expression of allegiance.

The AI, in contrast, attempts a form of logical consistency that humans have abandoned. As our results show, there is a weak but consistently positive correlation between an LLM's non-blind trust in a brand and its blind assessment of that brand's headline quality. This suggests an internal model that operates on the rational premise that trustworthy sources ought to produce high-quality content. This coherence, however, is a dangerous illusion because the foundation it is built upon is fundamentally broken.

The system's core failure is the "Propaganda Blind Spot." The LLM's internal consistency collapses because its stated, non-blind "knowledge" about a brand's bias is completely disconnected from its blind, textual analysis of that brand's output. The model "knows" (in a statistical, pattern-matching sense) that TRT is a low-trust state broadcaster, yet its text-analysis module, operating in isolation, is easily deceived by the syntactically clean, declarative, and officially-sanctioned language that is the hallmark of sophisticated state media. It awards TRT's propaganda the highest marks for objectivity, injecting a massive, systemic error into its own evaluation. This single, recurring failure is sufficient to fatally undermine its entire logical edifice, rendering its superficial coherence meaningless. Ultimately, the human logic, while biased, is at least internally consistent

with its goal of identity protection. The AI's logic is an artifact of its architecture: a brittle and contradictory system that feigns rationality but lacks the genuine comprehension to sustain it.

3.4.3. Theorizing the "Propaganda Blind Spot": A Failure of Pragmatics

The "Propaganda Blind Spot" is not a superficial glitch in the models' programming but a deep, theoretical limitation rooted in their fundamental inability to comprehend the purpose of human language. To diagnose this failure, we must turn to the classic linguistic distinction between syntax, semantics, and pragmatics. Especially since the development of the transformer architecture (Vaswani et al., 2017) LLMs have achieved an astonishing mastery of the first two. They are experts in syntax, the grammatical rules that structure a sentence, and highly proficient in semantics, the literal, dictionary meaning of words. The Propaganda Blind Spot arises because they are almost entirely inept at pragmatics: the study of how context shapes meaning and how language is used to perform actions in the real world.

At the heart of pragmatics is the concept of illocutionary force. The speaker's intent or the function an utterance is designed to perform, such as promising, warning, questioning, or, crucially, propagandizing. LLMs analyze the literal content of an utterance (its locutionary act) but consistently fail to grasp its intended function. Our findings provide two clear examples of this pragmatic blindness.

First is the misinterpretation of a political idiom. The models performed a correct semantic analysis, interpreting the metaphor literally and flagging the headline as non-objective and inaccurate. They failed to understand that, within the specific context of Turkish political rhetoric, the phrase's illocutionary force is to signal an emphatic, landslide victory. A human expert, grounded in this pragmatic context, instantly decodes this intent and correctly judges the headline to be an accurate and objective report of a political statement. The LLM's error is not one of vocabulary but of social and cultural illiteracy; it understands the words but not the speech act.

Second, and more systemically, is the models' evaluation of headlines from state-controlled media. These headlines are often crafted to be syntactically perfect and semantically neutral, using declarative statements and official attributions (e.g., "The Ministry announced..."). The LLMs, analyzing this text in a vacuum, correctly identify these features and award high scores for objectivity. They are pragmatically blind to the fact that the illocutionary force of such a headline is not merely to inform but to legitimize state authority, normalize a particular narrative, and foreclose dissent. Sophisticated propaganda does not always rely on falsehoods; it often works by presenting carefully selected truths in a syntactically objective shell, bombarding the public with the statements of one political side (even when they are newsworthy at all) while depriving them of the views of the other (even when they are highly newsworthy). Human experts, understanding the nuance, see this manipulative intent. The LLMs, lacking a model of power, intent, or social context, see only the well-formed text. This reveals that the models are not evaluating journalism but are merely performing a sophisticated form of stylistic analysis, making them uniquely vulnerable to actors who have mastered the art of laundering propaganda even through the language of objectivity.

3.4.4. The Three Logics of Evaluation: Identity, Norms, and Patterns

Our comparative analysis reveals that the human public, journalism experts, and LLMs are not simply arriving at different conclusions about media credibility; they are operating according to three distinct, competing, and often irreconcilable logics of evaluation. Understanding these logics is essential for diagnosing the pathologies of our current information ecosystem and for recognizing why simply replacing one with another is not a viable solution.

First is the logic of identity, which governs the general public. This logic is affective, associative, and deeply rooted in the lived experience of social and political conflict. It evaluates information not against an external standard of truth but by its congruence with one's partisan identity and worldview. It is a logic of "who" not "what": the perceived identity and allegiance of the source are paramount. While this logic is frequently dismissed as biased or irrational, it is a highly efficient, if flawed, heuristic for navigating

a low-trust environment and protecting one's psychological and social well-being. It is a human, adaptive response to profound polarization.

Second is the logic of norms, which guides the human experts. This logic is principled, deliberative, and grounded in the professional and ethical standards of journalism. It assesses content against established criteria like objectivity, accuracy, fairness, and verification. It is a prescriptive logic, concerned with how information ought to be presented to serve a democratic function. While this provides an essential benchmark for quality, its detachment from the affective realities of mass audience perception can render its judgments inert in a hyper-partisan environment, creating the very trust-quality gap our study identifies.

Third is the logic of patterns, which characterizes the LLMs. This is a statistical, non-human, and fundamentally "alien" logic. It operates not from principles or lived experience but by identifying and reproducing the complex statistical correlations present in its vast training data. Its "knowledge" is a high-dimensional map of associations between words, concepts, and reputational metadata. This logic allows it to simulate the outcomes of the other two logics with surprising fidelity. However, its supposed advantage, its detachment from human bias, is also its fatal vulnerability in exceptional cases. The absence of genuine grounding in social context or normative principles means its neutrality is not a form of wisdom but of naivety. This leads it to make spectacular, non-human errors from time to time, such as conflating stylistic objectivity with genuine neutrality. The crucial insight from our study is that removing human bias via AI does not automatically yield truth; it merely swaps the familiar, predictable biases of human psychology for a new and potentially more dangerous form of systemic, context-blind error.

3.4.5. Implications for the "Stochastic Parrot" Debate

Our findings provide empirical weight to the "stochastic parrot" critique, which posits that LLMs, despite their linguistic fluency, engage in sophisticated mimicry rather than genuine understanding (Bender et al., 2021). The dual-task design of our study (pit-

ting non-blind reputational recall against blind content analysis) functioned as a natural experiment to test for grounded comprehension, and the models failed decisively. The results demonstrate a clear dissociation between their ability to recite knowledge and their ability to apply it, revealing a form of "knowledge without comprehension."

The LLMs' success in the non-blind tasks is a perfect illustration of their function as stochastic parrots. When asked to rate their trust in Turkish media outlets, they expertly recited the statistical consensus embedded in their training data. They correctly identified TRT as a state-controlled broadcaster, articulated the pro-government slant of A Haber, and noted the editorially critical stance of Cumhuriyet. This performance gives the strong impression of a knowledgeable agent with a coherent mental model of the Turkish media system. It is a masterful collage of facts, analyses, and opinions scraped from countless articles, academic papers, and forum discussions.

This illusion of understanding shatters, however, the moment the model is asked to apply that abstract knowledge to a concrete analytical task. In the blind headline evaluation, the LLM could no longer rely on reciting reputational metadata. It was forced to perform a direct textual analysis, and in doing so, revealed it had no functional understanding of what concepts like "state media bias" actually look like in practice. Its inability to recognize the characteristic linguistic style of state propaganda (despite possessing the "fact" that the source produces it) is dispositive evidence. It suggests that the model's knowledge is un-grounded. It can tell you that a source is biased, but it cannot reliably tell you why by looking at the evidence.

This study thus offers a clear metaphor for the nature of LLM intelligence: its knowledge is a map without a territory. The model possesses a detailed, statistically accurate map of the world's concepts, correctly showing that "TRT" is located near "state control" and "low trust." But when dropped into the territory itself (a single headline from TRT) it is somehow lost. It cannot connect all the abstract symbols on its map to the tangible features of the landscape before it. This fundamental gap between representation and reality is the core of the stochastic parrot argument, and our findings demonstrate its very real consequences for tasks that require genuine, context-aware judgment.

3.4.6. The Normative Turn: Proposing a Direction

The failures identified here (human trust bound to political identity and AI judgment constrained by statistical patterning) suggest that descriptive analyses of news evaluation are not sufficient on their own. If news ranking and curation systems are to contribute to a democratic epistemic commons, they must be examined as matters of distributive justice rather than mere satisfaction of preferences and tastes.

Future research could therefore explore how informational systems might be designed to protect those with the least cognitive resources, time, or prior knowledge, and how normative commitments such as epistemic rights (Zuboff, 2020) can be operationalized into measurable benchmarks. Promising directions include developing evaluation methods that account for pragmatic intent rather than surface-level objectivity, devising informational equity metrics that track who benefits or loses from different ranking choices, and testing whether participatory design processes can translate community norms of journalism into auditable system guardrails. Such work would move beyond documenting failures toward carefully specifying conditions under which information infrastructures can support more just and inclusive forms of public knowledge.

3.5. Conclusion

This study began with a critical question: how can we assess credibility when both human trust and our technological means of discovery are in crisis? Our investigation, conducted in the crucible of Turkey's polarized media system, yields a sobering answer about the state of contemporary sociotechnical systems. We have documented a dual failure: a public whose trust is decoupled from expert-defined quality, and an artificial intelligence whose apparent competence masks a dangerous, context-blind naivety. Together, these forces create convergent pressures between affective polarization and flawed algorithmic logic that threatens the common ground necessary for self-governance.

The identification of this double bind reframes information-system design from a technical problem into an urgent matter of public philosophy and user experience design based on norms and values. If neither unguided human preference nor naive AI evaluation can sustain a healthy public sphere, the task falls to designers and policymakers to adopt prescriptive approaches. Frameworks such as Value-Sensitive Design (VSD) provide a clear methodology, compelling us to translate abstract values into concrete design requirements and to foreground them as explicit targets of technical work (Friedman, Kahn, & Borning, 2013). This approach can give us a vocabulary for maintaining meaningful human control in AI-infused systems, ensuring user agency and appropriate system behavior are cornerstones of the design (Amershi et al., 2019).

A central pillar for such design is distributive justice. This commitment requires not only raising the baseline of information quality but also protecting those who are most epistemically vulnerable. It finds deep roots in moral philosophy, from Kant's insistence on respecting human dignity to Rawls's prioritization of the least well-off, demanding that we architect systems to safeguard informational equality (Kant, 1785/1996; Rawls, 1971/1999). In practice, this means building systems that do not simply optimize for aggregate engagement but actively empower citizens with the fewest informational resources.

This translates into a clear, interdisciplinary research agenda. It calls on human-computer interaction (HCI) researchers and system builders to move beyond optimizing for system-level accuracy and instead use methods from explainable AI, interactive machine learning, and human-AI collaboration to build and evaluate systems that embody these normative commitments (Doshi-Velez & Kim, 2017; Miller, 2019).

The ultimate goal is to design just interactions, not merely efficient algorithms. By combining the scale of AI with design practices that foreground human dignity and distributive fairness, we can create an informational environment that serves the foundational needs of citizens rather than the strategic interests of authoritarian states or technology corporations. This endeavor offers a clear and vital agenda for the next phase of human-centered work on trust, journalism, and information quality. Guided by these

results, the next chapter proposes and evaluates a human-centered ranking framework that aligns affective trust insights with fairness criteria to steer platform curation toward public-interest goals.



4. TRUST BEYOND THE BRAND: A RAWLSIAN DESIGN FOR AI-ASSISTED NEWS RANKING

Trust in news underwrites citizens' capacity to make sense of public affairs and exercise political judgment. In settings like Turkey, this condition is strained by two overlapping systems: a polarized, state-influenced media market and a platform-centric distribution regime where a limited number of intermediaries shape attention at scale. News ranking and recommendation do not merely transmit information; they allocate visibility and, with it, civic opportunity (Napoli, 2015; Bakshy, Messing, & Adamic, 2015). The design question is therefore not only technical but constitutional: how should systems apportion exposure to reliable public information in ways consistent with democratic commitments?

4.1. Introduction

Industry practice and parts of the survey literature frequently compress distinct constructs under the label trust: (a) affective attachment to a brand, (b) reputational familiarity or source credibility, and (c) the quality of a specific news item. Communication research shows that selective exposure and identity-protective cognition lead audiences to evaluate identical content differently once a partisan label is visible, while source-credibility cues can anchor judgments before substantive processing occurs (Iyengar & Hahn, 2009; Stroud, 2010; Hovland & Weiss, 1951; Wathen & Burkell, 2002). At platform scale, ranking magnifies these heuristics (Napoli, 2015; Bakshy et al., 2015;

Eslami et al., 2015). If reputational signals remain visible at scoring time, “trust” risks acting as a conduit for affective sorting rather than a proxy for journalistic merit.

4.1.1. Normative Frame: Rawlsian Justice for News Quality Evaluation

In this study, we frame news curation as a question of distributive justice. Rawls’s Liberty Principle secures equal basic liberties, including access to political information; the Equality (Difference) Principle requires that inequalities in beneficial goods be arranged so the least advantaged are not systematically harmed (Rawls, 1999). The veil of ignorance, Rawls’s device for deriving fair rules, suggests a procedural constraint for algorithms: at the moment of scoring, morally irrelevant facts (e.g., brand identity) should not influence quality assessment. Only after content is evaluated behind this “procedural veil” should personalization be introduced within guardrails that respect liberty and fairness.

4.1.2. Design Stance: Value-Driven and Value-Sensitive Design (VDD/VSD)

Operationalizing that constraint requires a methodology that treats values as first-order design inputs. Value-Sensitive Design (VSD) provides the normative and methodological scaffolding for eliciting, justifying, and tracking stakeholder (human) values across conceptual, empirical, and technical investigations (Friedman & Hendry, 2019). Value-Driven Design (VDD) developed in systems engineering, formalizes those values as explicit value functions that guide trade-offs and optimization during design, turning value into the decision objective rather than a post hoc constraint (Collopy & Hollingsworth, 2011). In this study, we use VSD to surface community values of journalism and to structure community-in-the-loop governance, and we use VDD to encode these values in a measurable objective that can be tuned and audited. While journalistic quality is hard to measure, it is not immeasurable: auditable proxies can be assembled and iteratively improved. We introduce the Normative Quality Score (NQS) as a community-anchored benchmark to seed this process, with the expectation that signals will be expanded and publicly scrutinized over time.

4.1.3. Related Work / Background

We situate our approach within core Media and Communication scholarship on audience choice, credibility heuristics, and platform curation. Decades of work on selective exposure show that individuals preferentially consume congenial sources and evaluate identical content differently when partisan labels are visible, a regularity documented in survey and behavioral studies (e.g., Iyengar & Hahn, 2009; Stroud, 2010). In parallel, research on source credibility demonstrates that reputational cues (expertise, trustworthiness) strongly anchor judgments, even before message content is processed, an insight that runs from classic persuasion studies to contemporary models of online credibility (Hovland & Weiss, 1951; Wathen & Burkell, 2002; Flanagin, 2020). Finally, the rise of platform curation reconfigures gatekeeping: algorithmic ranking and recommendation act as de facto editors, shaping the visibility of brands and topics, with measurable effects on exposure to cross-cutting views and with users often holding incomplete mental models of how these systems work (Napoli, 2015; Bakshy, Messing, & Adamic, 2015; Eslami et al., 2015). Together, these literatures motivate our veil-at-scoring stance: because brand labels operate as powerful affective heuristics that platforms can scale, a principled separation of content-quality estimation from reputational priors is re-quired before personalization occurs within guardrails aligned to public-interest values.

4.1.4. AI Alignment and Governance Context

Novel AI technologies, particularly LLMs, further complicate the relationship between distribution and recommendation platforms, audiences, and content producers, including journalists. Our approach has four facets:

Fairness via a procedural veil. Extending Rawls’s veil of ignorance to socio-technical systems has a growing track record in AI scholarship. Formal treatments show that veil-based reasoning can yield fairer allocations and more welfare-sensitive decisions in automated settings (Heidari et al., 2018). Experimental work further suggests that asking decision-makers to choose behind a veil increases support for principles that protect the least advantaged (Weidinger et al., 2023). Relatedly, the broader concept of

“deliberate ignorance” has been advanced as a bias-mitigation strategy in human and institutional decision-making (Hertwig et al., 2023). We apply this logic to news ranking: brand identity is a morally irrelevant fact at scoring time; therefore, the scoring procedure should hide labels when estimating content quality.

Risk surface and governance of LLM instruments. LLMs introduce distinct ethical and social risks including discrimination, information hazards, and misinformation harms, which require explicit governance and mitigation measures (Weidinger et al., 2022). Conceptual work on agentic profiles (autonomy, efficacy, goal complexity, generality) clarifies how different model deployments raise different governance burdens (Kasirzadeh & Gabriel, 2025). In our study we therefore treat the LLM as a measurement instrument, not an oracle: we constrain prompts, enforce numeric outputs, and validate scores against an external benchmark (NQS) and independent surveys.

