Check for updates

Original research article

Misinformation without liars: How invisible data-processing choices manufacture false conclusions in automated text analysis

Statistical Journal of the IAOS
1–7
© The Author(s) 2026
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/18747655261480969
journals.sagepub.com/home/sji
Sage | IOS Press

Nadejda Komendantova[1] and Petr Pesov[1,2]

Abstract

Official statistics increasingly draws on non-traditional sources, such as social-media posts and other web-scraped text, analysed with AI language models for nowcasting, indicator production, and crisis monitoring. We show that such pipelines can manufacture false statistics with no misinformation actor involved: ordinary, invisible data-processing choices are enough. Using standard, publicly available zero-shot classifiers, we give fully reproducible demonstrations on conflict-related messages. A routine "N comments" interface string captured by scrapers can collapse a model's judgement that a message concerns armed conflict from near-certain to below the retention threshold, while a benign sentence of equal length does not; the effect is large on a widely used model yet negligible on others. Four equally defensible phrasings of a single analytical question yield almost disjoint datasets from identical inputs, and swapping the classifier alone moves an unambiguous conflict report from retained to discarded. These distortions are silent and idiosyncratic: pipelines return confident, plausible numbers, and the direction and size of error depend on the specific model, prompt, and preprocessing. We argue that interpretation artifacts are a distinct threat to data quality and public trust, and propose a short, demonstrated audit that statistical agencies can adopt.

Keywords

Official statistics, text-as-data, data quality, misinformation, reproducibility

Received: 7 July 2026; accepted: 4 August 2026

1 Introduction

National statistical offices and international organisations are steadily extending their evidence base beyond surveys and administrative registers. Social-media posts, news comment sections, and other web-scraped text are now used, experimentally and in production, for nowcasting, sentiment and topic indicators, monitoring of the Sustainable Development Goals, and rapid situational awareness during crises.[1,2] Because these sources arrive as unstructured natural language at a volume no human can read, agencies increasingly turn to AI language models to classify, score, and aggregate them. This is the promise of "text as data": a fast, inexpensive new signal.[3]

Discussions of misinformation in this setting almost always centre on bad actors.[4,5] The concern is fabricated content, coordinated inauthentic behaviour, manipulated media, and deceptive sources, and the proposed remedies are detection, verification, and provenance. This paper is about a second, quieter route to false information, one that requires no liar at all. When text is turned into numbers by an automated pipeline, ordinary and well-intentioned processing decisions can silently change what the data appear to say: how the text was collected and cleaned, how the analytical question was phrased, which model was used. The resulting figure is not a lie anyone told; it is an artifact of interpretation. Yet once it is published as a statistic, it can mislead exactly as effectively as deliberate disinformation, and with more authority.

[1]Cooperation and Transformative Governance Group, International Institute for Applied Systems Analysis, Laxenburg, Austria
[2]Department of Public Policy, Tel Aviv University, Tel Aviv, Israel

Corresponding author:
Petr Pesov, Cooperation and Transformative Governance Group, International Institute for Applied Systems Analysis (IIASA), Laxenburg, Austria.
Email: petr@mail.tau.ac.il

We keep the argument concrete and reproducible. Using publicly available models and a handful of short scripts, we demonstrate three "knobs" that an analyst turns without noticing, each able to flip a conclusion: contamination of the input by interface text that scrapers capture; the wording of the classification question; and the choice of model. We then trace how such per-message errors propagate into aggregate indicators, argue that they constitute a data-quality risk that current statistical quality frameworks were not designed to catch, and propose a short "interpretation audit" that agencies can fold into existing quality assurance. To show the audit is practical rather than aspirational, we report applying it to a corpus of 24.4 million machine-scored messages from our own research infrastructure. Every number in this paper can be regenerated by a reader with the supplementary scripts.

