A Five-Stage Pipeline for Automated Impartiality Analysis of Public Service Broadcasting: A 150-Country Corpus and a Single Analytical Framework

Methodology Version 3.8.2

David Schlaepfer

Schweizerischer Verein für ausgewogene Berichterstattung (Swiss Association for Balanced Reporting, SVFAB), Zurich, Switzerland

ORCID: 0009-0000-5671-9266 · Corresponding author: kontakt@svfab.ch

October 2026

Keywords: media bias, public broadcasting, AI-assisted content analysis, impartiality measurement, source credibility, cross-country comparison

Licence: CC BY-ND 4.0 (Attribution-NoDerivatives 4.0 International).

Plain Language Summary

Public broadcasters are required by law to report accurately and to reflect a diversity of views. This paper describes how a computer program checks broadcasts against these requirements using their subtitles: it counts, by fixed rules, who gets to speak, which terms are used and what is missing, and backs every count with a quotation and a time stamp. The procedure was compared with 159 broadcasts on which the Swiss complaints authority (UBI) had ruled: within the same programme format, broadcasts in which it found a violation receive higher scores on average. The procedure thus follows the view of the UBI: if the nine cases in which the Federal Supreme Court overturned its decision are counted as dismissed, the difference largely disappears. A single score varies between two runs; statements about many broadcasts are reliable.

Abstract

This paper presents the complete methodology of a system for automated impartiality analysis of public broadcasting content, applied to an archived corpus covering 150 countries in 70 languages. The system processes subtitles (SRT/VTT, partly from the broadcasters, partly from in-house machine transcription) of news and current affairs programmes through a structured analytical framework comprising 5 analytical steps, 18 criteria — 17 quality criteria (K1–K18; K9 has been repealed since version 4.0, 16 carry a score), each derived from an explicit counting rule rather than from a holistic judgement and classified as clearly countable or requiring interpretation, and as the eighteenth (K19) the political positioning of the parties —, a 6-dimensional source credibility assessment (Source Traffic Light) as a tool of expert selection, context-dependent balance assessment, and country-specific legal anchoring. Each broadcast receives a composite score — the mean of all applicable criteria — on a 0-10 scale (0 = neutral/balanced, 10 = strongly one-sided), reported together with the single most serious finding (peak value) and the mean of the three most serious findings among the criteria a regulator examines (severity). Since version 3.6, four criteria carry two figures: the reading applied by the Swiss supervisory authority and, stated first, the reading this project considers necessary — the latter explicitly a position, not a finding of unlawfulness. The framework is calibrated on, and compared with, decisions of the Swiss Independent Complaints Authority for Radio and Television (UBI): in the 14 programme formats in which both occur, the 34 broadcasts with an upheld complaint score on average 0.31 points higher in the composite score and

0.66 points higher in severity than the 81 with a dismissed complaint (0–10 scale); under the reading of the Federal Supreme Court, which overturned nine of the upheld complaints, the difference largely disappears. The procedure does not reliably classify the individual broadcast. The rules were developed on these decisions. Two runs of the same broadcast differ on average by 0.71 points in the composite score and 1.29 in severity. A single universal prompt template, parameterised by country-specific legal frameworks and political landscapes, ensures methodological consistency across all monitored countries. This report serves as the canonical methodological reference for all publications, analyses and API-based research access that build on it.

1. Introduction

Public service broadcasters in democratic societies operate under legal mandates requiring balanced, accurate, and pluralistic reporting. In Switzerland, Article 4 of the Radio and Television Act (RTVG) codifies this obligation. Similar provisions exist in virtually all countries with public broadcasting systems: the BBC Charter (UK), the Broadcasting Act (Canada), Medieansvarslov (Norway), Zákon o ČT (Czech Republic), 放送法第4条 (Japan), and many others.

Despite these legal requirements, systematic, continuous, and comparable measurement of individual broadcast compliance remains largely absent. Existing media monitoring — whether by regulators, academic institutes, or self-regulatory bodies — typically relies on small samples, manual coding by small teams, and qualitative assessments that resist quantification and cross-country comparison. The result is a regulatory gap: legal obligations exist, but operational measurement infrastructure does not.

The framework described here addresses this gap through a fully automated, AI-assisted pipeline that analyses broadcast transcripts against 18 criteria, produces structured reports with timestamped evidence, and generates comparable scores across countries. New items are added every night.
This paper provides the complete methodological description of the system at prompt version 4.14-detail (since 27 September 2026). It supersedes earlier descriptions in companion papers (Schlaepfer 2026a, 2026b, 2026c) and serves as the canonical citation for researchers, regulators, and institutions engaging with data produced by the system.

2. System Architecture

2.1 Pipeline Overview

The pipeline processes broadcast content in five stages:

- Acquisition: Subtitle files (SRT/VTT) are collected from public broadcast platforms or produced by in-house machine transcription and converted to timestamped plain text with [MM:SS] markers.
- Analysis: Each transcript is processed through the universal analysis prompt (Section 3), producing a structured report with scores, evidence, and legal assessment.
- Post-processing: Since prompt version 4.0 the model no longer assigns criterion scores at all. For each criterion it reports a counting line (the observable quantities) and an evidence line (one timestamp per counted position); the evaluation applies the scale anchor formula to those counts locally. The composite score, peak value, severity and own measure are likewise computed locally. Where the counted trigger quantity is zero, the criterion is reported as 'not applicable' rather than as zero, and the analysis must document which candidates were examined and rejected.
- Conversion: Reports are converted to standardised DOCX and PDF formats via a central converter (version 4.2) with consistent formatting. For reports of prompt version 4.x the converter reads the computed table and never recalculates; 'not applicable' is

rendered without a bar, and criteria carrying no score are rendered as a note. A release gate and a wording sweep run before conversion; only reports that pass both are published.