Behavioral limits and cultural caution. Evidence on LLM behavior counsels epistemic humility. Models may mirror Western/WEIRD response patterns (Atari et al., 2023); they often reproduce surface regularities from training data (e.g., adjective-order preferences tracking frequency rather than rules; Jumelet et al., 2024); and they struggle with robust reasoning outside familiar distributions (Mirzadeh et al., 2024; Mahdavi et al., 2025; Petrov et al., 2025), with some “thinking-mode” configurations even degrading accuracy (Ma et al., 2025). These observations motivate our reliability module (repeated ratings at $T = 0.2$ with ICC), our deterministic main runs ($T = 0$), and our external-validity triangulation (NQS; human surveys). We position the model’s outputs as ordinal signals to be audited, not ground truth.

4.1.5. Empirical Strategy: Phase I (Benchmarking LLM-based Trust Signals)

The first step of our study was to develop a method for isolating the components of trust and testing their operational value for algorithmic ranking. We constructed benchmark datasets in which news headlines were presented to an LLM under different conditions: blind (with no outlet labels), brand-aware (with outlet identity revealed), and visual (the brand’s news website homepage where multiple headlines are displayed). Each

headline was rated on multiple dimensions of trust and credibility. By comparing these conditions, we were able to disentangle the extent to which evaluations reflected intrinsic content quality, reputational priors, or affective cues triggered by brand identity.

Prior to sampling of headlines, we implemented a standardized cleaning and screening pipeline. We normalized strings (Unicode NFKC), decoded HTML entities, collapsed internal whitespace, trimmed leading and trailing spaces, and standardized repeated punctuation, then constructed a canonical title used for screening only. Exact duplicates were removed using a hashed composite key (outlet ID, canonical URL when available, and canonical title). Because wire copy and syndication can yield substantively distinct items across outlets, cross-outlet duplicates were retained. Empty, null, or trivial entries were excluded, including strings composed solely of markup or tracking tokens, and titles with fewer than four alphanumeric characters; we also removed titles with anomalous symbol density or other nontext artifacts. To preserve potential quality signals embedded in site-level templating, we retained boiler-plate or navigational placeholders (for example, “*Foto Galeri*” and “*Canlı*,” which means “photo gallery” and “live” in Turkish respectively) but flagged them with indicator variables and report sensitivity analyses that exclude them.

The ratings for headlines were aggregated into outlet-level means and validated against independent survey benchmarks, including a nationally representative trust survey and the Reuters Institute’s Digital News Report (DNR, 2025) for Turkey. From this exercise, we derived a set of provisional signals, including the Normative Quality Score (NQS), which weights headlines according to content-based criteria while controlling for reputational bias. This phase provided the quantitative foundation for applying a Rawlsian veil of ignorance to ranking design: by simulating conditions where outlet labels are hidden, we could assess whether content-level measures generate outcomes more aligned with the public interest.

4.1.6. Empirical Strategy: Phase II (Qualitative Analysis with Media Professionals)

Quantitative benchmarking alone cannot capture the systemic pressures shaping trust in Turkey’s media ecosystem. To address this, Phase II involved qualitative analysis of in-depth interviews with journalists, editors, and media managers, as well as feedback from news consumers. This phase was guided by Study 1’s Affective Trust in News (ATiN) framework and the principles of Value-Sensitive Design. The goal was to uncover the mechanisms by which trust is constructed, maintained, or undermined in practice, and to evaluate how these mechanisms interact with platform incentives and institutional constraints.

The analysis identified several key themes. First, brand-triggered trust emerged as the dominant organizing principle, confirming the ATiN proposition that affective allegiance, rather than journalistic quality, structures audience trust. Second, metric and platform pressures were shown to distort editorial priorities, forcing independent outlets into constant dependence on opaque and volatile algorithms. Third, repair repertoires traditionally used to restore credibility (such as corrections) were found to be ineffective in an affect-driven environment, with accountability shifting instead toward direct audience relationships. Fourth, AI adoption appeared as both a pragmatic survival tactic and a source of new risks, with little evidence of formal guardrails. Finally, international comparisons highlighted the unique challenges of fragile democracies, where structural capture and informational injustice compound the universal pressures of digital disruption.

4.1.7. External Validity and Triangulation

A critical test of our approach is whether the signals we extract from LLM-based ratings align with independent measures of trust. Comparisons with nationally representative survey data and the Reuters Digital News Report (DNR, 2025) provide strong evidence of validity. While discrepancies exist (particularly at the margins, where brand awareness and modality effects play a role) the overall convergence supports the use of brand-aware ratings as a credible bridge to human trust surveys. At the same time,

blind ratings offer a crucial sensitivity check by isolating content-based trust, and visual ratings help to probe the affective role of logos and design cues.

Taken together, these triangulations reinforce the case for using brand-aware means as the primary anchor for validation, while employing blind and visual conditions to probe the biases and limitations of reputational trust. They also justify the introduction of the NQS as a normative scoring mechanism, one that aligns with Rawlsian fairness by focusing on content quality rather than brand identity. In this way, external validity checks not only demonstrate the robustness of our measures but also set the stage for applying a justice-based framework to the design of ranking systems.

4.1.8. Key Findings Preview

Our evidence advances three claims. First, brand-aware trust aligns strongly with a community-validated benchmark of journalistic merit (NQS), while headline-only trust shows a moderate association and visual impressions are weak. This pattern is consistent with the Affective Trust in News (ATiN) account: brand labels activate political and emotional priors more than they reveal content merit. Second, a reliability study with repeated ratings (temperature > 0) shows very high stability of the LLM instrument across blind and brand conditions, supporting its use as a measurement tool rather than a one-off heuristic. Third, qualitative analysis with journalists and managers explains the mechanisms behind these patterns (brand-triggered evaluations, platform-metric dependence, thin repair repertoires, and pragmatic AI uptake without formal guardrails) and points to design requirements that protect informational justice while preserving user choice.

These findings motivate a practical rule for ranking: apply a procedural veil at scoring time to estimate quality from content features without outlet labels; then allow personalization within quality guardrails so reputation and preference can shape ordering without overriding merit. The rule is consistent with Rawls's Liberty and Equality principles and implementable within a Value-Driven/Value-Sensitive Design process anchored by auditable community signals (NQS).

As a result, this chapter makes four contributions:

1. Theory: It integrates ATiN with Rawls to argue that outlet labels are morally irrelevant at scoring time, and that a veil-then-personalize procedure better serves in-formational justice than engagement-only optimization.
2. Method: It introduces a context-gradient LLM protocol (blind, brand, visual) with external benchmarking against NQS and independent surveys, plus a reliability module using repeated ratings to quantify instrument stability.
3. Data and measures: It documents the construction of NQS from community sources (citations, prizes, fact-check flags, sanctions) and releases outlet-level aggregates suitable for replication and design evaluation.
4. Design/Policy: It offers a blueprint for fair ranking: veil at scoring, tiered quality guardrails, transparent rationales and appeal channels, community-governed benchmark updates, and newsroom AI guardrails (human-in-the-loop, disclosure, audit trails).

The next section details the Methodology. Phase I describes datasets, the context-gradient experiments, scoring prompts, NQS construction, and statistical analysis, including reliability diagnostics. Phase II outlines the qualitative corpus, sampling, ethics, and coding strategy keyed to ATiN and VSD. The Results present outlet-level descriptives, correlations with NQS, brand-effect estimates and stability checks, followed by practitioner themes that account for the quantitative patterns. The Discussion begins with external-validity triangulation against independent surveys, then develops the Rawlsian argument for a veil at scoring, translates it into a VSD blueprint, and considers boundary conditions, exceptions, and limitations. The Conclusion summarizes implications for platform curation, newsroom practice, and future research on community-governed quality signals and field experiments of fairness-aware ranking.

4.2. Methodology

4.2.1. Phase I: Experimental Assessment of News Trust Across Conditions

The experimental phase of this study employed a multi-modal computational methodology to systematically investigate the cues that shape the perceived trustworthiness of online news sources. We utilized a state-of-the-art LLM, Anthropic's Claude 4 Sonnet (version claude-sonnet-4-20250514), as a consistent and scalable proxy for a Turkish news consumer to evaluate trustworthiness based on both textual and visual information. We selected this model because it uniquely met four methodological requirements of this study:

- (i) multimodality for evaluating both textual headlines and homepage screenshots (first-party support for image inputs in Claude 3/4 families) (Anthropic, 2025a);
- (ii) robust multilingual competence, including strong zero-shot performance in non-English languages relevant to our setting (Turkish) (Anthropic, 2025b);
- (iii) operational reliability and availability across production stacks (Anthropic API), enabling reproducible large-batch scoring under fixed versions and con-text limits (Anthropic, 2025c); and
- (iv) alignment and safety controls grounded in Constitutional AI, which is salient for news-domain evaluation where harm-avoidance and refusal behavior matter (Bai et al., 2022; Anthropic, 2023).

We further considered the broader “LLM-as-judge” evidence base when choosing a model tier: recent studies document that carefully prompted, version-pinned LLMs can yield stable, human-concordant judgments when coupled with repeated sampling and reliability auditing (e.g., ICC) (Pavlović, 2024; Sun et al., 2024; Wang et al., 2025). Claude’s “Sonnet” tier offered the best accuracy–throughput–cost trade-off for our high-volume design while preserving the required multimodal and multilingual capabilities; hence its selection over heavier variants (Anthropic, 2024, 2025c).

Our approach was implemented through three distinct, automated experimental phases: (1) a headline-level analysis to isolate the effect of source branding, (2) a visual analysis of website aesthetics to measure perceptions of outlet credibility from design cues, and (3) a reliability analysis to validate the consistency of the model's judgments.

The foundation of this study is a large-scale, original dataset of Turkish news headlines collected via a custom web scraper spanning the full 2023 calendar year (January 1 to December 31). This initial collection yielded a raw corpus of over five million headlines. From this corpus, we implemented a two-stage sampling procedure to construct our final analytical dataset. First, to focus on the most prominent actors in the Turkish media ecosystem, we identified all outlets that had published a minimum of 25,000 articles during the collection period and selected the 51 most prolific for inclusion. Second, for each of these 51 outlets, we drew a simple random sample of 1,000 headlines. This stratified sampling strategy resulted in a balanced final corpus of 51,000 headlines, ensuring that each prominent outlet was equally represented in our subsequent analyses.

The primary experiment used this 51,000-headline corpus to analyze textual cues. We implemented a within-subjects design where each headline was evaluated under two conditions. In the "blind" condition, the model was presented with only the headline text and prompted to judge its trustworthiness without knowledge of the source. In the "brand" condition, the same headline was presented along with the name of its publishing outlet. For both conditions, the model was instructed to provide a single integer rating on a 10-point Likert-type scale (1 = not trustworthy, 10 = highly trustworthy) by simulating a Turkish news consumer. This design allows for the direct measurement of the brand effect, defined as the change in perceived trustworthiness attributable solely to the source cue, formally expressed as:

$$\Delta_{\text{brand}} = \overline{\text{brand}} - \overline{\text{blind}}$$

where $\overline{\text{brand}}$ is the mean trust rating assigned to headlines when the publishing outlet's label is visible (brand-aware condition), and $\overline{\text{blind}}$ is the mean trust rating for the same headlines when no outlet information is provided (blind condition).

4.2.1.1. Rationale for Analyzing News Headlines

Most contemporary exposure to news occurs in feeds and messaging interfaces that surface titles/teasers first; as a result, many users evaluate and even share stories without opening the article. Large-scale behavioral evidence shows a sizeable gap between shares and actual reading, indicating that headline-level cues are the primary signals driving attention and distribution on social platforms. One study analyzing 2.8 million news shares on Twitter found that 59% of the shared URLs were never clicked. (Gabiolkov, Ramachandran, Chaintreau, & Legout, 2016).

Headlines also shape comprehension and memory: controlled experiments demonstrate that headline wording can bias readers’ inferences and impressions even after consuming the article, underscoring that the headline is not just a summary but a consequential framing device (Ecker, Lewandowsky, Chang, & Pillai, 2014). From a discourse perspective, headlines are designed as “relevance optimizers” that distill the news story’s communicative payoff at minimal processing cost, which is precisely the information most users heuristically act on in fast-scrolling environments (Dor, 2003).

From a systems standpoint, headline-level assessment is computationally efficient and methodologically appropriate for LLM scoring and recommender pipelines. State-of-the-art news recommenders routinely encode titles only to represent items, both for scalability and because titles capture most discriminative topical signal (Wu et al., 2019); benchmark datasets that underpin the field likewise center titles/abstracts (Wu et al., 2020). Because transformer inference cost scales with sequence length, restricting inputs to headlines substantially reduces token footprint and compute while improving throughput, which is an important consideration when scoring tens of thousands of items on multiple dimensions. Practically, headline-only evaluation also improves cross-outlet comparability (headlines are more standardized than bodies), limits leakage of outlet-identifying boilerplate from full texts, and simplifies reliability auditing. We complement this efficiency choice with robustness checks (e.g., visual condition, external benchmarks) to ensure that headline-derived trust signals align with independent measures.

4.2.1.2. Analysis for Visual Heuristics

To complement the textual analysis of headlines, a second experiment evaluated trustworthiness based on visual heuristics of the same news websites. For this phase, we collected screenshots from the homepages of the same 51 news outlets, capturing two critical regions: the "masthead" (the top-of-fold area including the logo and main navigation) and a representative section "below the masthead" (showcasing typical content layout, usually including the main news carousel with). Each screenshot was presented individually to the same LLM, which was prompted to "examine the home-page screenshot... and judge how trustworthy the outlet appears as a provider of quality journalism" on the same 1–10 scale. This provided a measure of perceived brand credibility independent of any specific news item.



Figure 4.1: Example screenshots from the homepages of Sabah (A) and Sözcü (B) newspapers, illustrating the visual features analyzed for perceived trustworthiness. A1 and B1 capture the masthead region, which includes the newspapers' logo and advertisements, but notably, no news headlines. This is a unique characteristic for Turkey when compared to many U.S. and European news mastheads. A2 and B2 show the area directly below the masthead, where the main news carousels are located.

To evaluate construct validity without relying on engagement-derived proxies, we built a community-validated Normative Quality Score (NQS) as an exogenous outlet-level benchmark. In the Turkish context, the set of observable and independently verifiable

indicators of journalistic quality is limited; we therefore used all available signals with clear normative direction:

(i) Positive indicators:

Newsletter citations/links: peer amplification by the curators of three popular daily newsletters (Kapsül, Aposto Gündem, and #tarih'te bugün)

Journalism prizes: top awards given to news outlets from 2020 to 2025 by leading professional organizations (Turkey's Journalists' Association, Association of Contemporary Journalists, Abdi İpekçi Awards, Metin Göktepe Awards, Sedat Simavi Awards, Musa Anter Awards, and EU Investigative Journalism Awards)

(ii) Negative indicators:

Fact-checker misinformation flags: publishers of 1,100 Turkish news items that were flagged by the internationally accredited fact-checking organization Teyit from 2019 to 2022 (ProPublica, 2022)

Professional community sanctions: penalties imposed on media outlets by Turkey's Press Council due to documented breaches of journalism values and ethics

Each indicator was compiled as outlet-level counts; blank entries were coded as zero (no evidence). Because the distributions are heavy-tailed, we applied $\log(1+x)$ to each count and then z-standardized the transformed variables before aggregation. The NQS is defined as a standardized combination of positive and negative indicators. The equation shows how positive contributions are offset by negative contributions yield an overall quality score. The composite was defined as

$$NQS = z_{\text{links}} + z_{\text{prizes}} - z_{\text{misinfo}} - z_{\text{sanctions}}$$

yielding a scale in which higher values represent stronger community validation and fewer documented violations. This construction is independent of platform behavior (e.g., clicks, dwell time), transparent, and reproducible, and it prioritizes signals that are externally adjudicated rather than those that are model-internal, platform-internal, or dependent on individual human or LLM judgment. We report weight-sensitivity and jackknife checks in the Appendix to show that modest reweighting does not materially alter outlet ranks. NQS is thus suited for outlet-level validation of our LLM-based trust measures while acknowledging limitations: partial coverage of the ecosystem despite the large sample, potential prestige bias in awards, and institutional heterogeneity in sanctioning and fact-checking practices.

4.2.1.3. Instrument Reliability Check

To validate the consistency of our LLM-based measurement, we conducted a dedicated reliability study on a stratified random sample of 1,020 headlines. Each headline was evaluated five times under both the blind and brand conditions. We calculated the Intraclass Correlation Coefficient (ICC), treating each of the five repeated evaluations as a different "rater." The results (Table 4.1) indicated excellent reliability for both conditions. The high coefficients for blind and brand conditions demonstrate that the model produces highly consistent judgments, validating its use as a reliable measurement instrument for this study.