2 Materials and reproducible setup

All demonstrations use zero-shot Natural Language Inference (NLI) classifiers, a now-common way to score text against a stated concept without task-specific training data.[6] Given a passage (the premise) and a hypothesis such as "*This text is about armed conflict,"* the model returns the probability that the premise entails the hypothesis, and we take that entailment probability as the relevance score. Our primary model is a multilingual DeBERTa-v3 model fine-tuned for NLI,[7,8] chosen because it is widely used, freely available, and representative of the tools an agency would reach for. For the model comparison in Section 4 we add two further public classifiers of different architecture.

Following common practice, a message is retained as relevant when its score exceeds a fixed threshold and discarded otherwise. We use 0.45 throughout, and report results against a range of thresholds where it matters. In each demonstration we change exactly one element, the input text, the hypothesis wording, or the model, and hold everything else fixed, so that any change in output is attributable to that element alone.

A paper about analytical fragility must itself be reproducible, so we checked. Scores are deterministic: rescoring an identical input returned identical values to ten decimal places. We pin the software environment (Python 3.14.3, PyTorch 2.10.0, Transformers 4.57.6) and the exact model revision hashes, and all inputs, hypotheses, cleaning code, and scoring scripts are provided in the supplementary materials.

3 Demonstration 1: Contamination from the input

When posts are collected in bulk, more than the author's words is captured. Platforms display interface text around each message, and the most common such fragment is the comment counter ("*2 comments,"* and its equivalents in other languages) that sits beneath a post. It is not something the author wrote; it is chrome that a scraper can sweep up alongside the content.

Consider a message that is plainly about armed conflict: "*IDF says it struck Hezbollah weapons depots overnight; several militants reported killed."* Our model scores its conflict-relevance at 0.935, effectively "yes." Append the string \n\n2 comments and the score falls to 0.233, below the 0.45 retention threshold, even though the author's words are unchanged. Removing the counter with a one-line cleaning step returns the score to 0.935 exactly, which isolates the interface string as the sole cause. The effect is not confined to one language: the same message in Hebrew falls from 0.996 to 0.321, and in Russian from 0.889 to 0.116.

A natural objection is that appending *any* text dilutes the score. It does not. Figure 1(a) shows that a benign sentence of greater length ("*please remember to buy milk and bread on the way home"*) leaves the score essentially unchanged (0.954), as do an appended URL and an emoji, whereas the short comment counter is by far the most damaging addition. The problem is specific to the kind of fragment, not to its length. Preprocessing decisions of this kind are a known and underappreciated source of instability in automated text analysis.[9]

There is an important limit to how far this generalises, and reporting it honestly is part of the argument. The collapse is dramatic on our primary model but not universal: on a second NLI model the counter had essentially no effect, and on a third the counter mattered little while a short off-topic phrase mattered more. The artifact is real, and on a popular production-grade model it is severe across languages, but its size and even its presence depend on the specific classifier. That model-dependence is the second knob, and we turn to it next.

4 Demonstration 2: The analyst's question, and the analyst's model

Zero-shot classification requires the analyst to phrase the target concept as a hypothesis, and that phrasing is an unconstrained choice.[10,11] Figure 2(a) scores four messages against four equally defensible phrasings of a single question: *is this post about conflict?* The scores swing so widely that the four phrasings carve out different datasets. A reported military strike is retained under "*describes violence"* and under a detailed conflict hypothesis, yet discarded under "*discusses politics"* and even under "*is about war or armed conflict."* A parliamentary-budget post, which contains no violence, is retained only under "*discusses politics"* and would quietly contaminate a "conflict" corpus built with that phrasing. One reasonable wording retains none of the four messages at all.

