The Ethical Dimensions of Data Quality for Automated Fact-Checking

Laurence Dierickx
University of Bergen
Bergen, Norway
l.dierickx@uib.no

Carl-Gustav Lindén
University of Bergen
Bergen, Norway
carl-gustav.linden@uib.no

Andreas Lothe Opdahl
University of Bergen
Bergen, Norway
andreas.opdahl@uib.no

ABSTRACT

Automated fact-checking (AFC) has grown in popularity to address the online spread of misinformation, propaganda, and misinformation about critical contemporary issues. Various natural language processing, machine learning, knowledge representation and database techniques have been used in AFC, whereas, from an end-user perspective, little attention was paid to the quality of the datasets feeding these information systems. Considering the recognised need to blend AI-based tools with journalistic values, this research proposes a practical framework for assessing and improving data quality when developing or implementing AFC systems. Drawing on an interdisciplinary approach, it contributes to understanding how to better align AI-based solutions with ethical standards in journalism and fact-checking.

KEYWORDS

automated fact-checking, datasets, data quality, ethics

1 INTRODUCTION

Automated fact-checking (AFC) attracted growing interest in the wake of the online spreading of misinformation, disinformation, and propaganda on significant issues of our contemporary world, such as the presidential US elections, the COVID-19 pandemic, the global warming crisis, or the Russian-Ukraine war. Since online lies spread faster than the truth [50], automated fact-checking aims to provide practical answers to speed up a time-consuming process when performed manually [38]. AFC can be used for claim identification, evidence retrieval, which consists of finding information beyond the claim, and claim classification [48] [29] [38].

Research explored several tools and techniques based on natural language processing, machine learning, knowledge representation and databases, which play a pivotal role in claim detection and verification [21]. However, the journalistic field or the journalist as end-users were less considered. In a systematic literature review of papers devoted to AFC and published over the last five years, we only found 21 papers out of 267 that considered them. In these works, the focus was mainly on the complementarity between the journalist and the tool. Less attention was paid to the quality of the datasets that feed these systems, especially from an end-user perspective according to the fitness-for-use principle, which relates to data that adapt to the use of their final users. Therefore, this principle goes beyond the sole concerns of accuracy in data [44].

At the same time, there is a recognised need for embedding journalistic values within AI systems to integrate them into journalism workflows better [7] [34] [28]. AFC systems work well when the domains of facts are restricted and on English corpus, but they are not often scalable to real-time content spread on social media and pre-existing fact datasets appear as insufficient [31]. However, they are a means to feed helpfully the systems, insofar as information disorders are not solely agenda-related: for instance, conspiracy theories do not go away once they are debunked [19].

This research aims to question the quality of the data used in AFC systems and to define how to blend datasets with the professional values of their potential end-users. Hence, we have developed a data quality assessment to provide a method to evaluate issues and define the levels to improve when building (or using) datasets in automated fact-checking. This framework is grounded in data science and previous works on data quality in data-driven journalism [13]. From an end-user perspective, it is built on the ethical standards of journalism and fact-checking, to contribute to align AFC system with professional values. Therefore, it can be considered a practical tool to infuse end-users' values in AFC systems.

2 DATA QUALITY AND THE FITNESS-FOR-USE PRINCIPLE

The definition of data quality is protean insofar as it encompasses a set of complementary dimensions which were extended and refined over time. Accuracy was approached as a measure of agreement with an identified source [25], the level of precision and reliability of the data [18], or as the representation of a different real-world state from the one that should have been represented [52]. Scientific literature also refers to the completeness of a given dataset, its consistency (in terms of meeting formal requirements), timeliness and reliability. Considering that defining data quality remains a complex task due to the multidimensionality of the concept, an agreement was found on the fitness-for-use principle, according to which quality data meet explicit or implicit user needs [5]. In other words, data quality refers to data that adapt to their final use, also in terms of relevance and comprehensibility [53].