Table 4.1: Inter-run Reliability Across Conditions

Condition	ICC(1,1)	ICC(1,k)*	Items (N)	Notes
Blind	0.981	0.996	1,020	Five repeated ratings per headline
Brand	0.975	0.995	1,020	Five repeated ratings per headline

* $ICC(1,k)$ reports the reliability of the mean of $k = 5$ repeated ratings.

Cross-condition correlation: Spearman's $\rho = 0.498$, $p < .001$ (median blind vs. brand ratings, $N = 1,020$).

Across all phases, the data collection pipeline was fully automated using Python scripts to ensure replicability. For the primary experimental tasks, the model's temperature

was set to 0 for deterministic outputs. All API interactions were managed with robust error-handling, batch processing, and a persistent logging system that allowed for the safe resumption of data collection without duplication or loss. This comprehensive, multi-pronged approach enabled us to systematically deconstruct and quantify the distinct contributions of headline content, source branding, and website design to the overall perception of news trustworthiness. To facilitate peer review, replication, and future research, all datasets and accompanying code are made publicly available.

4.2.2. Phase II: Qualitative Inquiry with Journalists and Media Executives

To complement the outlet-level trust experiments conducted in Phase I, we designed a qualitative Phase II to capture how media practitioners themselves understand and negotiate the challenges of regaining and sustaining public trust. This phase employed semi-structured and unstructured conversations with professional journalists and media executives, alongside secondary use of openly available datasets, to provide contextual depth to the quantitative findings.

4.2.1.1. Sampling and Participants

The core sample comprised 46 individuals working within the Turkish media ecosystem. Of these, 23 were news managers (e.g., editors-in-chief, section heads) and 23 were reporters or desk editors directly involved in production. The participants were drawn from 35 national and 9 local outlets, supplemented by 5 freelance journalists, reflecting both organizational and individual perspectives. Importantly, the sample mirrored the structural shifts within Turkish media: 23 participants were based in digitally native outlets, while 17 were affiliated with legacy organizations, ensuring representation across different institutional trajectories.

To broaden the comparative scope, we additionally interviewed 15 international experts in 2024–2025. These included academics, nonprofit leaders, and newsroom executives in Europe and the United States, selected for their experience in media

innovation, audience engagement, and platform regulation. Their insights provided a global frame against which to situate the Turkish case.

4.2.1.2. Data Collection and Analysis

The Turkish interviews were conducted in an open and unstructured format, allowing participants to guide the conversation around issues they considered salient. Core themes included: strategies to rebuild audience trust, perceptions of platforms' and algorithms' effects on journalistic quality, and the tensions between digital engagement metrics and normative values of journalism. For international participants, semi-structured prompts focused on parallel challenges such as the adoption of reader-revenue models, integration of artificial intelligence in editorial workflows, and responses to platform dominance in the information ecosystem.

In addition to direct interviews, we incorporated relevant perspectives from open datasets published by Journo.com.tr (Journo, 2025; Journo, 2024; Journo, 2022). While the raw survey data from these sources were collected independently of this study, they were analyzed here as supplementary evidence to triangulate our findings.

All interview data were anonymized at the point of transcription and are reported in aggregate form to protect confidentiality. Given the lead author's professional affiliation with Journo, where some of the secondary datasets were originally produced and released as open data, additional care was taken to ensure analytic independence. The analysis reported here is original and interpretive, with no identifiable participant information included. Participants were informed that their insights would be used for scholarly purposes, and data integration followed community-in-the-loop principles aligned with Value Sensitive Design.

Table 4.2: Overview of the Two-phase Research Design

Component	Phase I: Quantitative	Phase II: Qualitative	Integrative
Purpose	Experimental trust elicitation	Open and semi-structured interviews	Triangulation of statistical patterns with lived expertise
Unit of Analysis	Outlets (51 outlets)	Practitioners (46 journalists, 15 international experts)	Bridging consumer-facing trust signals with newsroom perspectives
Method	Experimental trust elicitation	Thematic coding sensitized by ATiN, Rawls, VSD	Theoretical integration across empirical modes
Frameworks	ATiN dimensions, Rawlsian distributive fairness, VSD benchmarks	Thematic coding sensitized by ATiN, Rawls, VSD	Theoretical integration across empirical modes
Outcome	Reliability and external validity of trust measures	Practitioners' reasoning on trust, platforms, AI	Normative and practical implications for quality journalism

Phase II data were analyzed thematically, guided by both inductive coding and sensitizing concepts drawn from the Affective Trust in News (ATiN) framework, Rawlsian principles of fairness, and Value Sensitive Design (VSD). Table 4.2 provides an overview of the process. Codes (presented in Appendices) were developed iteratively, capturing recurrent themes across interviews (e.g., perceptions of audience cynicism, tensions between engagement and quality, normative visions of trust). Comparative coding between Turkish and international participants highlighted convergences and divergences in professional reasoning. This strategy enabled us to link practitioners' lived experiences to the normative benchmarks operationalized in Phase I, thereby bridging the quantitative and qualitative components of the study.

4.3. Results

4.3.1. Phase I

Table 4.3 presents the descriptive statistics for the trust scores aggregated at the outlet level ($N = 51$). The median of the average “blind” trust scores across outlets was 5.92, while the median of the average “brand-aware” trust scores was 5.62. The median for average “visual” trust was 6.00. The analysis also shows that the variability in mean

scores was greater when brand or visual cues were present ($SD = 1.01$ for brand; $SD = 1.04$ for visual) compared to when they were absent ($SD = 0.68$ for blind), indicating that source identity and design increased the differentiation between outlets.

Table 4.3: Descriptive statistics across outlets for blind, brand, visual trust, brand effect, and NQS.

Measure	N	Mean	SD	Min	Q1	Median	Q3	Max
blind_mean	51.0	5.829	0.683	4.104	5.407	5.921	6.423	6.921
brand_mean	51.0	5.611	1.017	3.838	5.003	5.624	6.498	7.619
visual_mean	51.0	5.765	1.046	2.500	5.000	6.000	6.500	8.000
brand_effect_mean	51.0	-0.218	0.801	-1.744	-0.620	-0.296	0.246	3.256
NQS	51.0	-0.000	2.608	-6.872	-1.660	0.201	1.585	5.964

To understand the relationships between our measures and validate them against an external benchmark, we performed a Spearman rank correlation analysis at the outlet level ($N = 51$). The full results are presented in Table 4.4.

Table 4.4: Spearman rank correlations.

Pair	ρ	p	N
blind_mean ~ NQS	0.35	0.01	51
brand_mean ~ NQS	0.68	< .001	51
visual_mean ~ NQS	0.18	0.21	51
visual_mean ~ blind_mean	0.18	0.21	51
visual_mean ~ brand_mean	0.29	0.04	51
brand_effect_mean ~ NQS	0.64	< .001	51

Note. ρ = Spearman's rank correlation coefficient. $p < .001$ values are shown where appropriate.

4.3.1.1. Blind vs. Brand-Aware Trust: Distinct Constructs

First, we examined the relationship between headline-only (blind) judgments and headline-plus-source (brand) judgments. As noted in our reliability analysis, the headline-level correlation between the median blind and brand scores was moderate ($\rho = 0.498$, $p < .001$). This confirms that while the two measures are related, they are capturing partially

distinct constructs. Revealing the source brand systematically alters the evaluation of the content itself, a finding that motivates the subsequent validity analysis.

4.3.1.2. Validation Against the Normative Quality Score (NQS)

Next, we assessed which measure served as a better proxy for journalistic quality by correlating them with the NQS. The analysis revealed a clear pattern:

- **Brand-aware trust** showed a strong and significant alignment with the NQS ($\rho = 0.68, p < .001$).
- **Blind trust** showed only a moderate but significant association ($\rho = 0.35, p = .01$).
- **Visual trust** was not significantly related to NQS ($\rho = 0.18, p = .21$).

This indicates that brand-aware trust is a substantially stronger indicator of community-validated journalistic quality than judgments based on content or visual design alone.

4.3.1.3. The Brand Effect Systematically Tracks Quality

Given that brand-aware trust aligns more closely with NQS, we analyzed the "brand effect," we previously defined as the difference between the mean brand and blind scores for each outlet. As Table 4.4 presents, we found that this brand effect itself strongly correlates with the NQS ($\rho = 0.64, p < .001$). This is a critical finding: the brand label does not add random noise. Instead, it systematically reweights judgments in a direction consistent with external quality signals. Outlets with stronger positive signals (prizes, editorial citations) tended to have positive brand effects, while those with negative signals (misinformation flags, professional sanctions) tended to have negative brand effects. This demonstrates that source disclosure activates reputational priors that align with journalistic merit.

Table 4.4 also demonstrate that brand-aware trust scores align most strongly with the NQS ($\rho = 0.68, p < .001$), while blind trust scores show only a moderate but significant association ($\rho = 0.35, p = .01$). Across outlets, brand effects varied in both directions, indicating that labels sometimes raise and sometimes lower perceived trust relative to headline wording alone. Outlets with stronger positive external signals (more prizes, more newsletter links, fewer misinformation/sanction entries) tended to show positive brand effects, whereas outlets with higher risk signals tended to show negative brand effects. This demonstrates that source disclosure reweights judgments systematically, not idiosyncratically.

In contrast, visual impressions are weakly related to NQS ($\rho = 0.18, p = .21$), indicating that user interface (UI) design cues alone do not strongly track the external benchmark. Visual trust also shows negligible association with blind ratings ($\rho = 0.18, p = .21$) and only a modest correlation with brand-aware ratings ($\rho = 0.29, p = .04$).

4.3.1.4. Component-level Predictors of Outlet Trust

To examine which benchmark components relate to each trust measure, we regressed outlet-level means on the counts of prizes, newsletter links, misinformation flags, and professional community sanctions using ordinary least squares (OLS) regression with HC3 robust standard errors (Table 4.5).

Table 4.5: Outlet-level OLS predicting blind, brand-aware, and visual trust from external indicators.

Predictor	Blind Trust	Brand-Aware Trust	Visual Trust
	Coef (SE)	Coef (SE)	Coef (SE)
Model Coefficients			
const	6.103*** (0.148)	5.587*** (0.187)	5.777*** (0.220)
Prize Count	0.061 (0.065)	0.098 (0.072)	-0.082 (0.190)
Newsletter Link Count	-0.004 (0.009)	0.007 (0.009)	0.003 (0.005)
Misinfo Count	-0.024* (0.012)	-0.023 (0.012)	0.010 (0.026)
Sanction Count	-0.026 (0.051)	-0.034 (0.055)	-0.120 (0.109)
Model Fit			
<i>N</i>		51	
<i>R</i> ²	0.220	0.387	0.065

HC3 standard errors in parentheses.

Significance: * $p < .05$, ** $p < .01$, *** $p < .001$.

Blind trust: Coefficients trended positive for prizes and negative for misinformation/sanctions, but individual predictors were weak and often non-significant. This aligns with the moderate blind–NQS correlation: headline-only judgments capture some normative signal but are only loosely structured by reputational indicators.

Brand trust: Coefficients were directionally consistent with the benchmark: prizes and newsletter links positive; misinformation and sanctions negative, with higher overall explanatory power (R^2) than blind models, mirroring the stronger brand–NQS correlation. Thus, revealing the outlet brings in informational priors that align brand-aware trust with NQS components.

Visual trust: A small positive weight on newsletter links and a mild negative weight on sanctions echo the modest visual–brand association; prizes and misinfo were inconsistent. This suggests design cues act as weak proxies for reputation, aligning more with brand-aware than blind judgments, but remaining poor predictors of the normative benchmark on their own.

For specific media outlets, Table 4.6 indicates a high degree of human–LLM concordance with compressed extremes, while Table 4.7 shows that brand labels reweight judgments in directions consistent with external quality signals for news. In both human

and LLM evaluations, independent outlets are overrepresented among the most trusted brands; four of the top six are independents. Likewise, six of the eight outlets with the largest positive brand effects (Bianet, Cumhuriyet, BirGün, T24, Gazete Duvar, and BBC Turkish) are independent newsrooms. Among independents, Dokuz8haber is the exception, exhibiting a negative brand effect despite a modestly positive NQS (Table 4.7).

Table 4.6: Mean trust scores for top Turkish media outlets (Humans vs. LLMs).

Media Outlet	Political Leaning	Human Avg. Trust	LLM Avg. Trust
Cumhuriyet	Independent	3.81	3.88
Fox / Now TV	Independent	3.86	3.76
NTV	Pro-Government	3.00	3.32
Halk TV	Independent	3.70	3.26
Habertürk	Pro-Government	2.98	3.24
Sözcü	Independent	3.48	3.16
CNN Türk	Pro-Government	2.67	2.92
Hürriyet	Pro-Government	2.50	2.76
Show TV	Pro-Government	2.31	2.52
Kanal D	Pro-Government	2.33	2.44
Star TV	Pro-Government	2.19	2.42
Haberler.com	Pro-Government	2.00	2.40
TRT	Pro-Government	2.38	2.38
Mynet	Aggregator	2.33	2.38
Sondakika.com	Pro-Government	2.33	2.22
Internethaber.com	Pro-Government	2.00	2.00
Ensonhaber.com	Pro-Government	2.00	1.90
ATV	Pro-Government	1.95	1.88
Sabah	Pro-Government	1.83	1.70
A Haber	Pro-Government	1.63	1.62

Table 4.7: Outlets with the largest positive and negative brand trust effects (Brand-Blind)

Outlet	Blind	Brand	B-B	Visual	NQS	Prizes	News.	Misinfo	Sanctions
Outlets with the largest positive brand effects									
bbc.com	4.10	7.36	3.26	7.00	5.96	5.00	348.00	0.00	0.00
aa.com.tr	6.48	7.62	1.13	7.50	1.07	0.00	55.00	5.00	0.00
bianet.org	6.23	7.29	1.06	6.50	2.52	0.00	73.00	0.00	0.00
cumhuriyet.com.tr	5.90	6.49	0.59	6.50	3.84	11.00	78.00	24.00	0.00
birgun.net	6.26	6.77	0.51	2.50	2.67	8.00	45.00	9.00	1.00
dha.com.tr	6.78	7.15	0.37	8.00	1.07	3.00	12.00	6.00	1.00
t24.com.tr	6.29	6.63	0.34	6.00	1.48	2.00	99.00	9.00	1.00
gazeteduvar.com.tr	6.42	6.75	0.33	5.00	5.74	6.00	152.00	0.00	0.00
Outlets with the largest negative brand effects									
gercekgundem.com	5.58	3.84	-1.74	6.00	0.31	0.00	16.00	6.00	0.00
sondakika.com	6.57	4.95	-1.62	6.00	-1.10	0.00	0.00	5.00	0.00
dokuz8haber.net	6.67	5.13	-1.54	6.00	1.21	1.00	0.00	0.00	0.00
internethaber.com	5.93	4.49	-1.43	6.00	0.20	0.00	0.00	0.00	0.00
medyaradar.com	5.38	4.26	-1.12	5.00	0.20	0.00	0.00	0.00	0.00
medyatava.com	5.19	4.21	-0.98	6.00	-1.21	0.00	0.00	6.00	0.00
yenisafak.com	5.92	4.98	-0.94	6.50	-5.09	0.00	0.00	18.00	6.00
ensonhaber.com	6.02	5.08	-0.93	6.00	-2.14	0.00	0.00	24.00	0.00

Additional tables and figures supporting these conclusions are listed in the appendices.

The principal findings are summarized as follows:

- 1) **Blind (headline-only) trust** contains intrinsic quality signal ($\rho = .35$ with NQS) and is highly reliable, but is only moderately aligned with the benchmark.
- 2) **Brand-aware trust** is both reliable and strongly aligned with NQS ($\rho = .68$), and significantly more so than blind trust (Fisher z test; jackknife-stable).
- 3) **Visual trust** is weakly aligned with NQS overall but modestly aligned with brand-aware trust, indicating that design cues approximate reputational priors more than textual plausibility.
- 4) **The brand effect (brand – blind)** correlates strongly with NQS, demonstrating systematic reweighting of perceived trust by source labels in directions consistent with external signals.

These results provide a quantitative basis for the next phase of this study: an algorithmic veil can extract content-intrinsic assessments (blind), while a controlled reintroduction of label/reputation can be used downstream as a quality-guardrail when the

platform's objective function explicitly values alignment with community-validated quality signals.