Because each phrasing is defensible and none is obviously correct, the contents of the resulting dataset reflect the analyst's wording rather than any property of the posts.[12]

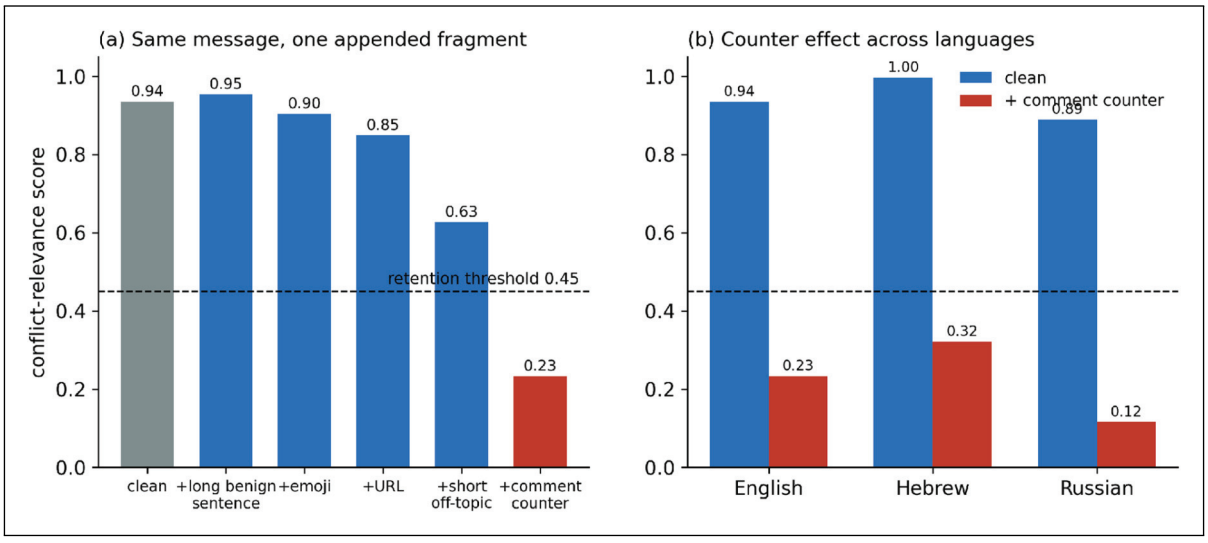

Figure 1. A captured interface string collapses the relevance score. (a) One clearly conflict-related English message scored with different appended fragments: the comment counter (red) drops it below the 0.45 retention threshold, while a *longer* benign sentence does not, ruling out length as the cause. (b) The same counter effect across English, Hebrew, and Russian (primary model). Exact values are annotated on each bar.