The rise of big data added extra layers to these concerns, as they challenge the quality dimensions of believability, verifiability and the reputation of the data in the context of data collected online or through sensors [4]. Beyond the correctness of the data, it is also a matter of trusting them [8] [33]. Considering that building trust is essential for adopting a machine learning application [42], all of these considerations are far from trivial in the wake of the growing development of artificial intelligence systems because of their strong dependence on data. Nonetheless, the system's performances also depend on the algorithm at work, which behaviour may also depend on the intrinsic characteristic of the data – especially in terms of volume and completeness [16] [22] [41]. These concerns often remain confined to specialised research areas, and journalistic aspects were little considered. In journalism studies, research on data-driven journalism recognised the structuring role played by computerised databases, which is probably exacerbated by introducing AI technologies in newsrooms. In these fields, the need for high quality data is a prerequisite because if the data are bad or biased, the information will be bad or biased too [1] [6] [15]. Nonetheless, aspects related to data quality have been little addressed, although it was also considered a critical issue [12] [35]. Furthermore, it was also suggested that data selection and evaluation should be journalistic, considering that these tasks are related to a journalistic human expertise, while validation, standardisation and normalisation should be programmers' domain [32].

2.1 Building the Assessment Framework

According to the fitness-for-use principle, data quality assessment is use and context-dependent. It encompasses various strategies, methods and techniques to identify erroneous data and measure their impact on the processes. Its objective is to improve the overall quality of the data [3] [9]. In this research, we defined data quality indicators that fit journalistic and fact-checking ethical values, considering that automated fact-checking systems are likely to be used by journalists and fact-checkers to support or augment their professional practices. Also, we considered that fact-checking activities relate to journalism practices as a distinct sub-genre and a form of accountable journalism [20] [36] [43].

The core ethical standards of journalism are grounded in the social responsibility of journalism, which indistinctly refers to the content of the news, the function of news media in society and the responsibility of news media towards society [2]. Although ethical journalism is first and foremost a matter of practice, it is framed by principles commonly acknowledged: the respect of the truth, which means providing verified facts based on reliable sources; reporting with accuracy; providing well-balanced information with fairness, independence and non-partisanship [23]. Objectivity is another standard promoted in journalism as a constitutive of professional self-perception and identity. However, this concept is regularly criticised as it appears as an ideal, or even a myth, because it relies on the individual subjectivity of the journalist [11] [37] [54]. Choosing a topic, an angle, sources, and the narrative also illustrate the impossibility of objectivity insofar as it implies human and organisational choices [45] [51] [55].

Considering that explaining these choices contributes to increasing the credibility of the news and to (re) building trust with audiences, transparency was presented as an alternative to the disputed concept of objectivity [10] [26] [27]. Transparency means that journalists remain "open and explicit about their processes, methods, limitations and assumptions" [49]: 1507. This concept gained interest in the context of digital environments, seen as a means to open the "black box" of professional practices. In data journalism, for instance, transparency is considered a normative value that contributes to open journalism [40]. Transparency is also at the heart of the guidelines promoted by the international fact-checking organisations – International Fact-Checking Network (IFCN) and European Fact-Checking Standard Network (EFCSN). Practically, their members must be transparent about their organisational structure, funding, partnerships and agreements. They must also be committed to non-partisanship and fairness. Last but not least, fact-checkers must provide their narratives with all the details, methods and sources to allow readers to replicate their work. Much more than a discursive stance, transparency rhymes with professional practices in fact-checking as it is a practical requirement.