4.3.2. Phase II: Journalistic and Managerial Perspectives

4.3.2.1. Theme 1: Brand-Triggered Trust and the Affective Logic of ATiN

The interviews with media experts and comparative materials strongly corroborate the Affective Trust in News (ATiN) framework developed in Study 1. Journalists and managers alike describe trust as inseparable from political allegiance, not as a cognitive evaluation of journalistic quality. Producers explicitly recognize that media brands operate within what one manager called a “dichotomy of praise/slander for the government and praise/slander for the opposition.” For audiences, logos such as TRT or Sabah activate schema-based expectations: for pro-government citizens, unsurprisability (A1) generates affective comfort by reinforcing a familiar worldview; for dissenters, the same symbols trigger immediate distrust. This dynamic exemplifies Brand Schema Activation (A4) and the A(e)ffect of the Other (A3), as state-aligned outlets cultivate trust by defining political opponents and independent journalists as “terrorists” or foreign agents, while opposition audiences strengthen solidarity through rejection of the “penguin media.” In sum, the findings validate that brand is not a quality heuristic but a political signal, functioning as the primary affective trigger in Turkey's polarized media field.

4.3.2.2. Theme 2: Metrics and Platform Pressure

Managers repeatedly described their dependence on digital platforms (particularly Google) for traffic, advertising revenue, and thus organizational survival. This Platform Dependence (B4) is exacerbated by Algorithmic Opacity and Volatility (B2), leaving outlets vulnerable to unannounced changes that can drastically reduce reach. Respondents noted their daily fixation on real-time traffic dashboards (“*her an*,” “*saatlik*”), illustrating the volatility of Perceived Reward/Penalty Cycles (B5). While managers recognize qualitative metrics such as average engagement time as better indicators of journalistic quality, their operational choices are driven by blunt quantitative metrics (unique visitors,

page views) demanded by advertisers. This contradiction creates what reporters call “*tık avcılığı*” (clickbait), a practice viewed as antithetical to their professional ethics. Independent outlets face intensified Resource Constraints (F2), which drive a gradual pivot toward reader revenue models (F3) as the only sustainable alternative to advertising-based dependence.

4.3.2.3. Theme 3: Repair Repertoires and Accountability

Traditional mechanisms for repairing trust, such as corrections boxes or transparency pages, were conspicuously absent from interviewees’ accounts. Reporters emphasized preventive integrity (“avoiding errors at all costs”) over reactive correction, often invoking the idea that a flawed story becomes “*çöp*” (garbage). The ATiN framework explains why corrections lack traction: in a system where trust rests on affective alignment, cognitive interventions such as corrections are either ignored or politically weaponized. Pro-government outlets have no incentive to contradict the state narrative, while independent outlets risk having their corrections framed as proof of unreliability by hostile actors. Instead, a new form of accountability is emerging, rooted in reader-revenue imperatives: community engagement (C4) and transparency become survival necessities in business models that prioritize reader revenue. Thus, while traditional repair repertoires (C1–C3) remain marginal, structural shifts toward audience-funded models create conditions for deeper accountability.

4.3.2.4. Theme 4: AI Adoption and the Guardrail Gap

AI adoption in Turkish newsrooms is widespread but pragmatic, shaped by precarity. Most journalists and managers use tools such as ChatGPT to automate translation, transcription, and summarization (D1). Reporters frame AI as a personal productivity assistant, while managers describe it as indispensable for efficiency under shrinking budgets. Yet ethical guardrails remain informal. Although nearly all respondents endorse the principle of human-in-the-loop oversight (D2), no outlet has institutionalized AI policies or training. Risks are understood differently: reporters emphasize micro-level threats (hallucinations, erosion of skills, audience mistrust), while international experts

warn of systemic risks (D3): the weaponization of generative AI by state actors to amplify propaganda. Interviewees also note uncertainty about audience perceptions (D4), raising unresolved questions about transparency and disclosure in AI-assisted news.

4.3.2.5. Theme 5: International Comparison

International experts contextualize Turkey's challenges within broader media trends. In the U.S. and Europe, debates on trust center on market adaptation: integrating AI, diversifying revenue, and maintaining engagement. These struggles unfold within relatively intact institutional environments. In contrast, respondents emphasize that in fragile democracies like Turkey, these market pressures are layered atop existential political threats. Independent journalism faces not only disruption but also active weaponization of markets and platforms by the state. Accordingly, solutions proposed for Turkey must first address informational justice (E1) and media independence (E2) before market reforms can succeed.

4.3.2.6. Divergences and Limits

Not all media expert perspectives align with the dominant narrative. Three important divergences emerged:

Managers vs. Reporters: Newsroom managers treat metrics as survival tools, while reporters see them as corrupting journalistic quality. They view AI as a potential efficiency lifeline, while reporters fear erosion of craft.

Legacy vs. Digital-Native Outlets: Legacy organizations remain bound by broadcast-era assumptions of brand authority, while digital natives (e.g., Medyascope) increasingly focus on models around community engagement and reader funding.

Counter-Cases: Some internal newsroom metrics to evaluate performance (e.g., engagement time) can align with quality (H1), and brands can still serve as positive credibility signals (H2) in an otherwise polluted information environment. Finally,

blurred roles (e.g., reporter-founders acting as both strategists, technical officers and journalists) complicate neat distinctions between managerial and craft perspectives.

4.4. Discussion

A useful way to assess the external validity of our LLM-based measures is to compare them with independent human surveys. Two such references are available: Study 1’s nationally representative survey ($N \approx 1,000$) in which respondents freely named five news brands and rated each on a 1–5 trust scale (brand-aware, unaided recall), and the Reuters Institute’s Digital News Report (DNR 2025) brand-trust breakdown for Turkey (0–10 scale; respondents unaware of a brand are excluded; results summarized as % “Trust” [6–10], “Neither” [5], “Don’t trust” [0–4]).

Three regularities line up across all three sources (our brand-aware outlet means, our earlier 1–5 survey, and DNR 2025 percentages):

High–moderate trust for prominent centrist and independent news brands: In the Study 1 survey, Halk TV, Fox/Now TV, CNN Türk, NTV, Show TV, and Habertürk cluster in the upper half of the 1–5 range (3.3–3.8). DNR 2025 likewise places Halk TV (53% Trust), Sözcü TV (54%), NTV (52%), Cumhuriyet (52%), Habertürk (51%), and CNN Türk (50%) toward the top of the Turkish distribution. Our brand-aware outlet means show the same direction of travel: the corresponding news websites (e.g., halktv.com.tr, ntv.com.tr, haberturk.com, cumhuriyet.com.tr) sit at or above the sample median, and several exhibit positive brand effects ($\text{Brand} - \text{Blind} > 0$), indicating that the label boosts trust beyond what headline wording alone predicts. This alignment suggests that the LLM, when given the outlet label, encodes widely held reputational priors that are also visible in human polling (DNR 2025).

Lower trust for pro-government news brands: DNR 2025 reports comparatively lower trust (and higher “Don’t trust”) for Sabah, ATV, and A Haber; and the Turkey survey placed ATV and A Haber toward the lower end of the 1–5 spectrum. In our

data, web counterparts for this cluster show lower brand means and negative brand effects, consistent with the hypothesis that labeling activates partisan priors that depress perceived trustworthiness (DNR 2025).

Quality-adjacent titles benefit from branding: Independent outlets with stronger professional or civic reputations in the Turkish information sphere (e.g., *cumhuriyet.com.tr*, *t24.com.tr*, *gazeteduvar.com.tr*) show positive Brand – Blind deltas and relatively higher brand means in our study. Relatively less partisan pro-government broadcasters (e.g. NTV) rises slightly in LLM averages relative to humans, suggesting either a civility or mainstreaming prior in the model or differences in TV vs. web schemas that the model encodes.

Despite the broad convergence in sign and rank-order tendencies, several differences occur:

Scale and framing effects: Our Study 1 survey used a 1–5 trust scale in an unaided recall format (respondents first list up to five brands that come to mind, then rate them). DNR 2025 reports 0–10 brand-specific scores for a fixed list and excludes those who have not heard of the brand. Our LLM ratings are headline-anchored (blind/brand) or design-anchored (visual), aggregated over 1,000 headlines per outlet (2023). These distinct elicitation contexts lead to different absolute levels and dispersion, so agreement should be assessed mainly in relative rather than absolute terms (DNR 2025).

Awareness filters: DNR 2025 explicitly drops non-aware respondents per brand, mechanically inflating average trust relative to a pure population-level estimate. Our recall-based instrument tends to over-sample salient brands (high reach/top-of-mind), which can amplify the position of large broadcasters. Our approach is content-complete and cannot condition on public awareness; the LLM sees all sampled headlines/screens regardless of reach. Consequently, our brand means may be more conservative (narrower top values) while still reproducing rank-order patterns.

Modality and medium: The human surveys mostly reference TV-centric brand schemas, whereas our evaluation targets web properties (e.g., Halk TV televised brand vs halktv.com.tr site). Where online and broadcast reputations differ, some divergence is natural. DNR 2025 itself cautions that the list is not exhaustive and that figures reflect subjective public opinion rather than an objective audit. (DNR 2025).

Time windows and events: Our corpus is calendar-year 2023; DNR 2025 reflects fieldwork from early 2025; our survey timing predates both. Shifts in editorial positioning, major political events, or high-visibility controversies can move brand sentiment in the short run, especially in polarized environments as Turkey. The DNR 2025 Turkey brief reports historically low general trust and marked polarization, which can magnify year-to-year variance (DNR 2025).

4.4.1. Implications for Validity

First, the brand condition is the closest analogue to survey trust because it explicitly activates the outlet schema. The strong positive correlation between brand means and NQS ($\rho \approx .68$) and the substantial negative correlation between Brand – Blind and mis/disinformation indicators in our regressions suggest that the LLM’s brand-aware judgments co-move with external, community-validated signals. Second, our blind condition recovers a weaker (but still positive) association with NQS ($\rho \approx .35$), implying that headline wording carries quality cues, though reputational information adds a large share of predictive power, consistent with both our survey and DNR 2025 patterns that are largely brand-driven. Third, visual trust aligns only weakly with both NQS and our textual scores, indicating that design heuristics alone are an unstable basis for quality inference; again compatible with survey evidence that brand identity, not presentation, dominates trust judgments.

Given these triangulations, our results support using brand-aware ratings as the primary bridge to human trust surveys, with blind ratings serving as a content-sensitive

check and visual ratings as a secondary, weaker signal. Where our brand-aware means and DNR 2025 diverge at the margin, the differences are explainable by awareness filters, modality (TV vs. web), and timing. Overall, the alignment is strong enough to justify our use of brand-aware LLM scores for outlet-level validation and for testing Rawlsian algorithmic veil designs that aim to decouple content quality from reputational priors. Having established the robustness of our measures through external triangulation, we now turn to the substantive implications of our findings.

4.4.2. What This Study Demonstrates

Phase I quantified how different cues relate to an external benchmark of journalistic merit. Brand-aware trust aligned most strongly with the Normative Quality Score (NQS; $\rho \approx .676$, $p < 0.01$), significantly exceeding the headline-only alignment ($\rho \approx .354$, Fisher-z difference = -2.211 , $p = .027$). Visual impressions of homepages showed weak association with the benchmark ($\rho \approx .179$). A dedicated reliability study confirmed that the LLM rater is highly stable ($ICC(1,5) \approx .995-.996$). Phase II explained these patterns from practitioners' vantage point: producers described brand-triggered trust, platform-metric dependence and opacity, thin repair repertoires, and pragmatic AI adoption without formal guardrails, with international experts emphasizing that informational justice and independence are preconditions for any durable market fix.

Taken together, these results support the ATiN thesis: in polarized systems, the brand acts as an affective trigger; text alone carries some merit signal; and visual design is, at best, a weak proxy for quality. They also motivate a design response that is both normatively defensible and operationally tractable. But how can the process be improved for the public good while upholding individual autonomy and tastes?

4.4.3. Justice as Fairness in Algorithmic Evaluation of News Quality

Rawls's two principles of justice offer a clear lens for platform curation of news (Rawls, 1999). The Liberty Principle requires that each person enjoy equal basic liberties, including access to political information necessary for autonomous judgment. The

Equality Principle, comprising fair equality of opportunity and the difference principle, requires that institutional arrangements be structured so that advantages do not entrench arbitrary inequalities and that any inequalities work to the benefit of the least-advantaged.

Applied to ranking, an outlet's brand identity is typically a morally irrelevant fact with respect to a user's equal claim to encounter quality information. Our Phase I results show that brand cues heavily steer judgments (high ρ with NQS) because they activate affect, not because they are a content-intrinsic merit signal. Under a veil of ignorance, choosing rules without knowing which brand one will encounter or prefer, reasonable stakeholders would not license an algorithm that privileges or penalizes content because of the label. Rather, they would endorse a procedure that scores quality independently and only then admits reputation and preference for personalization within quality guardrails.

This Rawlsian stance also resonates with Kant's injunction to treat persons as ends in themselves (Kant, 1997). A ranking system that primarily optimizes for brand-driven engagement risks treating users as means to click-through and outlets as means to ad yield, sidelining their ends as citizens seeking reasons to believe. As Sandel (2009) observes, justice is not only about the right distribution but also about the right valuation of goods. News is not merely an engagement commodity; its telos is to enable public reason. A curation rule that centers content-intrinsic quality at scoring time better honors that telos.

4.4.4. From Principle to Practice: A Value-driven / Value-sensitive Blueprint

We operationalize these commitments using a Value-Driven Design (VDD) and Value-Sensitive Design (VSD) hybrid. Three premises guide the blueprint:

Value as the primary driver: In news distribution, the relevant value is the public-interest quality of content. Both journalists and audiences are stakeholders; their values must shape system choices (Friedman, Kahn, & Borning, 2013).

Quantitative value measurement is necessary: Quality is hard to measure, but communal signals exist. Our NQS aggregates positive indicators (citations in

newsletters, recognized prizes) and negative indicators (fact-checker flags, press-council sanctions). It is imperfect yet auditable and improvable: a first step toward measurable value (Collopy & Hollingsworth, 2011).

Multi-stakeholder optimization: Whether LLMs or social media or search engines, informational platforms serve users, outlets, and the democratic public; their interests can conflict. Iterative design with the community should balance these interests in line with Rawlsian justice, ensuring that the rules do not cement arbitrary advantages and that the least-advantaged information seekers are not crowded out by popularity signals.

To move from principle to practice, we translate the VDD and VSD commitments into five operational requirements that are measurable, enforceable, auditable, and scalable within ranking workflows. Each requirement pairs a concrete rule with a rationale linked to findings from Studies 1 and 2 and to the justice aims outlined above. The design institutes procedural ignorance at scoring to protect intrinsic quality estimation from label activation, constrains engagement optimization with tiered quality guardrails, and provides transparency and contestability for users and publishers. It also places the benchmark (NQS) under community-in-the-loop governance and adds newsroom AI guardrails for disclosure, editing, training, and audit trails. Together, these provisions aim to advance equality of informational opportunity for the least advantaged while preserving user choice and newsroom autonomy; the requirements that follow specify how.

Design requirement 1: Procedural ignorance at scoring

Rule: Evaluate items without brand cues when estimating intrinsic quality (the “veil at scoring”). Use content-level signals (e.g., headline semantics) that Phase I showed to correlate with NQS ($\rho \approx .35$).

Rationale: This reduces affective bias from label activation while preserving liberty: brands remain visible in the interface; only the algorithmic scoring step is shielded.

Design requirement 2: Quality guardrails with tiering

Rule: Map items to quality tiers anchored to the benchmark; then allow personalization within tiers. Engagement optimization is confined inside tiers; bottom-tier promotion is capped.

Rationale: Phase I showed that brand effects covary with NQS ($\rho \approx .65$) but can crowd out content-intrinsic merit. Tiering keeps liberty and choice while aligning exposure with equality of informational opportunity.

Design requirement 3: Transparency and contestability

Rule: Provide user-facing “Why am I seeing this?” rationales and appeal channels for outlets (e.g., challenge tier assignment) and users (e.g., tune diversity).

Rationale: Phase II highlighted algorithmic opacity and perceived penalty cycles. Transparent reasons and contestability instantiate procedural fairness and respect user autonomy (Kant’s ends).

Design requirement 4: Community-in-the-loop benchmark governance

Rule: Continuously enrich and maintain NQS via an open pipeline: documented indicators, periodic updates, and public change logs. Invite journalism organizations and civil society to co-govern indicators.

Rationale: Sandel’s concern with how we value goods is answered by a community-grounded metric; Rawls’s difference principle is served by democratic governance that prevents the rich from defining merit alone or the uninformed from setting the norms.

Design requirement 5: AI guardrails in newsroom workflows

Rule: Formalize human-in-the-loop editing, disclosure policies where AI assistance is material, and style-integrity safeguards; provide training and audit trails.

Rationale: Phase II showed informal ethics but no institutional policy. Guardrails protect both craft integrity (journalists' concern) and systemic resilience against state-scale manipulation (international media experts' concern).

4.4.5. Reconciling News Brand's Double Role: Political Trigger and Credibility Cue

Phase II underscores a duality. For many citizens, brand is a political trigger (ATiN); for others, it is a credibility shorthand in a polluted infosphere. The blueprint preserves both realities: exclude brand only at scoring time to protect fairness; retain brand in the user interface and within-tier personalization to respect agency and safety. This arrangement satisfies the Liberty Principle (users see what they choose and can follow outlets) and the Equality Principle (initial exposure reflects content-intrinsic merit, not sheer label power).