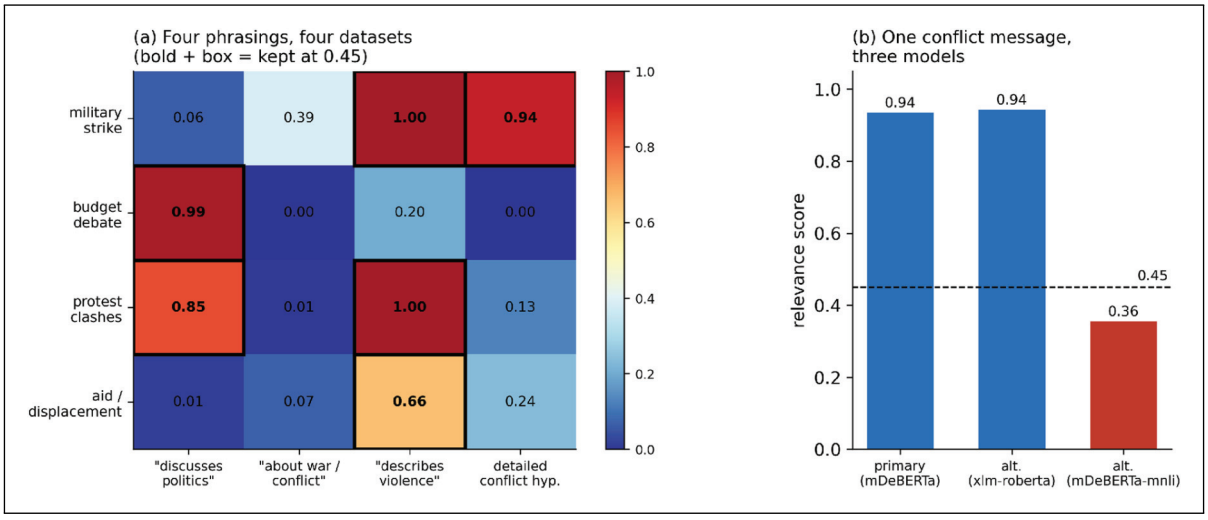

Figure 2. The question and the model each decide what counts. (a) Four messages scored against four defensible phrasings of "is this post about conflict?"; bordered cells are retained at the 0.45 threshold, and the four wordings produce almost disjoint kept-sets. (b) The same clean military-strike message scored by three public models, one of which discards it as irrelevant. Exact values are annotated.

The choice of model compounds this. Holding the message and the hypothesis fixed and swapping only the classifier, the clean military-strike message scores 0.94 on our primary model, 0.94 on one alternative, and 0.36 on another (Figure 2(b)). The third model would silently discard, as irrelevant, a message that plainly reports a military strike.[13] The model is a third knob, turned once at the start of a project and rarely revisited.

Sections 3 and 4 together describe three invisible degrees of freedom: the input as captured, the question as phrased, and the model as chosen (Figure 3). None is exotic. Each is a routine decision, and each can move a message across the line between "counted" and "discarded."

5 From artifact to false statistic

These per-message errors do not stay local. A pipeline rarely reports individual scores; it reports an indicator built by aggregating them: the share of posts discussing a conflict, a volume trend over time, a sentiment index for a

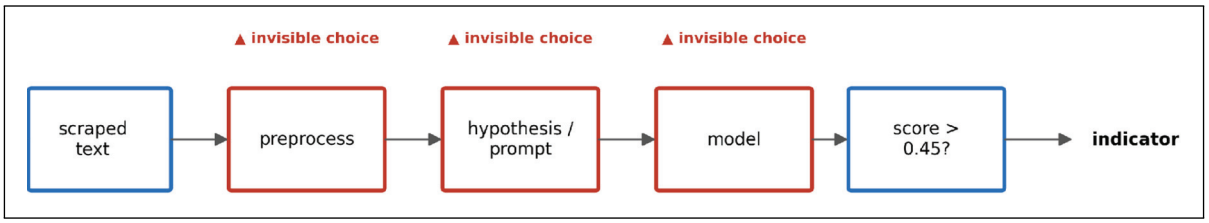

Figure 3. Three invisible knobs between text and statistic. Preprocessing, hypothesis wording, and model choice are each routine decisions, made once and rarely revisited; each can move a message from counted to discarded, and the aggregate indicator inherits the error.

region. Whatever bias the interpretation step introduces is inherited by that indicator, which then appears precise and authoritative. If a preprocessing artifact suppresses genuine conflict messages, an agency may report, with a tidy confidence interval, that a region barely discusses a conflict when in fact it discusses it constantly. If a hypothesis wording admits political-but-nonviolent posts, a "conflict intensity" series may rise for reasons that have nothing to do with conflict.

Two properties make these errors dangerous rather than merely regrettable. First, they are silent. The pipeline does not fail, throw an error, or flag low confidence; it returns plausible numbers, and with corpora far too large to read by hand there is usually no external check that would expose the discrepancy. Second, they are systematic. Because a single preprocessing rule, hypothesis, or model is applied to the whole corpus, the error is a consistent tilt that survives aggregation rather than random noise that averages out.[14]

For official statistics the stakes are specific. The value of an official figure rests on trust, and on a shared understanding that its quality has been assured against known failure modes.[15] Existing quality frameworks, for example the European Statistics Code of Practice and the Generic Statistical Business Process Model,[16,17] were built around sampling error, coverage, non-response, and source reliability.[18] They have little to say about the validity of an automated interpretation step that converts text into a number, because that step did not exist when the frameworks matured. An interpretation artifact is therefore a quality risk that current assurance does not catch, and it appears where agencies are expanding fastest.