Table 1: Assessment of the data quality dimensions

Dimension	Verification
TRUTH	
Accuracy	Level of interoperability, standardisation
	Ratio accurate values/total values (measure of erroneous data)
	Uniqueness (measurement of duplicate entries and redundancies)
	No encoding problems, no information overload
Consistency	Well defined data structure (percentage of data with consistent format and values)
	Homogeneity in the format, structure, and values
	Unambiguous and explicit labelling
Correctness	Identifying abnormal values
	Identifying the causes of NULL values
	Spelling coherence
	Data documented with metadata
	Compliance with metadata
Comprehensibility	The extent to which data are understandable by the end-user
FAIRNESS	
Timeliness	Currentness (percentage of updated data)
Completeness	Appropriate amount of data (ratio missing values/total values - ratio NULL values/total values)
Accessibility	Right to use the data
	Level of retrievability of the data
Objectivity	Unbiased data (size and representativity of the sample)
	Identification of human bias (data and/or annotations)
Relevance	The extent to which the data are relevant for the purpose
	Newsworthiness
	Data scarcity (fraction of data containing relevant information)
Usability	Making sense in a journalistic context
TRANSPARENCY	
Reliability	Authenticity (source)
	Authority and reputation (source, annotators)
Credibility	Degree of believability and expertise (data source, data, annotation process, annotators)
Verifiability	Fact-checking the source, the data and the annotation process

Table 2: Sample of Fact-Checking Datasets

Authors	Description/URL
Alhindi et al., 2021	Multidomain dataset based on 4K+ claim–article pairs from diverse sources. https://github.com/Tariq60/arastance
Arslan et al., 2020	Dataset of 23K+ statements extracted from U.S. general election presidential debates, annotated by human coders. https://zenodo.org/record/3609356
Drchal et al., 2022	Derived from the FEVER dataset, CsFEVER contains 127K+ claims. CTKFacts contains 3K+ claims from a corpus of more than two million Czech News Agency news reports. https://huggingface.co/ctu-aic/
Sepúlveda-Torres et al., 2021	Content 7K+ news items classified as Compatible, Contradiction, or Unrelated. https://zenodo.org/record/4596394
Samarinas et al., 2020	Large-scale dataset based on the FEVER dataset, used for evidence-retrieval, and MSMARCO, a collection of large-scale datasets for deep learning. https://github.com/algoprog/Quin
Shahi and Nandini, 2020	Multilingual cross-domain dataset of 5K+ fact-checked news articles on COVID-19, collected from 04/01/2020 to 01/07/2020. https://gautamshahi.github.io/FakeCovid/
Kotonya and Toni, 2020	Dataset based on 11,8K claims collected from 5 fact-checking websites. https://github.com/neemakot/Health-Fact-Checking
Sathe et al., 2020	Dataset of 124k+ triples consisting of a claim, context and evidence document extracted from English Wikipedia articles and citations, and 34k+ manually written claims refuted by evidence documents. https://github.com/wikifactcheck-english/wikifactcheck-english/
Gupta and Srikumar, 2021	Multilingual dataset for factual verification of naturally existing real-world claims composed of 38K+ short statements. https://github.com/utahnlp/x-fact/

2.2 Method

Data quality assessments usually consists of defining data quality indicators and providing tools for measurement [18]. However, data quality also depends on the design and production processes at work to generate the data [52]. Also, in a data quality assessment, subjective considerations intertwine with objective ones, insofar as it reflects human needs, experiences and contexts of [39].

The framework to assess data quality for automated fact-checking is built upon three core ethical principles in journalism and fact-checking (Table 1): the principle of "truth" relates to the data quality dimensions of accuracy, consistency, correctness and comprehensibility; the principle of "fairness" encompasses the dimensions of timeliness, completeness, accessibility, objectivity, relevance and usability; the principle of "transparency", as a lever for trust, is related to the reliability, credibility and verifiability of the data. This three-level segmentation assumes that telling the truth involves the knowledge of the application domain the data refers to, that being fair refers to unbiased and well-balanced information, and that transparency gathers the means for remaining trustworthy.