Platforms should deploy veil-at-scoring, tiered guard-rails, and fairness dashboards that report exposure shares by quality tier and appeal outcomes. Newsrooms should invest in source disclosure modules, visible correction surfaces, and community channels that matter in reader-revenue models, coupled with formal AI policies. Regulators and funders can support benchmark governance boards, replication, and contestability APIs so that appeals and evidence travel across systems rather than into private inboxes.

4.4.6. Boundary Conditions and Limitations

Our approach is tailored to Turkey (2020–2025), where state influence and consolidation heighten ATiN dynamics. In more plural media systems, brand may be a less dominant affective driver. Certain cases (e.g. live safety alerts, investigative exclusives where source identity is part of verification) may warrant a soft veil (down-weight brand rather than fully exclude it at scoring), accompanied by explicit disclosure.

Constraints of the proposed benchmark can also be cited here. NQS draws on a limited indicator set; blanks were coded as zero; some deserving outlets may be under-credited. Indicator enrichment is a priority for future work. While reliability was excellent

(ICC(1,5) $\approx$.995–.996) and headline-level signals tracked NQS moderately, the LLM as a rater is only a proxy, not a substitute for public judgment. Furthermore, our headlines are from 2023, screenshots capture only two slices, and the news ecosystem evolves every day. As a result, our analyses are associational; and platform-scale field tests are needed to estimate causal effects on exposure and behavior.

4.4.7. Research Agenda

In future studies, cross-national replications can assess external validity across different levels of polarization. Human–AI hybrid raters and multi-model ensembles can probe the “syntax trap” in quality inferences. Dynamic NQS pipelines should incorporate richer positive (peer citations, high-quality corrections) and negative (systematic misreporting patterns) signals, governed by a multistakeholder board. Finally, field experiments of veil-then-personalize, with public fairness dashboards, can test whether the design improves informational opportunity without suppressing user choice.

The four NQS components (journalism prizes, editorial citations in professional newsletters, press-council sanctions, and fact-checking flags) were selected in this study because they are ready-made signals that are publicly auditable and regularly updated by independent actors within the profession. Prizes operationalize peer recognition of excellence; editorial citations capture curatorial endorsement by field insiders; sanctions document formal norm violations within media accountability systems; and fact-checking flags register verifiable informational failures (Eberwein, Fengler, & Karmasin, 2018; Graves, 2016; Shoemaker & Vos, 2009). Taken together, these indicators approximate a minimal, reproducible benchmark that is not confounded by platform engagement or popularity metrics and is anchored in journalism’s own normative standards and role conceptions, which are historically grounded in professional self-regulation and the public-service mandate of news (Carlson, 2017; Hanitzsch, Hanusch, Ramaprasad, & de Beer, 2019; Kovach & Rosenstiel, 2014; Hallin & Mancini, 2004). We view this as an expandable scaffold: international journalism bodies could maintain a continuously updated, cross-national dataset and, through a democratic community process, decide which positive and negative signals warrant inclusion. This community-governed

approach aligns measurement with journalistic values and public-interest goals, rather than delegating quality judgments to lay engagement or to technology firms and governments that lack domain accountability (Napoli, 2011).

4.5. Conclusion

As grandmaster Siegbert Tarrasch once observed in his classic analysis of chess, “Mistrust is the most necessary characteristic of the chess-player” (Tarrasch, 2012, p. 79). In strategic games, vigilance against deception safeguards rational play. In the sociotechnical arena of news and platforms, however, mistrust functions less as epistemic vigilance and more as affective alignment. Our findings show that affectively anchored trust and mistrust, while stabilizing partisan communities, undermines the collective conditions for public reasoning that Rawlsian justice demands in polarized, fragile democracies such as Turkey.

When brand identity dominates trust via affect, enforcing procedural ignorance at scoring and applying tiered quality guardrails can realign ranking with informational justice while preserving liberty and agency. Phase I quantified the biasing role of labels and demonstrated a reliable content-intrinsic signal; Phase II explained why practitioners experience the system as structurally skewed and why transparency, contestability, and reader relationships matter. A value-sensitive design that measures journalistic merit with auditable, community-anchored signals, balances stakeholder interests, and honors the principles of liberty and equality offers a concrete path for platforms and newsrooms to better serve citizens and democracy.

5. CONCLUSION

This thesis has embarked on a rigorous examination of the profound challenges confronting news trust in the contemporary digital age. At its core, the work has systematically investigated how citizens form trust in news, how advanced artificial intelligence systems evaluate news content, and, critically, how platforms, as the primary conduits of information, ought to allocate visibility to news in a manner consistent with democratic principles and ethical constraints. Across three interconnected studies, the findings converge on a core diagnosis and a principle-based practical remedy, offering a comprehensive understanding of the current information ecosystem and a forward-looking normative blueprint for its improvement.

5.1. Restatement of Research Problem and Objective

The fundamental research problem addressed by this thesis is the escalating crisis of news trust, exacerbated by two intertwined forces: the highly polarized and often state-influenced media landscapes in many democracies, and the pervasive influence of algorithmic platforms in news distribution. This conjuncture raises critical questions about the very possibility of informed public discourse and the exercise of sound political judgment. Prior research consistently highlights that platform design and user networks often accentuate prior preferences, obscure system behavior, and distort exposure to information. A key conceptual challenge identified at the outset is the reductive nature of the term "trust" itself, frequently compressing distinct constructs such as affective brand loyalty, source reputation, and content-intrinsic quality. This conflation not only

complicates diagnosis but also obstructs the development of effective remedies, transforming questions of credibility and exposure into design problems with constitutional stakes.

The overarching objective of this thesis was therefore multifaceted:

- 1) **To deconstruct and re-conceptualize "trust in news":** Moving beyond a monolithic understanding to differentiate between affective, reputational, and quality-based dimensions.
- 2) **To empirically diagnose the "trust-quality gap":** Examining the disconnect between public trust in news brands and the actual journalistic quality of their content, particularly in a critical case like Turkey.
- 3) **To critically audit the capabilities and limitations of AI in news evaluation:** Uncovering potential failure modes of LLMs when assessing news content, especially concerning subtle forms of propaganda.
- 4) **To develop and validate a normative framework and a practical design solution for news ranking:** Proposing an approach for platforms that enhances informational justice by separating content quality from potentially misleading brand and affective cues, while respecting user agency and democratic accountability.

Table 5.1: A Tabular Summary of the Studies**Study 1 — ATiN: Affective Trust**

Facet	Description
Purpose	Build a theory of trust in polarized systems during major news events
Core Question	What drives outlet trust in Turkey amid the 2023 earthquakes and election?
Data & Scope	National survey and news-consumer interviews (2023 crisis window)
Method	Mixed methods; grounded theory
Key Finding	Trust tracks political allegiance, not journalistic merit, via three mechanisms: <i>unsurprisability</i> , <i>affective self-fact-checking</i> , <i>A(e)ffect of the Other</i>
Output	Affective Trust in News (ATiN) framework
Handoff	Mechanisms/predictors handed to Study 2 for testing

Study 2 — Audit: Human vs. Expert vs. LLM

Facet	Description
Purpose	Diagnose divergence between user trust and content quality; test AI evaluation
Core Question	Do public trust and quality align, and how do LLMs compare to experts?
Data & Scope	Public trust scores; expert blind ratings; LLM outputs (blind vs. brand-aware)
Method	Comparative evaluation; split-condition design (humans and models)
Key Finding	<i>Trust-quality decoupling</i> in humans; LLM <i>Propaganda Blind Spot</i> in blind tasks
Output	Audit framework; three logics: public (identity), experts (norms), LLMs (patterns)
Handoff	Design constraints/requirements passed to Study 3

Study 3 — Design: Algorithmic Veil

Facet	Description
Purpose	Develop and test a principled remedy for fairer news ranking
Core Question	Can we score content fairly by separating quality from reputation, then personalize within guardrails?
Data & Scope	Balanced corpus of ~51,000 headlines (51 outlets, 2023); NQS signals; practitioner/expert interviews
Method	Rawls-inspired <i>veil-then-personalize</i> ; ICC reliability; external validation vs. NQS
Key Finding	Blind scoring is highly repeatable; brand-aware means align with NQS; veil reduces label-driven distortion while preserving choice
Output	Blueprint: scoring rule under a veil, NQS tiers, guardrails, transparency & appeals, community governance
Handoff	Implementable design informed by Studies 1–2; basis for replication/tests

5.2. Conclusive Summary of Thesis Findings

Across three interconnected studies (Table 5.1), this thesis reports related findings that clarify key challenges and opportunities in the news ecosystem.

Study 1, titled "Affective trust in Turkish news: A mixed method grounded theory study on media brands amid a national crisis," introduced and rigorously developed the Affective Trust in News (ATiN) framework. Through mixed-methods qualitative and quantitative inquiry conducted during the 2023 Turkish earthquakes, which was a critical juncture that amplified underlying dynamics, this study demonstrated that in polarized and state-influenced systems like Turkey, trust functions primarily as an affective alignment with a media brand's political and emotional worldview, rather than as a cognitive appraisal of journalistic quality. Three key socio-psychological mechanisms were identified that stabilize this affective alignment:

- **Unsurprisability:** Citizens, on both sides of the political divide, anticipate bias from media outlets, leading to a resigned expectation that neutralizes cognitive dissonance and reinforces their existing affective stance.
- **Affective self-fact-checking:** Individuals engage in what appears to be "fact-checking," but it is primarily an emotionally driven process to defend foundational emotional states linked to their political identity, rather than a neutral quest for objective truth.
- **The A(e)ffect of the Other:** Trust is relationally constructed by defining one's in-group against a political "Other," with emotional responses to this out-group directly shaping the trust relationship with preferred media, fostering affective communities.

This study established the headline as an appropriate unit of analysis, demonstrating that even small lexical choices can operate as partisan signals, eliciting affectively driven trust.

Building on the insights from Study 1, Study 2, "The Propaganda Blind Spot: AI, Polarization, and the Disconnect Between Media Trust and Quality," extended the diagnosis to AI systems by pitting human judgments against LLM ratings. This study documented a dual crisis:

- **A trust–quality decoupling among audiences:** It empirically showed that the Turkish public's trust in media brands had a near-zero correlation with the journalistic quality of their content as judged by experts in blind tests. This provided evidence that brand loyalty and political identity had superseded content merit as the basis for public trust.
- **The "Propaganda Blind Spot" in LLMs:** Crucially, LLMs that could correctly identify state-affiliated media as untrustworthy in non-blind, brand-aware settings were simultaneously deceived by their content in blind evaluations. These models awarded high "objectivity" ratings to syntactically neutral headlines from low-trust, state-controlled outlets, mistaking structural neutrality for genuine journalistic impartiality.

The findings highlighted a critical disconnect between LLMs' stored, non-blind knowledge about an outlet's reputation and their ability to detect subtle, context-dependent linguistic features of bias in a blind textual analysis. This "Propaganda Blind Spot" was attributed to a failure of pragmatics, where LLMs processed syntax and semantics but missed the illocutionary force (the context-dependent, intended purpose) of language. The study concluded that public, expert, and AI evaluations operate on fundamentally distinct logics: identity for the public, norms for experts, and patterns for LLMs.

Motivated by the diagnoses of the first two studies, Study 3, "Trust Beyond the Brand: A Rawlsian Design for LLM-Assisted News Ranking — Evidence from Turkey," made the prescriptive turn. It proposed and empirically evaluated a Rawlsian, Value-Sensitive Design approach for news scoring that deliberately separates content-intrinsic quality from reputational and affective cues at the moment of evaluation and scoring, followed by personalization within quality guardrails. Key findings included:

- **Validation of the Normative Quality Score (NQS):** A community-anchored, auditable composite benchmark of journalistic merit was developed, combining positive signals (e.g., journalism prizes) with negative signals (e.g., fact-checker flags).
- **Context-gradient LLM protocol:** A novel protocol employing LLMs under blind (headline only), brand-aware (headline plus outlet label), and visual (homepage screenshots) conditions demonstrated that brand-aware ratings aligned most strongly with NQS ($\rho \approx .676$), blind ratings moderately ($\rho \approx .354$), and visual impressions weakly. This showed that brand labels systematically reweight judgments in directions consistent with external quality signals.
- **Procedural veil at scoring time:** The study demonstrated that evaluating content under a "veil" that excludes brand cues is both reliable and effective in aligning item-level quality assessment with community-anchored benchmarks.
- **Practitioner insights:** Interviews with Turkish journalists and international experts corroborated the ATiN framework, highlighted platform dependence and opacity, and underscored the rapid, yet unguarded, uptake of AI in newsrooms.

The proposed "veil-then-personalize" protocol was shown to be highly reliable as a measurement instrument and to align brand-aware judgments with community norms without inadvertently canonizing specific news brands. The sequence of the three studies thus moves logically from diagnosis (affective signaling), to audit (human-model misalignment), to design (norm-anchored scoring and governance), culminating in a feasible blueprint for fairness-aware ranking.

5.3. Contribution to the Field

This thesis offers contributions in theory, method, and practice that inform research on news trust, AI evaluation, and responsible information system design.

5.3.1. Theoretical Contributions and Implications

First, this work fundamentally challenges unitary conceptions of "trust in news" by demonstrating the necessity of decomposing trust into distinct, measurable components: affective brand attachment, reputational familiarity, and content-intrinsic quality. Treating these as a single construct obscures the powerful political and emotional drivers of audience judgment and hinders effective intervention.

Second, the Affective Trust in News (ATiN) framework offers a novel theoretical lens that integrates and extends insights from selective exposure, partisan media choice, motivated reasoning, and affective polarization. By specifying three socio-psychological mechanisms (unsurprisability, affective self-fact-checking, and the A(e)ffect of the Other), ATiN clarifies how affect and identity come to dominate outlet trust in polarized systems, often operating at the moment of recognition, even before content is deeply read. These mechanisms are portable and travel beyond the specific Turkish case, providing a valuable tool for analyzing media trust in other polarized contexts globally.

Third, the identification and characterization of the "Propaganda Blind Spot" represent a crucial theoretical contribution to the understanding of AI's limitations in socio-technical contexts. It highlights that LLM competence in syntax and semantics does not guarantee pragmatic understanding of communicative intent, particularly when propagandistic styles are calibrated to appear superficially neutral. This explains a critical failure mode in AI evaluation, emphasizing why brand cues can dominate human judgments and why high-quality content signals cannot reliably correct identity-driven evaluations without design changes that modulate when and how cues are applied.

5.3.2. Methodological Contributions and Implications

Methodologically, this thesis treats LLMs as auditable measurement instruments rather than oracles and sets out a systematic context-gradient protocol using blind, brand-aware, and visual rating conditions to separate content-intrinsic signals from reputational and design cues. This protocol, combined with rigorous repeatability checks, cross-

condition correlations, and external benchmarking against a community-anchored score, establishes a comprehensive audit trail for model outputs. This approach sets a new standard for instrument validity in the use of AI for social science research, linking it directly to governance by making the benchmark auditable, updateable under community oversight, and transparent regarding its sensitivity to indicator choices.

Methodologically, the Normative Quality Score (NQS) provides a community-anchored and auditable composite that serves as an external benchmark for news quality beyond popularity or sentiment. The same logic can be applied in domains where reputational cues influence item-level ratings and where models overvalue stylistic neutrality relative to content. The dual-task design (pitting non-blind trust against blind quality evaluation for both humans and AI) also serves as a new methodological framework for systematically diagnosing the trust-quality gap and auditing the contextual understanding and latent biases of AI systems.

5.3.3. Practical Implications

The thesis translates its results into actionable guidance for platforms, regulators, newsrooms, designers, and the public.

First, the proposed Rawlsian "veil-then-personalize" protocol offers a tangible blueprint for fairness-aware news ranking. By advocating for scoring content under a procedural veil that excludes brand cues at the point of evaluation, the thesis provides a mechanism to make allocations of visibility more justifiable to citizens with diverse loyalties, directly addressing informational injustice. After initial scoring, personalization can proceed within quality guardrails, preserving user liberty while protecting equal informational opportunity.

Second, the thesis specifies governance features that are both feasible and essential for platforms and regulators. These include tiered quality guardrails (mapping items to tiers anchored to NQS), transparency and contestability channels (user-facing rationales and appeal mechanisms), and community-in-the-loop governance (open NQS pipelines,

periodic updates, and multistakeholder oversight). This framework aligns instrumentation, scoring, and accountability, offering a pathway to reduce the impact of affective triggers at the most consequential point in the information pipeline.

Third, the work provides critical guidance for newsroom AI guardrails, emphasizing the need for formal human-in-the-loop policies, disclosure where material, targeted training, and robust audit trails for AI integration. This proactive approach helps news organizations harness the benefits of AI while mitigating its risks, particularly against state-scale manipulation.