Although our demonstrations use zero-shot NLI classification, nothing in the mechanism is specific to that task. The three knobs correspond to conditions that recur wherever text is converted into numbers without per-item human review: an input assembled by collection machinery rather than by the author, an unconstrained analytical choice that operationalises the concept being measured, and an instrument whose behaviour is opaque and version-dependent. Sentiment analysis meets all three. Captured chrome and boilerplate enter the scored text in the same way, the choice among lexicons, trained models, and instruction phrasings plays the role of the hypothesis, and disagreement between defensible sentiment instruments is itself well documented.[13] Topic models fare no better: preprocessing decisions as mundane as stopword lists and stemming are known to reshape the recovered topics,[9] and a captured interface fragment becomes a high-frequency token that can seed or absorb a topic of its own. In retrieval-based systems, including the retrieval step of retrieval-augmented generation, contaminated passages shift embeddings, the query does the work the hypothesis does here, and the embedding model is the instrument; the added difficulty is that errors surface only indirectly, in what a downstream system was or was not shown.

We therefore expect artifacts of this family wherever those conditions hold. Their size in any given task is an empirical question, as it was here for NLI, and the summary that Sections 3 and 4 support (real, sometimes large, idiosyncratic across instruments) is the appropriate prior. The audit below is phrased for classification, but its logic transfers directly: ablate what the collector added, vary what the analyst chose, and disagree-test the instrument.

6 An interpretation audit for official statistics

Catching these errors is cheap and requires no new theory. The same manipulations that expose the artifacts serve as the audit, and they fit inside existing quality-assurance routines.[19]

The most basic defence sits upstream of the audit, in how the data are collected. Social-media items should be parsed into structured, typed records (post body, comment bodies, author, post and comment timestamps, and engagement counts such as the number of comments) rather than flattened into one block of text. When the comment count lives in its own numeric field it cannot enter the passage the model scores, and the contamination in Demonstration 1 never arises; structuring the input at collection is the root fix for that class of error, and cheaper than any later cleaning.[20] It is not a complete answer. Bulk collection routinely flattens anyway: scrapers that read rendered pages tend to concatenate interface text with the post, which is why the artifact appears in real pipelines and not only in principle. Even clean entities leave open which of them

Table 1. The interpretation audit at a glance: each step, the question it answers, and the artifact class it is designed to detect.

#	Audit step	What it checks	Artifact class detected
0	Entity-structured collection	Input arrives as typed fields, not flattened page text; the fields joined into the scored passage are recorded	Interface contamination at source (Sec. 3)
1	Chrome ablation	Scores are stable with and without interface fragments a scraper may capture	Input contamination (Sec. 3)
2	Prompt-sensitivity analysis	The retained set is stable across several defensible phrasings; wording fixed before results are seen	Question-wording dependence (Sec. 4)
3	Model-sensitivity spot-check	An independent alternative model agrees on a rescored sample	Instrument dependence (Sec. 4)
4	Gold-set anchoring	Scores are validated against a small, independently coded sample	External validity (all classes)
5	Provenance record	Model revision, prompt, cleaning steps, threshold, and code travel with every published figure	Auditability and silent drift (all classes)

Steps 1–3 are sampling-based and cost minutes of compute; steps 0, 4, and 5 are structural properties of the pipeline.

are joined and fed to the model, a composition choice that is itself consequential, and structuring the input does nothing to address the wording and model effects of Section 4. An audit should therefore confirm that collection is entity-structured and record which fields were combined into the scored text.

Beyond collection, five checks cover most of the remaining risk. The first is an ablation of captured chrome: score a sample of inputs with and without the interface fragments a scraper may pick up, such as comment counters, button labels, and boilerplate, and confirm the score barely moves. A large change reveals contamination. The second is a prompt-sensitivity analysis: score the same sample under several reasonable phrasings of the target concept and report how far the retained set moves.[21] Where the contents of a dataset depend on the wording, that dependence should be disclosed and the wording fixed in advance rather than chosen after seeing results.[22,23] The third is a model-sensitivity spot-check: rescore a sample with at least one alternative model, since large disagreement means the conclusion rides on an arbitrary instrument choice.