An extensive literature review of papers, pre-prints and proceedings published between 2020 and 2022 allowed us to identify a sample of nine datasets developed for automated fact-checking, which are publicly available (Table 2). This sample only included textual data because a corpus of images involves other types of considerations related to the intrinsic characteristics of images in terms of blur, noise, contrast, format and compression [14]. Nonetheless, data quality challenges also encompass the diversity of datasets, and the quality of annotations [57]. Google Refine was used as a data quality tool for data profiling to identify the overall data quality challenges from formal and empirical perspectives [30]. Due to the vast amount of data to assess, we considered the Pareto principle relevant, as "most of the errors are attributable to just a few variables" [47]: 237.

3 MAIN FINDINGS

The analysis aimed to identify the limitations or issues regarding the ethical principles of truth, fairness and transparency. As the purpose is not to attribute good and bad points to each examined dataset, this analysis adopted a transversal approach.

3.1 Truth

The nine datasets of our corpus have different characteristics in terms of size, domains, languages and format (JSON, CSV, TXT, TSV), which do not seem an obstacle to reusing them. However, cross-domain approaches (e.g., politics, sports, health) appear as the most challenging to deal with, considering the knowledge required to handle each domain well. Four datasets were not documented by metadata or lacked explicit labelling. The use of a sentiment score in one dataset was unclear, as well as the labelling used to assess the validity of a claim. Three datasets contained NULL values, which may have various causes and require human knowledge (e.g. the NULL values are equal to zero, the information exists but is not known or irrelevant to the variable). The overall understandability of the datasets was not always granted because of a lack of documentation, although academic papers documented processes. As they relied on textual data, the question of the standardisation and harmonisation of the language arose, also in multilingual datasets.

3.2 Fairness

In terms of relevancy, the language and context-dependency of the datasets raised the issue of using them in other languages or national contexts. The datasets' usability (and reusability) is also challenged by the dimension of accessibility, as most of the datasets did not have an attached licence. The dimension of timeliness is also problematic for several reasons: missing dates (1 dataset), no mention of the last update (1 dataset), and corpus collected over a limited period (3 datasets). Hence, the currentness of the datasets is not always guaranteed and raised questions about the relevance of their reusability, despite they can be useful to fact-check old propaganda discourses or conspiracy theories. However, the lack of maintenance of the datasets remains an obstacle to meeting the two ethical principles of truth and fairness since information disorders are also a dynamic phenomenon that can vary or change over time, and this also applies to concepts and definitions, considering that the construction of knowledge is an ongoing process. In addition, a cross-domain approach made it difficult to assess the completeness dimension. We also found two datasets with missing values, with a respective proportion of 11.37% and 24.53%. Nevertheless, the completeness of the datasets remained difficult to evaluate, whether for recent or older phenomena, because there is no absolute referral to assess it. As a corollary, the dimension of objectivity appeared problematic when looking at the annotations used for classification purposes: from "True" to "False", "Half-true", "Unproven", "Contradiction", "Compatible" or "Unrelated", there was no consensus among researchers.

3.3 Transparency

The majority of the datasets had no issues related to the source trustworthiness, as they mostly relied on specialised fact-checking and news websites. The pitfalls underlined in previous research were globally avoided, considering that several potential data quality issues will likely appear with open data, user-generated data and data from multiple sources [24]. However, three datasets used Wikipedia as a primary source and raised questions related to their reliability, credibility and verifiability but they also questioned the fairness principle, in terms of objectivity and relevance. In journalism, Wikipedia is taken with caution as the content comes from users of whom nothing is known about their expertise [46]. Also, the Wikipedan - or encyclopaedic - writing style differs from journalistic writing, making it less useful for training. The same applies to social media content used in one dataset, and it is perhaps exacerbated by the unknown and volatile nature of the users. The annotation processes did not appear particularly problematic. Datasets were mostly well documented, except one with no indication on the level of the human expertise for annotations. In this regard, research emphasised that, whether manual or automated, annotations are inherently error-prone and that, when performed manually, human subjective factors should also be considered [17] [22] [41].