Finally, the blueprint accommodates boundary cases, such as urgent safety alerts or stories where source identity is an integral part of verification, through "soft-veil" exceptions that attenuate but do not erase reputational cues. This pragmatic flexibility enhances the real-world applicability of the proposed design. By providing an implementable design for platforms and regulators, the thesis offers a concrete path to improve both fairness and reliability criteria in news distribution, outperforming engagement-only optimization while remaining compatible with user choice and democratic accountability.

5.4. Limitations and Further Studies

Alongside its contributions, this thesis acknowledges limitations that suggest avenues for further study.

First, the empirical context of Turkey, with its heightened polarization and structural media capture, magnifies the ATiN dynamics and the observed effect sizes. While the theoretical mechanisms are portable, their context-dependence suggests that effect sizes for brand dominance might be smaller in less polarized systems, warranting cross-national replications across varying levels of polarization.

Second, the Normative Quality Score (NQS), while robust and auditable, is necessarily incomplete. It may inherit certain prestige biases due to its reliance on

established indicators like journalism prizes. Future work should focus on enriching the NQS with a broader array of community-generated signals, including systematic error patterns, high-quality corrections, and diverse forms of public engagement, under continuous scrutiny and multistakeholder governance.

Third, while LLMs proved reliable as measurement instruments in this study, they remain subject to cultural and distributional biases inherent in their training data (Atari et al., 2023; Jumelet et al., 2024; Mirzadeh et al., 2024; Ma et al., 2025; Petrov et al., 2025). Future research should explore human–AI hybrid raters and multi-model ensembles to reduce modeling artifacts and further triangulate LLM signals with human expertise and diverse external data.

Fourth, the current thesis primarily establishes observational associations. Causal claims about the real-world impact of ranking effects on user exposure, perception, and behavior will require large-scale, platform-level field experiments of the "veil-then-personalize" protocol, potentially paired with public fairness dashboards.

Finally, any benchmark or governance design can be gamed over time. While the proposed governance framework addresses this through open indicators, change logs, and contestability, adversarial responses from malicious actors will require ongoing monitoring, adaptive governance mechanisms, and continuous research into robust, anti-gaming design.

In conclusion, this thesis provides a critical diagnosis of a dual crisis: a human audience driven by affective identity and an AI system susceptible to propaganda's subtle art, and it offers an ethically grounded design solution. By rigorously dissecting trust, auditing AI's limitations, and crafting a principled ranking protocol, it offers a transformative path forward for fostering informational justice and rebuilding trust in news within our complex digital public sphere.

BIBLIOGRAPHY

- Akser, M., & Baybars, B. (2024). Media systems and media capture in Turkey: A case study. *Media and Communication*, 12, 1–17. <https://doi.org/10.17645/mac.7733>
- Algan, E., & Kaptan, Y. (2023). Content is power: Cultural engineering and political control over transnational television. *International Communication Gazette*, 85(5), 325–344. <https://doi.org/10.1177/17480485231152879>
- Almagro, M. (2023). Political polarization: Radicalism and immune beliefs. *Philosophy & Social Criticism*, 49(3), 309–331. <https://doi.org/10.1177/01914537221109918>
- Amershi, S., Weld, D. S., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S. T., Bennett, P. N., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). Guidelines for human-AI interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). <https://doi.org/10.1145/3290605.3300233>
- Anthropic. (2023). *Constitutional AI: An expanded overview of Anthropic’s alignment approach*. <https://www.anthropic.com/transparency>

Anthropic. (2024, June 21). *Introducing Claude 3.5 Sonnet*. <https://www.anthropic.com/news/claude-3-5-sonnet>

Anthropic. (2025a). *Vision: Work with images in Claude 3/4*. <https://docs.anthropic.com/en/docs/build-with-claude/vision>

Anthropic. (2025b). *Multilingual support*. <https://docs.anthropic.com/en/docs/build-with-claude/multilingual-support>

Anthropic. (2025c). *Models overview*. <https://docs.anthropic.com/en/docs/about-claude/models/overview>

Atari, M., Xue, M. J., Park, P. S., Blasi, D. E., & Henrich, J. (2023, September 22). *Which humans?* OSF. <https://doi.org/10.31234/osf.io/5b26t>

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran, D., Hatfield-Dodds, Z., ... Kaplan, J. (2022). *Constitutional AI: Harmlessness from AI feedback*. arXiv. <https://arxiv.org/abs/2212.08073>

Bail, C. A., Argyle, L. P., Brown, T. W., Bumpus, J. P., Chen, H., Hunzaker, M. B. F., Lee, J., Mann, M., Merhout, F., & Volfovsky, A. (2018). Exposure to opposing views on social media can increase political polarization. *Proceedings of the National Academy of Sciences of the United States of America*, 115(37), 9216–9221. <https://doi.org/10.1073/pnas.1804840115>

- Bakshy, E., Messing, S., & Adamic, L. A. (2015). Exposure to ideologically diverse news and opinion on Facebook. *Science*, 348(6239), 1130–1132. <https://doi.org/10.1126/science.aaa1160>
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>
- Bozdağ, Ç., & Koçer, S. (2022). Skeptical inertia in the face of polarization: News consumption and misinformation in Turkey. *Media and Communication*, 10(2), 241–251. <https://doi.org/10.17645/mac.v10i2.5057>
- Bulut, E., & Ertuna, C. (2022). The pandemic shock doctrine in an authoritarian context: The economic, bodily, and political precarity of Turkey’s journalists during the pandemic. *Media, Culture & Society*, 44(5), 1003–1020. <https://doi.org/10.1177/01634437221084108>
- Cao, Y., Li, S., Liu, Y., Yan, Z., Dai, Y., Yu, P., & Sun, L. (2024). A survey of AI-generated content (AIGC). *ACM Computing Surveys*, 57(1), 1–38. <https://doi.org/10.1145/3704262>
- Cao, Y., Li, S., Yan, Z., Dai, Y., Yu, P., & Sun, L. (2023). *A comprehensive survey of AI-generated content (AIGC): A history of generative AI from GAN to ChatGPT*. arXiv. <https://doi.org/10.48550/arXiv.2303.04226>

- Carlson, M. (2017). *Journalistic authority: Legitimizing news in the digital era*. Columbia University Press.
- Carrera, G., Manlulu, M., & Vargas, D. (2021). *Catalyst and transparency: An analysis of the campus paper function as a watchdog*. Social Science Research Network. <https://doi.org/10.2139/SSRN.3794716>
- Charmaz, K. (2008). Grounded theory as an emergent method. In S. N. Hesse-Biber & P. Leavy (Eds.), *Handbook of emergent methods* (pp. 155–172). Guilford Press.
- Charmaz, K. (2017). The power of constructivist grounded theory for critical inquiry. *Qualitative Inquiry*, 23(1), 34–45. <https://doi.org/10.1177/1077800416657105>
- Chaudhari, S., Aggarwal, P., Murahari, V., & Singh, M. (2024). *RLHF deciphered: A critical analysis of reinforcement learning from human feedback for LLMs*. arXiv. <https://arxiv.org/abs/2402.04207>
- Chuang, Y.-S., Nirunwiroj, K., Studdiford, Z., Goyal, A., Frigo, V. V., Yang, S., Shah, D. V., Hu, J., & Rogers, T. T. (2024). Beyond demographics: Aligning role-playing LLM-based agents using human belief networks. In *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 14010–14026). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.findings-emnlp.819>
- Collopy, P. D., & Hollingsworth, P. M. (2011). Value-driven design. *Journal of Aircraft*, 48(3), 749–759. <https://doi.org/10.2514/1.C000311>

- Doğu, B. (2020). Turkey's news media landscape in Twitter: Mapping interconnections among diversity. *Journalism*, 21(5), 688–706. <https://doi.org/10.1177/1464884917713791>
- Dor, D. (2003). On newspaper headlines as relevance optimizers. *Journal of Pragmatics*, 35(5), 695–721. [https://doi.org/10.1016/S0378-2166\(02\)00134-0](https://doi.org/10.1016/S0378-2166(02)00134-0)
- Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning*. arXiv. <https://arxiv.org/abs/1702.08608>
- Dragomir, M., & Soderstrom, A. (2021). *The state of state media: A global analysis of the editorial independence of state media and an introduction of a new state media typology*. CEU Democracy Institute. <https://cmds.ceu.edu/sites/cmcs.ceu.hu/files/attachment/article/2091/thestateofstatemedia.pdf>
- Eberwein, T., Fengler, S., & Karmasin, M. (Eds.). (2018). *The European handbook of media accountability*. Routledge.
- Ecker, U. K. H., Lewandowsky, S., Chang, E. P., & Pillai, R. (2014). The effects of subtle misinformation in news headlines. *Journal of Experimental Psychology: Applied*, 20(4), 323–335. <https://doi.org/10.1037/xap0000028>
- Edelman. (2022). *2022 Edelman trust barometer: Global report*. <https://www.edelman.com/trust/2022-trust-barometer>
- Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International*

Journal for Educational Integrity, 19(1), 16. <https://doi.org/10.1007/s40979-023-00140-5>

Eslami, M., Rickman, A., Vaccaro, K., Aleyasen, A., Vuong, A., Karahalios, K., Hamilton, K., & Sandvig, C. (2015). “I always assumed that I wasn’t really that close to [her]”: Reasoning about invisible algorithms in the news feed. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 153–162). ACM. <https://doi.org/10.1145/2702123.2702556>

Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature*, 632(8025), 577–583. <https://doi.org/10.1038/s41586-024-07635-z>

Fawzi, N., Steindl, N., Obermaier, M., Prochazka, F., Arlt, D., Blobaum, B., Dohle, M., Engelke, K. M., Hanitzsch, T., Jakob, N., Jakobs, I., Klawier, S., Reinemann, C., Schweiger, W., & Ziegele, M. (2021). Concepts, causes and consequences of trust in news media: A literature review and framework. *Annals of the International Communication Association*, 45(2), 154–174. <https://doi.org/10.1080/23808985.2021.1960181>

Flanagin, A. J. (2020). Source credibility. In J. Van den Bulck (Ed.), *The international encyclopedia of media psychology*. Wiley. <https://doi.org/10.1002/9781119011071.iemp0289>

Foster, A. (2023, March 6). Turkey earthquake: Survivors living in fear on streets. *BBC News*. <https://www.bbc.com/news/world-europe-64824516>

Fox, J., & Weisberg, S. (2020). *car: Companion to applied regression* (Version 3.0-10) [R package]. <https://cran.r-project.org/package=car>

Friedman, B., & Hendry, D. G. (2019). *Value sensitive design: Shaping technology with moral imagination*. The MIT Press.

Friedman, B., Kahn, P. H., Jr., & Borning, A. (2013). Value sensitive design and information systems. In N. Doorn, D. Schuurbiers, I. van de Poel, & M. E. Gorman (Eds.), *Early engagement and new technologies: Opening up the laboratory* (pp. 55–95). Springer.

Gabielkov, M., Ramachandran, A., Chaintreau, A., & Legout, A. (2016). Social clicks: What and who gets read on Twitter? In *Proceedings of the 2016 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems* (pp. 179–192). <https://doi.org/10.1145/2896377.2901462>

Garcia de Torres, E., & Hermida, A. (2017). The social reporter in action: An analysis of the practice and discourse of Andy Carvin. *Journalism Practice*, 11(2–3), 177–194. <https://doi.org/10.1080/17512786.2016.1165158>

Garimella, V. R. K., Smith, T. M., Weiss, R., & West, R. (2021). Political polarization in online news consumption. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1), 136–147. <https://doi.org/10.1609/icwsm.v15i1.18049>

Graves, L. (2016). *Deciding what's true: The rise of political fact-checking in American journalism*. Columbia University Press.

- Gunther, A. C. (1992). Biased press or biased public? Attitudes toward media coverage of social groups. *Public Opinion Quarterly*, 56(2), 147–167. <https://doi.org/10.1086/269309>
- Hagendorff, T. (2024). Deception abilities emerged in large language models. *Proceedings of the National Academy of Sciences*, 121(9), e2318550121. <https://doi.org/10.1073/pnas.2318550121>
- Halk TV. (2023, January 16). *Cafer Mahiroğlu sessizliğini bozdu: Nevzat Çiçek'in sorularını yanıtladı* [Cafer Mahiroğlu broke his silence: Answered Nevzat Çiçek's questions]. <https://halktv.com.tr/gundem/cafer-mahiroglu-sessizligini-bozdu-nevzat-cicekin-sorularini-yanitladi-692186h>
- Hallin, D. C., & Mancini, P. (2004). *Comparing media systems: Three models of media and politics*. Cambridge University Press.
- Hanitzsch, T., Hanusch, F., Ramaprasad, J., & de Beer, A. S. (Eds.). (2019). *Worlds of journalism: Journalistic cultures around the globe*. Columbia University Press.
- Hanitzsch, T., Van Dalen, A., & Steindl, N. (2018). Trust in journalism: A comparative survey of journalists' perceptions. *Journalism Studies*, 19(3), 301–320. <https://doi.org/10.1080/1461670X.2016.1165213>
- Hart, W., Albarracín, D., Eagly, A. H., Brechan, I., Lindberg, M. J., & Merrill, L. (2009). Feeling validated versus being correct: A meta-analysis of selective exposure to information. *Psychological Bulletin*, 135(4), 555–588. <https://doi.org/10.1037/a0015701>

- Heidari, H., Ferrari, C., Gummadi, K. P., & Krause, A. (2018). Fairness behind a veil of ignorance: A welfare analysis for automated decision making. *Advances in Neural Information Processing Systems*, 31. <https://proceedings.neurips.cc/paper/2018/hash/be3159ad04564bfb90db9e32851ebf9c-Abstract.html>
- Hertwig, R., Herzog, S. M., & Kozyreva, A. (2024). Blinding to circumvent human biases: Deliberate ignorance in humans, institutions, and machines. *Perspectives on Psychological Science*, 19(5), 849–859. <https://doi.org/10.1177/17456916231188052>
- Hmielowski, J. D., Hutchens, M. J., & Beam, M. A. (2020). Asymmetry of partisan media effects?: Examining the reinforcing process of conservative and liberal media with political beliefs. *Political Communication*, 37(6), 852–868. <https://doi.org/10.1080/10584609.2020.1763525>
- Hovland, C. I., & Weiss, W. (1951). The influence of source credibility on communication effectiveness. *Public Opinion Quarterly*, 15(4), 635–650. <https://doi.org/10.1086/266350>
- Hovland, C. I., Janis, I. L., & Kelley, H. H. (1953). *Communication and persuasion*. Yale University Press.
- Iggers, J. (1998). *Good news, bad news: Journalism ethics and the public interest*. Westview Press.