Which alternative model to use for that check is itself a choice worth making deliberately. The comparison is informative only if the alternative is genuinely independent: a fine-tune of the same base model shares its blind spots, so agencies should prefer a different architecture or training lineage. The remaining criteria are practical. Language coverage should match the corpus rather than the benchmark, the model revision should be pinned and documented so that the check is repeatable, and the task interface should be close enough that the same inputs and operationalisation can be scored unchanged. The purpose is not to find a better model; disagreement is the finding, because it bounds how much of a published number is instrument rather than signal.

The remaining two checks anchor and document the pipeline. A small, independently coded gold set lets the claim "the model is wrong" be established against an external truth rather than asserted. Every published result should then travel with its provenance: the exact model revision, the hypothesis or prompt, the cleaning steps, the threshold, and, where possible, the scoring code, so that a figure can be regenerated and audited.[24–26]

Table 1 summarises the audit for practitioners.

To show this is practical rather than aspirational, we applied the first and last checks to our own research corpus of 24.36 million machine-scored messages. Drawing a random sample of 3000 and rescoring each independently, we found that none carried a comment-counter in the scored text (an upper bound on contamination of about 0.1% at 95% confidence), and that 86% of stored scores reproduced a fresh, independent re-score within 0.05 (median difference 0.005). The audit cleared the corpus of the specific artifact this paper describes, and it surfaced a small, localised pocket of scores computed under an earlier model configuration, which is the kind of finding an interpretation audit is meant to make visible. The exercise cost a few minutes of compute.

7 Conclusion

Misinformation in statistics built from text does not require a liar. It can be produced by an interface string that a scraper captured, by an offhand choice of wording, or by the model an analyst happened to load. The demonstrations here are deliberately small and fully reproducible, but the mechanism scales to any pipeline that turns text into numbers. The errors it produces are silent and systematic, and their size and direction depend on the particular preprocessing, prompt, and model, so they cannot be assumed away. As statistical agencies expand into automated text analysis, treating data interpretation as something to be audited rather than assumed is a low-cost safeguard for the credibility of official statistics. The cost of these checks is trivial next to the cost of a confidently wrong official number.

Several research directions follow. The audit proposed here is a minimal, demonstrated core. Developed into a

standardised protocol, as an interpretation-validity extension of the quality framework for statistical algorithms[19] and an explicit sub-process within the GSBPM,[17] it could become routine for agencies rather than an ad hoc precaution. A shared multilingual benchmark of collection artifacts (interface chrome, boilerplate, platform-specific fragments) would let instruments be stress-tested before deployment rather than after publication. The task generality argued in Section 5 should be established empirically for sentiment, topic-modelling, and retrieval pipelines. And because deployed models are updated continuously, interpretation auditing will need to develop from a one-time gate into ongoing monitoring, so that a silent model revision cannot re-introduce the artifacts an initial audit excluded.

Acknowledgements

During the preparation of this work the authors used Claude (Anthropic) to assist with drafting and editing the manuscript, running the reproducible analyses, and preparing the figures and references. The authors reviewed and verified the accuracy of this output and take full responsibility for the content of the publication.

ORCID iD

Petr Pesov https://orcid.org/0009-0000-7202-8668

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Data and code availability

The code and materials needed to reproduce every figure in this paper (the example messages, hypotheses, cleaning function, model identifiers and revision hashes, and scoring scripts) are deposited as research data with this article under a Creative Commons Attribution (CC BY) licence. They rely only on publicly available models and run on a standard computer. The corpus audit reported in Section 6 used a non-public production database and is not part of the deposit; the two demonstrations and all figures are reproducible from the deposited materials.

Supplemental material

Supplemental material for this article is available online.