4 DISCUSSION AND CONCLUSION

Results showed that adapting data to the ethical values of journalists and fact-checkers does not only mean ensuring the reliability and credibility of the data source as well as the accuracy of the data. One of the main challenges is related to the maintenance over time in regard to the dimension of actuality insofar as information disorders are an ongoing process. However, several examined datasets might be useful for older cases, considering that history might be repeating. Still, the question of maintenance remains critical as domains and concepts evolve over time. Further, fact-checking requires a critical approach toward the source of the data, including annotated data. Datasets based on Wikipedia and on social media raised questions about their fairness and trustworthiness. Acknowledging that the relationship between journalists and AI-driven systems is built on trust, the data that feed these systems should also be trusted.

Despite limitations due to its normative lenses and the sample size, the data quality assessment framework developed in this research aimed to provide clues to improve the overall data quality when using technologies that rely so heavily on large volumes of data. In many ways, the developed approach shares common concerns with computer and data science, such as it is set in the FAIR principles, which propose guidelines for improving the findability, accessibility, interoperability and reuse of digital assets [56]. As end-users of AI-based systems, journalists and fact-checkers are not always aware or informed about the data that feed the systems they use. At the same time, their expertise in the data source's reliability and credibility and their knowledge of the context should not be overlooked. Therefore, better fine-tuning AI-based systems with their end users would strengthen collaborations and favour cross-discipline approaches.

5 FUNDING

The research was funded by EU CEF grant number 2394203.

REFERENCES

[1] C.W. Anderson. 2018. *Apostles of Certainty: Data Journalism and the Politics of Doubt.* Oxford University Press. https://doi.org/10.1093/oso/9780190492335.001.0001
[2] Jo Bardoel and Leen dHaenens. 2004. Media Responsibility and Accountability. New Conceptualizations and Practices. *Communications* 29, 1 (2004). https://doi.org/10.1515/comm.2004.007
[3] Carlo Batini. 2009. Data Quality Assessment. In *Encyclopedia of Database Systems.* Springer US, 608–612. https://doi.org/10.1007/978-0-387-39940-9_107
[4] Carlo Batini, Anisa Rula, Monica Scannapieco, and Gianluigi Viscusi. 2015. From Data Quality to Big Data Quality. *Journal of Database Management* 26, 1 (2015), 60–82. https://doi.org/10.4018/jdm.2015010103
[5] Isabelle Boydens and Seth van Hooland. 2011. Hermeneutics Applied to the Quality of Empirical Databases. *Journal of Documentation* 67, 2 (2011), 279–289. https://doi.org/10.1108/00220411111109476
[6] Paul Bradshaw. 2017. Data journalism. In *The Online Journalism Handbook.* Routledge, 250–280. https://doi.org/10.4324/9781315761428-10
[7] Meredith Broussard, Nicholas Diakopoulos, Andrea L. Guzman, Rediet Abebe, Michel Dupagne, and Ching-Hua Chuan. 2019. Artificial Intelligence and Journalism. *Journalism & Mass Communication Quarterly* 96, 3 (2019), 673–695. https://doi.org/10.1177/1077699019859901
[8] Li Cai and Yangyong Zhu. 2015. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. *Data Science Journal* 14, 0 (2015), 2. https://doi.org/10.5334/dsj-2015-002
[9] Corinna Cichy and Stefan Rass. 2019. An Overview of Data Quality Frameworks. *IEEE Access* 7 (2019), 24634–24648. https://doi.org/10.1109/access.2019.2899751
[10] Stephanie Craft and Tim P. Vos. 2021. The Ethics of Transparency. In *The Routledge Companion to Journalism Ethics.* Routledge, 175–183. https://doi.org/10.4324/9780429262708-24

[11] Mark Deuze. 2005. What Is Journalism?: Professional Identity and Ideology of Journalists Reconsidered. *Journalism* 6, 4 (2005), 442–464. https://doi.org/10.1177/1464884905056815

[12] Nicholas Diakopoulos. 2019. *Automating the News*. Harvard University Press. https://doi.org/10.4159/9780674239302