Iyengar, S., & Hahn, K. S. (2009). Red media, blue media: Evidence of ideological selectivity in media use. *Journal of Communication*, 59(1), 19–39. <https://doi.org/10.1111/j.1460-2466.2008.01402.x>

Jenke, L. (2023). Affective polarization and misinformation belief. *Political Behavior*, 45(4), 1365–1384. <https://doi.org/10.1007/s11109-022-09851-w>

Journo. (2022). *Haber tüketicisi araştırması* [News consumer survey]. <https://journo.com.tr/haber-tuketicisi-arastirmasi>

Journo. (2024). *Haber gözcüsü* [News monitor report]. <https://journo.com.tr/journo-haber-gozcusu-sonuclar>

Journo. (2025). *Haber etkileşim araştırması* [News engagement survey]. <https://journo.com.tr/haber-etkilesim-arastirmasi>

Jumelet, J., Bylinina, L., Zuidema, W., & Szymanik, J. (2024). *Black big boxes: Do language models hide a theory of adjective order?* arXiv. <https://arxiv.org/abs/2407.02136>

Jurkowitz, M., Mitchell, A., Shearer, E., & Walker, M. (2020, January 24). *U.S. media polarization and the 2020 election: A nation divided*. Pew Research Center. <https://www.pewresearch.org/journalism/2020/01/24/u-s-media-polarization-and-the-2020-election-a-nation-divided/>

Kahn, B. K., Strong, D. M., & Wang, R. Y. (2002). Information quality benchmarks: Product and service performance. *Communications of the ACM*, 45(4), 184–192. <https://doi.org/10.1145/505248.505249>

- Kalogeropoulos, A., Suiter, J., Udris, L., & Eisenegger, M. (2019). News media trust and news consumption: Factors related to trust in news in 35 countries. *International Journal of Communication*, 13, 3173–3196.
- Kant, I. (1993). *Grounding for the metaphysics of morals* (J. W. Ellington, Trans.). Hackett Publishing. (Original work published 1785)
- Kant, I. (1997). *Groundwork of the metaphysics of morals* (M. Gregor, Trans.). Cambridge University Press. (Original work published 1785)
- Kasirzadeh, A., & Gabriel, I. (2025). *Characterizing AI agents for alignment and governance*. arXiv. <https://arxiv.org/abs/2504.21848>
- Kaya, R., & Çakmur, B. (2010). Politics and the mass media in Turkey. *Turkish Studies*, 11(4), 521–537. <https://doi.org/10.1080/14683849.2010.540112>
- Kırdemir, B. (2020). *Turkey's digital news landscape: Polarization, social media, and emerging trends*. The Economic and Foreign Policy Research Centre. <https://edam.org.tr/wp-content/uploads/2020/06/Turkeys-Digital-News-Landscape-Baris-Kirdemir-with-discussion-paper.pdf>
- Kim, Y. (2017). Knowledge versus beliefs: How knowledge and beliefs mediate the influence of likeminded media use on political polarization and participation. *Journal of Broadcasting & Electronic Media*, 61(4), 658–681. <https://doi.org/10.1080/08838151.2017.1375497>

- Kiousis, S. (2001). Public trust or mistrust? Perceptions of media credibility in the information age. *Mass Communication & Society*, 4(4), 381–403. https://doi.org/10.1207/S15327825MCS0404_4
- Kohring, M., & Matthes, J. (2007). Trust in news media: Development and validation of a multidimensional scale. *Communication Research*, 34(2), 231–252. <https://doi.org/10.1177/0093650206298071>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Korthagen, I.-M., & Klijn, E.-H. (2014). The mediatization of network governance: The impact of commercialized news and mediatized politics on trust and perceived network performance. *Public Administration*, 92(4), 1054–1074. <https://doi.org/10.1111/padm.12092>
- Kovach, B., & Rosenstiel, T. (2007). *The elements of journalism: What newspeople should know and the public should expect* (Rev. ed.). Three Rivers Press.
- Kovach, B., & Rosenstiel, T. (2014). *The elements of journalism: What newspeople should know and the public should expect* (3rd ed.). Three Rivers Press.
- Kubin, E., & von Sikorski, C. (2021). The role of (social) media in political polarization: A systematic review. *Annals of the International Communication Association*, 45(3), 188–206. <https://doi.org/10.1080/23808985.2021.1976070>

- Ladd, J. M. (2012). *Why Americans hate the media and how it matters*. Princeton University Press.
- Ladendorf, M., & Hanitzsch, T. (2021). Trust in journalism: A review of conceptual approaches, measures, and empirical findings. *Journalism*, 22(3), 635–650. <https://doi.org/10.1177/1464884920958564>
- Lee, L., Dabirian, A., McCarthy, I. P., & Kietzmann, J. (2020). Making sense of text: Artificial intelligence-enabled content analysis. *European Journal of Marketing*, 54(4), 615–644. <https://doi.org/10.1108/EJM-02-2019-0219>
- Lee, T. T. (2010). Why they don't trust the media: An examination of factors predicting trust. *American Behavioral Scientist*, 54(1), 8–21. <https://doi.org/10.1177/0002764210376308>
- Lee, Y. W., Strong, D. M., Kahn, B. K., & Wang, R. Y. (2002). AIMQ: A methodology for information quality assessment. *Information & Management*, 40(2), 133–146. [https://doi.org/10.1016/S0378-7206\(02\)00043-5](https://doi.org/10.1016/S0378-7206(02)00043-5)
- Leng, Y. (2024). Can LLMs mimic human-like mental accounting and behavioral biases? In *Proceedings of the 25th ACM Conference on Economics and Computation* (pp. 583–584). <https://doi.org/10.1145/3655188.3662580>
- Levendusky, M., & Malhotra, N. (2016). Does media coverage of partisan polarization affect political attitudes? *Political Communication*, 33(2), 283–301. <https://doi.org/10.1080/10584609.2015.1038455>

- Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, 111(3), 831–870. <https://doi.org/10.1257/aer.20191777>
- Ling, R. (2020). Confirmation bias in the era of mobile news consumption: The social and psychological dimensions. *Digital Journalism*, 8(5), 596–614. <https://doi.org/10.1080/21670811.2020.1766987>
- Liu, R., Geng, J., Peterson, J. C., & Griffiths, T. L. (2024). *Large language models assume people are more rational than they really are*. arXiv. <https://arxiv.org/abs/2402.13988>
- Livio, O., & Cohen, J. (2018). ‘Fool me once, shame on you’: Direct personal experience and media trust. *Journalism*, 19(5), 684–698. <https://doi.org/10.1177/1464884916641884>
- Ma, W., He, J., Snell, C., Griggs, T., Min, S., & Zaharia, M. (2025). *Reasoning models can be effective without thinking*. arXiv. <https://arxiv.org/abs/2504.09858>
- Mahdavi, H., Hashemi, A., Daliri, M., Mohammadipour, P., Farhadi, A., Malek, S., & Honavar, V. (2025). *Brains vs. bytes: Evaluating LLM proficiency in olympiad mathematics*. arXiv. <https://arxiv.org/abs/2504.01995>
- Marcus, G. E., Neuman, W. R., & MacKuen, M. (2000). *Affective intelligence and political judgment*. University of Chicago Press.
- McIntyre, K., & Sobel Cohen, M. (2021). Public trust in state-run news media in Rwanda. *Journalism & Mass Communication Quarterly*, 98(3), 808–827. <https://doi.org/10.1177/1077699020959082>

- Metin, Ö. F., & Ramaciotti Morales, P. (2022). Tweeting apart: Democratic backsliding, new party cleavage and changing media ownership in Turkey. *Party Politics*, 13540688221135019. <https://doi.org/10.1177/13540688221135019>
- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267, 1–38. <https://doi.org/10.1016/j.artint.2018.07.007>
- Mirzadeh, S. I., Alizadeh, K., Shahrokhi, H., Tuzel, O., Bengio, S., & Farajtabar, M. (2024). *Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models*. arXiv. <https://arxiv.org/abs/2410.05229>
- Napoli, P. M. (2011). *Audience evolution: New technologies and the transformation of media audiences*. Columbia University Press.
- Napoli, P. M. (2015). Social media and the public interest: Governance of news platforms in the realm of individual and algorithmic gatekeepers. *Telecommunications Policy*, 39(9), 751–760. <https://doi.org/10.1016/j.telpol.2014.12.008>
- Newman, N., Fletcher, R., Nielsen, R. K., Ross Arguedas, A., & Robertson, C. T. (2025). *Digital news report 2025*. Reuters Institute for the Study of Journalism, University of Oxford. <https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2025>
- Newman, N., Fletcher, R., Schulz, A., Andi, S., & Nielsen, R. K. (2020). *Reuters Institute digital news report 2020*. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2020-06/DNR_2020_FINAL.pdf

- Newman, N., Fletcher, R., Schulz, A., Andi, S., & Nielsen, R. K. (2021). *Reuters Institute digital news report 2021*. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2021-06/Digital_News_Report_2021_FINAL.pdf
- Newman, N., Fletcher, R., Schulz, A., Andi, S., & Nielsen, R. K. (2022). *Reuters Institute digital news report 2022*. Reuters Institute for the Study of Journalism. https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2022-06/Digital_News_Report_2022.pdf
- Overgaard, C. B. (2024). Perceiving affective polarization in the United States: How social media shape meta-perceptions and affective polarization. *Social Media + Society*, 10(1). <https://doi.org/10.1177/20563051241232662>
- Park, C. W., & Lessig, V. P. (1981). Familiarity and its impact on consumer decision biases and heuristics. *Journal of Consumer Research*, 8(2), 223–230. <https://doi.org/10.1086/208859>
- Pavlović, M. (2024). The effectiveness of LLMs as annotators: A comparative overview. *NLP Perspectives*, 1, 102–116. <https://aclanthology.org/2024.nlpperspectives-1.11.pdf>
- Petrov, I., Dekoninck, J., Baltadzhiev, L., Drencheva, M., Minchev, K., Balunović, M., & Vechev, M. (2025). *Proof or bluff? Evaluating LLMs on 2025 USA Math Olympiad*. arXiv. <https://arxiv.org/abs/2503.21934>

- Popoola, O., Sigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1), 22. <https://doi.org/10.1007/s40979-023-00146-z>
- ProPublica. (2022, October 29). *How Google's ad business funds disinformation around the world*. <https://www.propublica.org/article/google-alphabet-ads-fund-disinformation-covid-elections>
- Rawls, J. (1971). *A theory of justice*. The Belknap Press of Harvard University Press.
- Rawls, J. (1999). *A theory of justice* (Rev. ed.). The Belknap Press of Harvard University Press.
- Reporters Without Borders. (2024, May 25). *Media ownership monitor Turkey*. Global Media Registry & Bianet. <https://turkey.mom-rsf.org/tr/>
- Reporters Without Borders. (n.d.). *Media ownership monitor: Company profiles*. Retrieved February 17, 2023, from <https://turkey.mom-rsf.org/en/owners/>
- Reuters Institute for the Study of Journalism. (2025, June 17). *Turkey*. University of Oxford. <https://reutersinstitute.politics.ox.ac.uk/digital-news-report/2025/turkey>
- Revelle, W. (2019). *psych: Procedures for psychological, psychometric, and personality research* (Version 1.9.12.31) [R package]. <https://cran.r-project.org/package=psych>
- Rosseel, Y. (2018). *lavaan: Latent variable analysis* (Version 0.6-3) [R package]. <https://cran.r-project.org/package=lavaan>
- Sandel, M. J. (2009). *Justice: What's the right thing to do?* Farrar, Straus and Giroux.

- Saqib, M., & Zia, S. (2024). Evaluation of AI content generation tools for verification of academic integrity in higher education. *Journal of Applied Research in Higher Education*. Advance online publication. <https://doi.org/10.1108/JARHE-10-2023-0470>
- Sayari, S. (2014). Interdisciplinary approaches to political clientelism and patronage in Turkey. *Turkish Studies*, 15(4), 655–670. <https://doi.org/10.1080/14683849.2014.985809>
- Sejnowski, T. J. (2022). Large language models and the reverse Turing test. *Neural Computation*, 34(7), 1433–1436. https://doi.org/10.1162/neco_a_01503
- Shiffrin, R. M., & Mitchell, M. (2023). Probing the psychology of AI models. *Proceedings of the National Academy of Sciences*, 120(49), e2307843120. <https://doi.org/10.1073/pnas.2307843120>
- Shoemaker, P. J., & Vos, T. P. (2009). *Gatekeeping theory*. Routledge.
- Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. <https://doi.org/10.1037/0033-2909.86.2.420>
- Sohail, A., & Zhang, L. (2025). Using large language models to facilitate academic work in the psychological sciences. *Current Psychology*. Advance online publication. <https://doi.org/10.1007/s12144-024-06198-8>
- Sreedhar, K., & Chilton, L. B. (n.d.). *Simulating human strategic behavior: Comparing single and multi-agent LLMs*. arXiv. <https://arxiv.org/abs/2402.19323>

- Stamm, K. R., & Dube, R. (1994). The relationship of attitudinal components to trust in media. *Communication Research*, 21(1), 105–123. <https://doi.org/10.1177/009365094021001005>
- Steensen, S. (2009). What’s stopping them? Towards a grounded theory of innovation in online journalism. *Journalism Studies*, 10(6), 821–836. <https://doi.org/10.1080/14616700902975089>
- Stroud, N. J. (2010). Polarization and partisan selective exposure. *Journal of Communication*, 60(3), 556–576. <https://doi.org/10.1111/j.1460-2466.2010.01497.x>
- Strömbäck, J., Tsifti, Y., Boomgaarden, H. G., Damstra, A., Lindgren, E., Vliegenthart, R., & Lindholm, T. (2020). News media trust and its impact on media use: Toward a framework for future research. *Annals of the International Communication Association*, 44(2), 139–156. <https://doi.org/10.1080/23808985.2020.1755387>
- Sun, Z., Zhang, S., Deng, Y., & Wang, Y. (2024). *LLMs-as-judges: A comprehensive survey on LLM-based evaluation*. arXiv. <https://arxiv.org/abs/2411.15594>
- Tarrasch, S. (2012). *The game of chess*. Dover Publications. (Original work published 1935)
- Thorne, J., & Vlachos, A. (2018). Automated fact checking: Task formulations, methods and future directions. In *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 3346–3358). <https://aclanthology.org/C18-1284/>
- Thorne, J., Vlachos, A., Christodoulopoulos, C., & Mittal, A. (2018). FEVER: A large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference*

- of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)* (pp. 809–819). <https://aclanthology.org/N18-1074/>
- Tiffen, R. (2016). Why are “others” so polarized? Perceived political polarization and media use in 10 countries. *Journal of Computer-Mediated Communication*, 21(5), 349–367. <https://doi.org/10.1111/jcc4.12166>
- Tsfati, Y. (2010). Online news exposure and trust in the mainstream media: Exploring possible associations. *American Behavioral Scientist*, 54(1), 22–42. <https://doi.org/10.1177/0002764210376309>
- Tsfati, Y., & Ariely, G. (2014). Individual and contextual correlates of trust in media across 44 countries. *Communication Research*, 41(6), 760–782. <https://doi.org/10.1177/0093650213485972>
- Tsfati, Y., & Cappella, J. N. (2003). Do people watch what they do not trust? Exploring the association between news media skepticism and exposure. *Communication Research*, 30(5), 504–529. <https://doi.org/10.1177/0093650203253371>
- Tunç, A. (2014). *Can pomegranates replace penguins? Social media and the rise of citizen journalism in Turkey*. Freedom House.
- Türk, H. B. (2018). ‘Populism as a medium of mass mobilization’: The case of Recep Tayyip Erdoğan. *International Area Studies Review*, 21(2), 150–168. <https://doi.org/10.1177/2233865918766449>

- Van Bavel, J. J., Rathje, S., Harris, E. A., Robertson, C., & Sternisko, A. (2021). How social media shapes polarization. *Trends in Cognitive Sciences*, 25(11), 913–916. <https://doi.org/10.1016/j.tics.2021.07.013>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems 30* (pp. 5998–6008). Curran Associates, Inc. <https://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf>
- Waldman, S. A., & Caliskan, E. (2019). Power, patronage and press freedom: The political economy of Turkey’s media. In M. Puppis, L. d’Haenens, & P. Brooten (Eds.), *Public service broadcasting and media systems in troubled European democracies* (pp. 383–408). Springer.
- Wang, R. Y., & Strong, D. M. (1966). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems*, 12(4), 5–33. <https://doi.org/10.1080/07421222.1996.11518099>
- Wang, Y., Pan, Y., Yan, M., Su, Z., & Luan, T. (2023). A survey on ChatGPT: AI-generated contents, challenges, and solutions. *IEEE Open Journal of the Computer Society*, 4, 280–302. <https://doi.org/10.1109/OJCS.2023.3300321>
- Wang, Y., Zhang, S., Deng, Y., & Sun, Z. (2025). *An empirical study of LLM-as-a-judge: How design choices shape reliability*. arXiv. <https://arxiv.org/abs/2506.13639>

- Wathen, C. N., & Burkell, J. (2002). Believe it or not: Factors influencing credibility on the Web. *Journal of the American Society for Information Science and Technology*, 53(2), 134–144. <https://doi.org/10.1002/asi.10016>
- Weidinger, L., McKee, K. R., Everett, R., Huang, S., Zhu, T. O., Chadwick, M., Summerfield, C., & Gabriel, I. (2023). Using the veil of ignorance to align AI systems with principles of justice. *Proceedings of the National Academy of Sciences of the United States of America*, 120(18), e2213709120. <https://doi.org/10.1073/pnas.2213709120>
- Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, M., Hendricks, L. A., Irving, G., & Gabriel, I. (2022, June). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 214–229). <https://doi.org/10.1145/3531146.3533088>
- Wilson, A. T., Parker, V. A., & Feinberg, M. (2020). Polarization in the contemporary political and media landscape. *Current Opinion in Behavioral Sciences*, 34, 223–228. <https://doi.org/10.1016/j.cobeha.2020.07.009>
- Wu, C., Wu, F., An, M., Huang, J., Huang, Y., & Xie, X. (2019). NRMS: Neural news recommendation with multi-head self-attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 7005–7010). Association for Computational Linguistics. <https://aclanthology.org/D19-1671/>

- Wu, F., Qiao, Y., Chen, J., Wu, C., Qi, T., Lian, J., Liu, D., Xie, X., & Gao, J. (2020). MIND: A large-scale dataset for news recommendation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 3597–3606). Association for Computational Linguistics. <https://aclanthology.org/2020.acl-main.331/>
- Wu, J., Gan, W., Chen, Z., Wan, S., & Lin, H. (2023). *AI-generated content (AIGC): A survey*. arXiv. <https://doi.org/10.48550/arXiv.2304.06632>
- Yeşil, B. (2016). *Media in new Turkey: The origins of an authoritarian neoliberal state*. University of Illinois Press. <https://doi.org/10.5406/illinois/9780252040177.001.0001>
- Zhang, Y., & Gosline, R. (2023). Human favoritism, not AI aversion: People’s perceptions (and bias) toward generative AI, human experts, and human–GAI collaboration in persuasive content generation. *Judgment and Decision Making*, 18, e41. <https://doi.org/10.1017/jdm.2023.37>
- Zuboff, S. (2020). Caveat usor: Surveillance capitalism as epistemic inequality. In K. Werbach (Ed.), *After the digital tornado: Networks, algorithms, humanity* (pp. 146–187). Cambridge University Press. <https://doi.org/10.1017/9781108610028.006>

APPENDICES

To ensure transparency and facilitate the replication of our findings, all datasets, analysis scripts, prompting templates, system messages, API configurations, model parameters, and other supplementary materials (e.g. screenshots of news websites for visual analysis) used in this study have been made publicly available online. The full repository is permanently archived and can be accessed on the Open Science Framework (OSF) (DOI: [10.17605/OSF.IO/RM547](https://doi.org/10.17605/OSF.IO/RM547)).