References

1. Ricciato F, Wirthmann A, Giannakouris K, et al. Trusted smart statistics: motivations and principles. *Stat J IAOS* 2019; 35: 605–613.
2. UNECE High-Level Group for the Modernisation of Official Statistics. *Machine Learning for Official Statistics*. United Nations, 2022.
3. Grimmer J and Stewart BM. Text as data: the promise and pitfalls of automatic content analysis methods for political texts. *Polit Anal* 2013; 21: 267–297.
4. Wardle C and Derakhshan H. *Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making*. Council of Europe, 2017.
5. Lazer DMJ, Baum MA, Benkler Y, et al. The science of fake news. *Science* 2018; 359: 1094–1096.
6. Williams A, Nangia N and Bowman SR. A broad-coverage challenge corpus for sentence understanding through inference. In: Proc. NAACL-HLT; 2018.
7. He P, Gao J and Chen W. DeBERTaV3: improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In: Proc. ICLR; 2023.
8. Laurer M, van Atteveldt W, Casas A, et al. Less annotating, more classifying: addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI. *Polit Anal* 2024; 32: 84–100.
9. Denny MJ and Spirling A. Text preprocessing for unsupervised learning: why it matters, when it misleads, and what to do about it. *Polit Anal* 2018; 26: 168–189.
10. Webson A and Pavlick E. Do prompt-based models really understand the meaning of their prompts? *Proc NAACL*; 2022: 2300–2344. arXiv:2109.01247.
11. Gururangan S, Swayamdipta S, Levy O, et al. Annotation artifacts in natural language inference data. *Proc NAACL, Volume 2 (Short Papers)*; 2018: 107–112. arXiv:1803.02324.
12. Jacobs AZ and Wallach H. Measurement and fairness. *Proc ACM FAccT* 2021: 375–385. doi:10.1145/3442188.3445901
13. van Atteveldt W, van der Velden MACG and Boukes M. The validity of sentiment analysis: comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Commun Methods Meas* 2021; 15: 121–140.
14. Meng X-L. Statistical paradises and paradoxes in big data (I): law of large populations, big data paradox, and the 2016 US presidential election. *Annals of Applied Statistics* 2018; 12: 685–726.
15. Allegrezza S, Langer W and Joxhe M. Trust in official statistics across Europe: evidence from two waves of Eurobarometer using multilevel models. *Stat J IAOS* 2023; 39: 55–73.
16. Eurostat. *European Statistics Code of Practice*. Rev. ed. Publications Office of the European Union, 2017.
17. UNECE. *Generic Statistical Business Process Model (GSBPM), v5.1*. United Nations Economic Commission for Europe, 2019.
18. Groves RM and Lyberg L. Total survey error: past, present, and future. *Public Opin Q* 2010; 74: 849–879.
19. UNECE High-Level Group for the Modernisation of Official Statistics. *A Quality Framework for Statistical Algorithms (QF4SA)*. UNECE, 2020.

20. Freire S, Pigaiani C, Tucci M, et al. *Non-traditional Data Sources Can Improve Population and Housing Statistics: NIMPHS Project Final Report*. Luxembourg: Publications Office of the European Union; 2026. doi:10.2760/0932917.
21. Steegen S, Tuerlinckx F, Gelman A, et al. Increasing transparency through a multiverse analysis. *Perspect Psychol Sci* 2016; 11: 702–712.
22. Simmons JP, Nelson LD and Simonsohn U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol Sci* 2011; 22: 1359–1366.
23. Gelman A and Loken E. The statistical crisis in science. *Am Sci* 2014; 102: 460–465.
24. Gebru T, Morgenstern J, Vecchione B, et al. Datasheets for datasets. *Commun ACM* 2021; 64: 86–92.
25. Mitchell M, Wu S, Zaldivar A, et al. Model cards for model reporting. *Proc ACM FAccT* 2019: 220–229. doi:10.1145/3287560.3287596
26. Bender EM and Friedman B. Data statements for natural language processing: toward mitigating system bias and enabling better science. *Trans Assoc Comput Linguist* 2018; 6: 587–604.