[13] Laurence Dierickx. 2017. News Bot for the Newsroom: How Building Data Quality Indicators Can Support Journalistic Projects Relying on Real-Time Open Data. In *Global Investigative Journalism Conference 2017 Academic Track*. https://ijec.org/2018/02/02/research-news-bot-for-the-newsroom-how-building-data-quality-indicators-can-support-journalistic-projects-relying-on-real-time-open-data/

[14] Samuel Dodge and Lina Karam. 2016. Understanding How Image Quality Affects Deep Neural Networks. In *2016 8th International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE. https://doi.org/10.1109/qomex.2016.7498955

[15] Konstantin Nicholas Dörr and Katharina Hollnbuchner. 2016. Ethical Challenges of Algorithmic Journalism. *Digital Journalism* 5, 4 (2016), 404–419. https://doi.org/10.1080/21670811.2016.1167612

[16] Lisa Ehrlinger, Verena Haunschmid, Davide Palazzini, and Christian Lettner. 2019. A DaQL to Monitor Data Quality in Machine Learning Applications. In *Lecture Notes in Computer Science*. Springer International Publishing, 227–237. https://doi.org/10.1007/978-3-030-27615-7_17

[17] Harald Foidl and Michael Felderer. 2019. Risk-Based Data Validation in Machine Learning-Based Software Systems. In *Proceedings of the 3rd ACM SIGSOFT International Workshop on Machine Learning Techniques for Software Quality Evaluation*. ACM. https://doi.org/10.1145/3340482.3342743

[18] Christopher Fox, Anany Levitin, and Thomas Redman. 1994. The Notion of Data and its Quality Dimensions. *Information Processing & Management* 30, 1 (1994), 9–19. https://doi.org/10.1016/0306-4573(94)90020-5

[19] Ted Goertzel. 1994. Belief in Conspiracy Theories. *Political Psychology* 15, 4 (1994), 731. https://doi.org/10.2307/3791630

[20] Lucas Graves and CW Anderson. 2020. Discipline and Promote: Building Infrastructure and Managing Algorithms in a "Structured Journalism" Project by Professional Fact-Checking Groups. *New Media & Society* 22, 2 (2020), 342–360. https://doi.org/10.1177/1461444819856916

[21] Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A Survey on Automated Fact-Checking. *Transactions of the Association for Computational Linguistics* 10 (2022), 178–206. https://doi.org/10.1162/tacl_a_00454

[22] Nitin Gupta, Shashank Mujumdar, Hima Patel, Satoshi Masuda, Naveen Panwar, Sambaran Bandyopadhyay, Sameep Mehta, Shanmukha Guttula, Shazia Afzal, Ruhi Sharma Mittal, and Vitobha Munigala. 2021. Data Quality for Machine Learning Tasks. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. ACM. https://doi.org/10.1145/3447548.3470817

[23] Kai Hafez. 2002. Journalism Ethics Revisited: A Comparison of Ethics Codes in Europe, North Africa, the Middle East, and Muslim Asia. *Political Communication* 19, 2 (2002), 225–250. https://doi.org/10.1080/10584600252907461

[24] Joseph F. Hair and Marko Sarstedt. 2021. Data, Measurement, and Causal Inferences in Machine Learning: Opportunities and Challenges for Marketing. *Journal of Marketing Theory and Practice* 29, 1 (2021), 65–77. https://doi.org/10.1080/10696679.2020.1860683

[25] YU Huh, FR Keller, TC Redman, and AR Watkins. 1990. Data Quality. *Information and Software Technology* 32, 8 (1990), 559–565. https://doi.org/10.1016/0950-5849(90)90146-i

[26] Michael Karlsson. 2020. Dispersing the Opacity of Transparency in Journalism on the Appeal of Different Forms of Transparency to the General Public. *Journalism Studies* 21, 13 (2020), 1795–1814. https://doi.org/10.1080/1461670x.2020.1790028