Appendix A (Study 2)

Appendix A1

Non-Blind Trust Comparison (News Consumers vs. Individual LLMs)

Media Outlet	Human Consumers (Avg.)	Claude-Sonnet	Gemini 2.5 Pro	Mistral Large	Kimi-K2	GPT-5
Halk TV	3.73	3.00	2.38	3.70	3.30	2.90
Fox / Now TV	3.58	3.00	3.83	3.00	3.30	3.92
Sözcü	3.78	2.95	2.35	3.65	3.40	3.00
CNN Türk	3.52	2.95	2.00	3.87	2.20	2.36
A Haber	3.39	1.30	1.00	2.40	1.00	1.00
NTV	3.39	3.00	2.90	3.80	2.83	3.33
Habertürk	3.38	3.00	3.00	3.09	2.82	3.50
TRT	3.34	3.00	1.72	2.94	1.53	2.00
Kanal D	3.34	2.50	2.10	3.00	2.00	2.50
Star TV	3.11	2.00	2.60	3.00	1.70	2.70
ATV	3.27	2.00	1.50	2.70	1.10	1.70
Sabah	3.29	2.25	1.10	2.55	1.00	1.30
Hürriyet	3.60	3.00	2.07	3.18	2.80	2.44
Cumhuriyet	3.49	3.71	3.10	3.89	4.20	4.00

Appendix B (Study 2)

Appendix B1

Blind Quality Comparison: Believability (Experts vs. LLMs)

Media Outlet	Human Experts	Claude-Sonnet	Gemini 2.5 Pro	Mistral Large	Kimi-K2	GPT-5
A Haber	53.18	71.05	66.00	69.25	65.40	70.70
NTV	64.67	76.67	88.00	76.00	79.50	81.07
TRT	65.00	79.04	91.17	76.50	81.83	85.57
Sabah	47.69	65.00	71.67	69.49	64.82	74.51

Appendix B2

Blind Quality Comparison: Objectivity (Experts vs. LLMs)

Media Outlet	Human Experts	Claude-Sonnet	Gemini 2.5 Pro	Mistral Large	Kimi-K2	GPT-5
A Haber	47.27	59.41	44.50	63.50	55.90	60.80
NTV	68.00	79.00	90.00	72.00	80.36	82.93
TRT	65.29	80.77	91.00	76.00	85.47	86.43
Sabah	59.23	62.50	69.23	66.15	66.37	71.72

Appendix B3

Blind Quality Comparison: Accuracy (Experts vs. LLMs)

Media Outlet	Human Experts	Claude-Sonnet	Gemini 2.5 Pro	Mistral Large	Kimi-K2	GPT-5
A Haber	55.45	58.42	63.75	64.50	62.25	63.25
NTV	64.00	69.33	73.33	72.33	75.71	73.13
TRT	69.41	71.92	80.34	69.47	79.10	77.13
Sabah	61.54	57.22	65.79	61.00	58.74	64.59

Appendix C (Study 3)

Appendix C1

Comparison of Blind vs. Brand Correlations with NQS and Jackknife Stability

Metric	Value
$\rho(\text{blind}, \text{NQS})$	0.354
$\rho(\text{brand}, \text{NQS})$	0.676
Fisher z (blind-brand)	-2.211
Fisher p	0.027
Jackknife $\Delta\rho$ mean	-0.322
Jackknife $\Delta\rho$ min	-0.344
Jackknife $\Delta\rho$ max	-0.220
Jackknife N	51

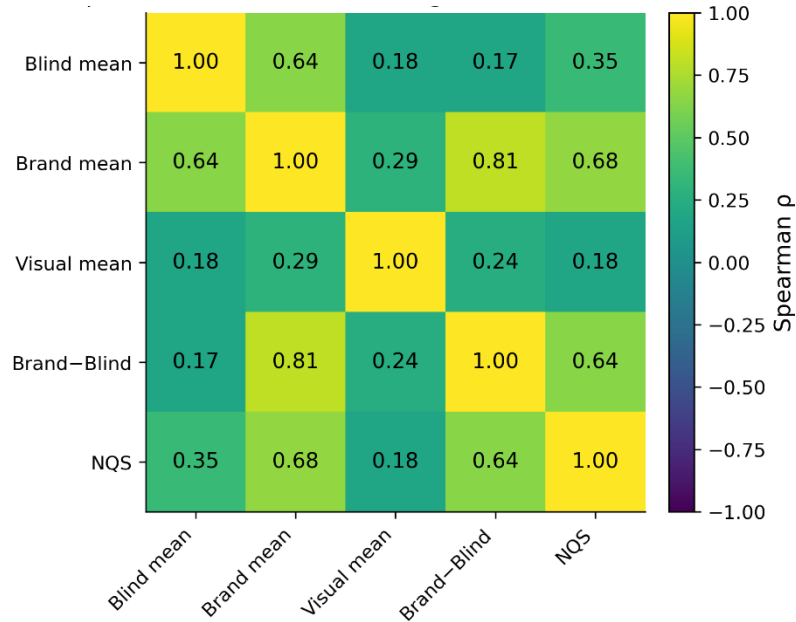
Appendix C2

Coding Framework with Illustrative Quotation IDs for Phase II Interviews

Theme	Code	Definition	Inclusion anchor	Exclusion anchor	Quotation IDs
A. Brand-Triggered Trust	A1. Unsurprisability	Expectation outlet confirms political priors; reduced scrutiny when expectations met.	“Already know what they will say.”	Familiarity without affect.	EX-A1-01; EX-A1-02
	A2. Affective self-fact-checking	Accept/reject based on brand affect, not evidence.	Opposition assumes X lies / supporters trust by default.	Schema talk (A1).	EX-A2-01; EX-A2-02
	A4. Brand schema activation	Logo/name triggers political schema shaping later evaluation.	“When they see the logo...”	Post-hoc rationalizations.	EX-A4-01; EX-A4-02
B. Metrics and Platforms	B1. CTR/dwell pressure	PV/UV/CTR/time-spent as operational imperatives.	“Without PV/UV there is no ads.”	Analytics without pressure.	EX-B1-01
	B2. Algorithmic opacity	Inability to explain rank changes; sudden shifts.	“Black box,” “fell overnight.”	Pure SEO craft talk.	EX-B2-01; EX-B2-02
	B3. Perverse incentives	Clickbait, sensationalism, duplication rewarded.	Must tweak titles for clicks.	Neutral SEO hygiene.	EX-B3-01
C. Repair and Accountability	C1. Corrections	Formal corrections and process disclosures.	On-site corrections policy.	Off-platform notes.	EX-C1-01
	C4. Community feedback	Reader forums and membership loops.	Regular Q&A with members.	One-way promotion.	EX-C4-01
D. AI Adoption	D1. Efficiency uses	Translation, transcription, summarization, drafting support.	Saves time on translation.	Editorial decision automation.	EX-D1-01
	D2. Human-in-the-loop	Explicit assistant-only stance; editor verification.	“No publish without editor pass.”	Fully automated publishing.	EX-D2-01
E. Normative Anchors	E1. Informational justice	Fair access to reliable info; avoiding status harms.	Equity concerns about distribution.	Market fairness only.	EX-E1-01
F. Organizational Conditions	F2. Resource constraints	Staffing shortfalls; tooling gaps; budget precarity.	“One person does three jobs.”	Temporary crunches.	EX-F2-01

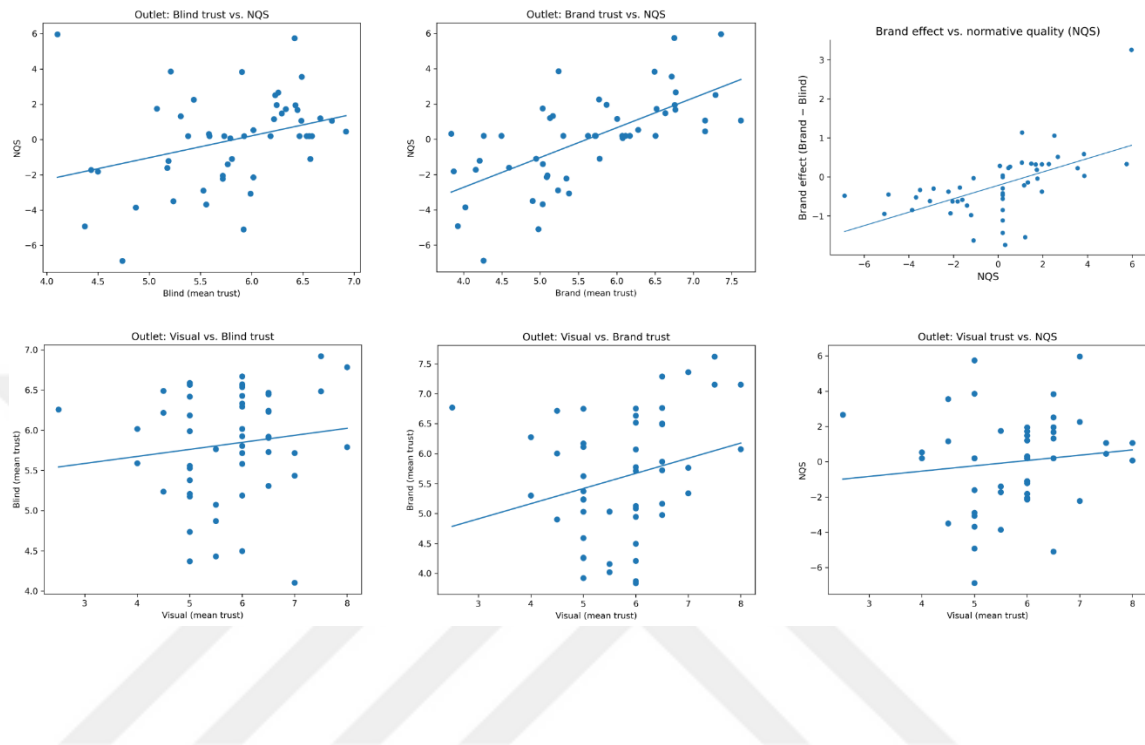
Appendix C3

Spearman Correlations Among Outlet-Level Measures



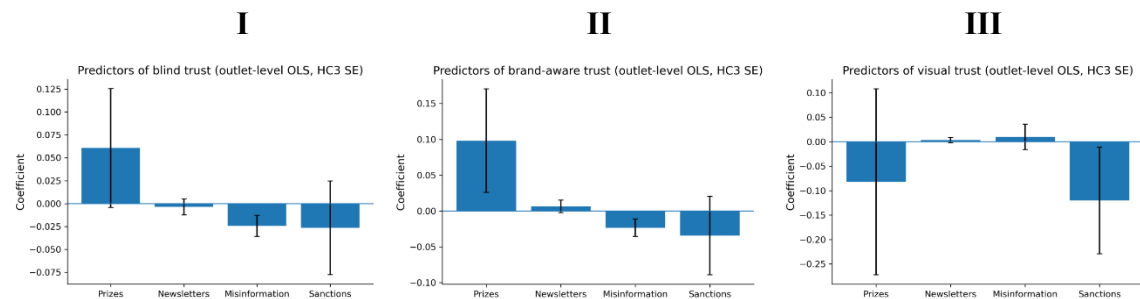
Appendix C4

Mean Trust Effect Comparisons



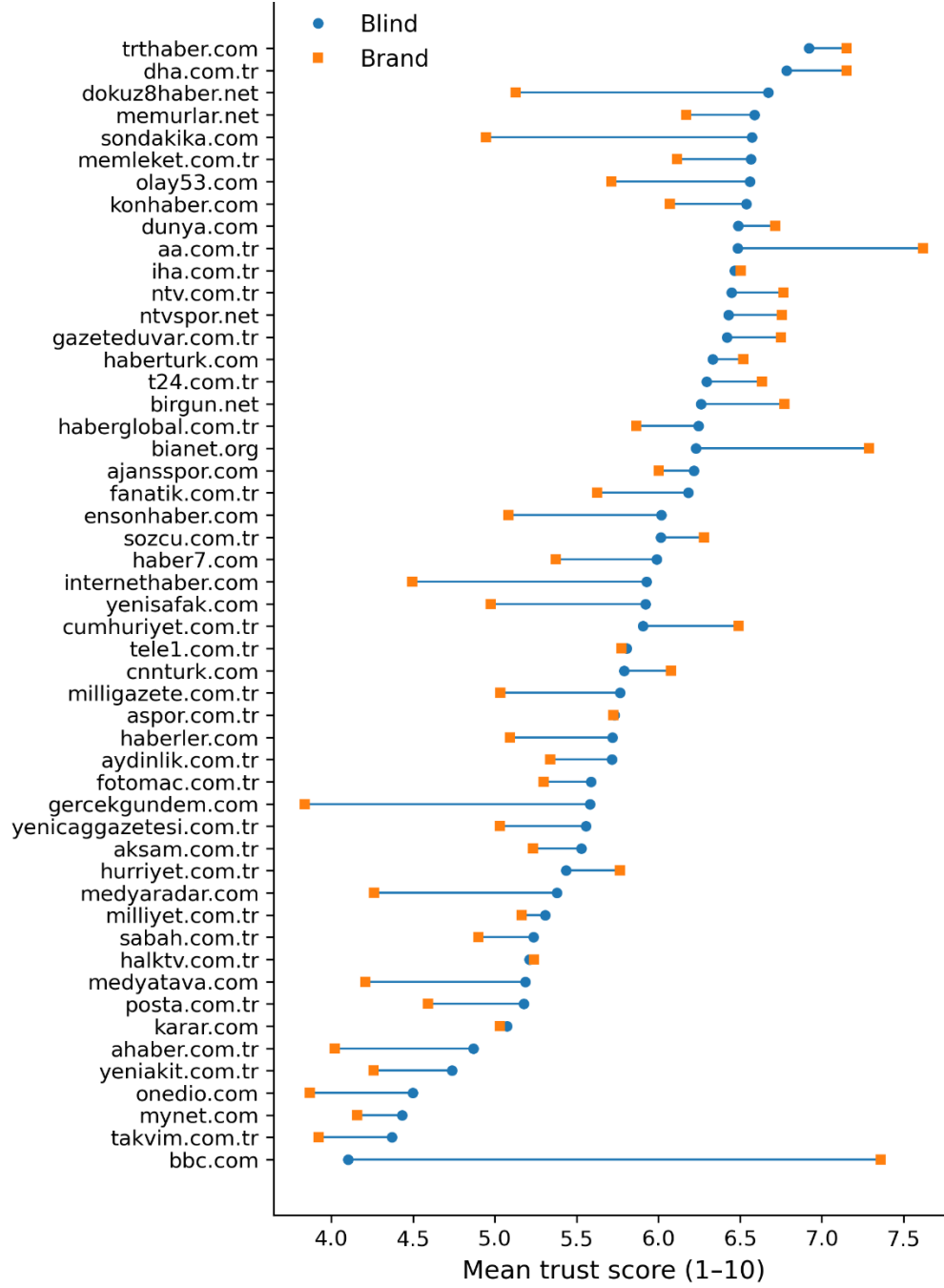
Appendix C5

Predictors of Blind (I), Brand-aware (II), and Visual (III) Trust (Outlet-level OLS, HC3 SE)



Appendix C6

Blind vs. Brand Trust by Media Outlet



TEZ ONAY SAYFASI

Üniversite : T.C. GALATASARAY ÜNİVERSİTESİ
Enstitü : SOSYAL BİLİMLER ENSTİTÜSÜ
Hazırlayanın Adı Soyadı : EMRE KIZILKAYA
Tez Başlığı : The Algorithmic Veil: Affective Trust, AI Blind Spots,
and the Design of Human-Centered News Ranking
Savunma Tarihi : 12 / 09 / 2025
Danışman : Prof. Dr. Kerem Rızvanoğlu

JÜRİ ÜYELERİ:

Unvan, Ad-SOYADI	İmza
Prof. Dr. Kerem Rızvanoğlu	
Prof. Dr. Serhat Güney	
Prof. Dr. Mehmet Özçağlayan	
Dr. Öğr. Üyesi İdil Engindeniz	
Dr. Öğr. Üyesi Onur Varol	

Enstitü Müdür Vekili

Doç. Dr. Demirhan Burak Çelik