[27] Michael Koliska. 2022. Trust and Journalistic Transparency Online. *Journalism Studies* 23, 12 (2022), 1488–1509. https://doi.org/10.1080/1461670x.2022.2102532

[28] Tomoko Komatsu, Marisela Gutierrez Lopez, Stephann Makri, Colin Porlezza, Glenda Cooper, Andrew MacFarlane, and Sondess Missaoui. 2020. AI Should Embody Our Values: Investigating Journalistic Values to Inform AI Technology Design. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. ACM. https://doi.org/10.1145/3419249.3420105

[29] Lev Konstantinovskiy, Oliver Price, Mevan Babakar, and Arkaitz Zubiaga. 2021. Toward Automated Factchecking: Developing an Annotation Schema and Benchmark for Consistent Automated Claim Detection. *Digital Threats: Research and Practice* 2, 2 (2021), 1–16. https://doi.org/10.1145/3412869

[30] Tien Fabrianti Kusumasari and Fitria. 2016. Data Profiling for Data Quality Improvement With Openrefine. In *2016 International Conference on Information Technology Systems and Innovation (ICITSI)*. IEEE. https://doi.org/10.1109/icitsi.2016.7858197

[31] Eric Lazarski, Mahmood Al-Khassaweneh, and Cynthia Howard. 2021. Using NLP for Fact Checking: A Survey. *Designs* 5, 3 (2021), 42. https://doi.org/10.3390/designs5030042

[32] Carl-Gustav Lindén. 2016. Decades of Automation in the Newsroom. *Digital Journalism* 5, 2 (2016), 123–140. https://doi.org/10.1080/21670811.2016.1160791

[33] Jianzheng Liu, Jie Li, Weifeng Li, and Jiansheng Wu. 2016. Rethinking Big Data: A Review on the Data Quality and Usage Issues. *ISPRS Journal of Photogrammetry and Remote Sensing* 115 (2016), 134–142. https://doi.org/10.1016/j.isprsjprs.2015.11.006

[34] Marisela Gutierrez Lopez, Colin Porlezza, Glenda Cooper, Stephann Makri, Andrew MacFarlane, and Sondess Missaoui. 2022. A Question of Design: Strategies for Embedding AI-Driven Tools into Journalistic Work Routines. *Digital Journalism* (2022), 1–20. https://doi.org/10.1080/21670811.2022.2043759

[35] Wilson Lowrey, Ryan Broussard, and Lindsey A. Sherrill. 2019. Data Journalism and Black-Boxed Data Sets. *Newspaper Research Journal* 40, 1 (2019), 69–82. https://doi.org/10.1177/0739532918814451

[36] Paul Mena. 2018. Principles and Boundaries of Fact-checking: Journalists' Perceptions. *Journalism Practice* 13, 6 (2018), 657–672. https://doi.org/10.1080/17512786.2018.1547655

[37] Juan Ramón Muñoz-Torres. 2012. Truth and Objectivity in Journalism. *Journalism Studies* 13, 4 (2012), 566–582. https://doi.org/10.1080/1461670x.2012.662401

[38] Preslav Nakov, David Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated Fact-Checking for Assisting Human Fact-Checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*. International Joint Conferences on Artificial Intelligence Organization. https://doi.org/10.24963/ijcai.2021/619

[39] Leo L. Pipino, Yang W. Lee, and Richard Y. Wang. 2002. Data Quality Assessment. *Commun. ACM* 45, 4 (2002), 211–218. https://doi.org/10.1145/505248.506010

[40] Colin Porlezza and Sergio Splendore. 2019. From Open Journalism to Closed Data: Data Journalism in Italy. *Digital Journalism* 7, 9 (2019), 1230–1252. https://doi.org/10.1080/21670811.2019.1657778

[41] Fakhitah Ridzuan, Wan Mohd Nazmee Wan Zainon, and Mohd Zairul. 2021. A Thematic Review on Data Quality Challenges and Dimension in the Era of Big Data. In *Lecture Notes in Electrical Engineering*. Springer Singapore, 725–737. https://doi.org/10.1007/978-981-16-2406-3_56

[42] Keng Siau and Weiyu Wang. 2018. Building Trust in Artificial Intelligence, Machine Learning, and Robotics. *Cutter business technology journal* 31, 2 (2018), 47–53. https://www.cutter.com/article/building-trust-artificial-intelligence-machine-learning-and-robotics-498981

[43] Jane B Singer. 2020. Border Patrol: The Rise and Role of Fact-Checkers and Their Challenge to Journalists' Normative Boundaries. *Journalism* 22, 8 (2020), 1929–1946. https://doi.org/10.1177/1464884920933137

[44] Giri Kumar Tayi and Donald P. Ballou. 1998. Examining Data Quality. *Commun. ACM* 41, 2 (1998), 54–57. https://doi.org/10.1145/269012.269021

[45] Jingrong Tong and Landong Zuo. 2019. The Inapplicability of Objectivity: Understanding the Work of Data Journalism. *Journalism Practice* 15, 2 (2019), 153–169. https://doi.org/10.1080/17512786.2019.1698974

[46] Khonzodakhon Umarova and Eni Mustafaraj. 2019. How Partisanship and Perceived Political Bias Affect Wikipedia Entries of News Sources. In *Companion Proceedings of The 2019 World Wide Web Conference*. ACM. https://doi.org/10.1145/3308560.3316760

[47] Richard D. De Veaux and David J. Hand. 2005. How to Lie with Bad Data. *Statist. Sci.* 20, 3 (2005). https://doi.org/10.1214/088342305000000269

[48] Andreas Vlachos and Sebastian Riedel. 2014. Fact Checking: Task definition and dataset construction. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science*. Association for Computational Linguistics. https://doi.org/10.3115/v1/w14-2508

[49] Tim P. Vos and Stephanie Craft. 2016. The Discursive Construction of Journalistic Transparency. *Journalism Studies* 18, 12 (2016), 1505–1522. https://doi.org/10.1080/1461670x.2015.1135754

[50] Soroush Vosoughi, Deb Roy, and Sinan Aral. 2018. The Spread of True and False News Online. *Science* 359, 6380 (2018), 1146–1151. https://doi.org/10.1126/science.aap9559

[51] Karin Wahl-Jorgensen. 2013. Subjectivity and Story-telling in Journalism. *Journalism Studies* 14, 3 (2013), 305–320. https://doi.org/10.1080/1461670x.2012.713738

[52] Yair Wand and Richard Y. Wang. 1996. Anchoring Data Quality Dimensions in Ontological Foundations. *Commun. ACM* 39, 11 (1996), 86–95. https://doi.org/10.1145/240455.240479

[53] Richard Y. Wang and Diane M. Strong. 1996. Beyond Accuracy: What Data Quality Means to Data Consumers. *Journal of Management Information Systems* 12, 4 (1996), 5–33. https://doi.org/10.1080/07421222.1996.11518099

[54] Stephen J. A. Ward. 2019. Journalism Ethics. In *The Handbook of Journalism Studies*. Routledge, 307–323. https://doi.org/10.4324/9781315167497-20

[55] Charlotte Wien. 2005. Defining Objectivity within Journalism. *Nordicom Review* 26, 2 (2005), 3–15. https://doi.org/10.1515/nor-2017-0255

[56] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak ..., and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 1 (2016). https://doi.org/10.1038/sdata.2016.18

[57] Yazhou Yao, Jian Zhang, Fumin Shen, Li Liu, Fan Zhu, Dongxiang Zhang, and Heng Tao Shen. 2020. Towards Automatic Construction of Diverse, High-Quality Image Datasets. *IEEE Transactions on Knowledge and Data Engineering* 32, 6 (2020), 1199–1211. https://doi.org/10.1109/tkde.2019.2903036