- Aggregation: Individual scores feed into country-level and cross-country dashboards, meta-analyses, and trend reports.

All arithmetic operations (score averaging, trend calculations, statistical tests) are performed locally by the evaluation program, never by the language model. This design decision follows from documented systematic errors in LLM arithmetic.

2.2 Model Selection

Production under version 4 uses Claude Opus 4.6 (claude-opus-4-6) at temperature 0; older analyses partly come from Claude Sonnet 4.6. Values from different models are not pooled, because the criterion profile depends on the model. Temperature 0 largely removes sampling randomness, but it does not make the output identical run after run; the procedure is near-deterministic. How far two runs of the same broadcast lie apart, and how precise comparisons between groups of broadcasts are, is quantified in Section 8.3. Claims about a broadcaster over a series of programmes rest on the group.

2.3 Universal Prompt Architecture

A single prompt template serves all countries. Country-specific parameters are injected at runtime:

- Broadcaster and programme format (e.g., main news, debate, magazine)
- Applicable legal framework (short reference and full description)
- Political landscape: governing coalition, opposition, party table with ideological positioning (CHES scores where available), current fault lines
- Language instruction (analysis language matches broadcast language)
- Party-specific evaluation table (adapted to national party system)

The analytical structure, criteria definitions, scoring scale, and output format are identical across all countries. This ensures that a score of 5.0 in Norway carries the same methodological meaning as a score of 5.0 in Japan or Argentina. Format rules for machine-parseable labels (dimension codes, traffic light colours, score headers) are enforced in every language — labels are never translated.

3. Analytical Framework: Five Steps

Each analysis follows a fixed five-step structure. The ordering reflects a deliberate methodological sequence: political context first (to establish the reference frame), then content inventory, then criterion-by-criterion assessment, then synthesis, then legal evaluation.

Step 1 — Party-Political Bias Assessment

For each party in the national political system, the analysis evaluates whether the party's official programme positions were accurately represented in the broadcast. The reference standard is the party's published platform, not media commentary about the party. Each party receives a score from -5 (positions actively distorted) through 0 (not mentioned) to +5 (positions accurately and completely represented). A left-right tendency score aggregates the overall directional bias of the broadcast. This political positioning is the eighteenth criterion (K19; K9 remains unassigned). Where a party stands on the left-right axis is determined not by the model but by the Chapel Hill Expert Survey (CHES 2024); for countries not covered by CHES, expert estimates are used.

Step 2 — Broadcast Information and Thematic Frame

This step inventories the broadcast: participants (with function, affiliation, political spectrum), main topic, societal context, and — critically — a Source Traffic Light assessment for every protagonist (Section 5). The step then identifies 8-10 relevant perspectives that a balanced broadcast on this topic should include (derived from general knowledge, independent of the transcript), and checks each against the actual broadcast content. Missing voices are listed with their expected contribution. This inventory feeds the reference grid of criterion K4, which is the scored instrument; Step 2 itself produces no score.

Context-dependent balance assessment (Section 6) determines the applicable balance standard based on the topic type (scientific, societal, or political).

Step 3 — 17 Quality Criteria**: Detailed Analysis**

The core of the analysis applies the 17 quality criteria (K1–K18, K9 repealed), each classified as clearly countable or requiring interpretation (Section 4). For each criterion, the analysis identifies up to 3 evidenced instances (the strongest), each with: timestamp, direct quote, description of the technique, and explanation of why it is problematic. The evaluation derives each score from the reported counts on a scale from 0 (neutral/balanced) to 10 (strongly one-sided); one criterion, K15, is reported without a score.

Step 4 — Overall Assessment

Identifies the 3 dominant techniques, extracts the broadcast's implicit messages (content-level, personal-level, societal-level), and assigns the finding grade on a five-tier scale: balanced reporting (0-2), slight tendency (3-4), clear one-sidedness (5-6), systematic imbalance (7-8), extreme imbalance (9-10).

Step 5 — Legal Assessment

Maps identified findings to specific provisions of the applicable national broadcasting law. For each potential violation: the norm, the factual basis (with timestamp and quote), and the legal assessment. This step anchors the analysis in the country's regulatory framework and enables regulatory bodies to evaluate findings within their jurisdiction.

4. Quality Criteria (K1–K18)

The system measures 18 criteria: the 17 quality criteria of this section and, as the eighteenth, the political positioning of the parties (K19; Step 1, Section 3). Criterion K9 was repealed in version 4.0; the number remains unassigned so that older analyses stay readable. K15 is described but carries no score; sixteen criteria carry a score. The Source Traffic Light (Section 5) is not a criterion of its own but a tool of expert selection (K1). All scored criteria rest on counted observations; they differ in how tightly the transcript fixes the count. Since version 4.14 each criterion therefore carries a classification: 'clearly countable' where what is counted stands in the text so that one can point to it, or 'requires interpretation' where one must first decide what counts. The basis is the measured difference between two runs of the same broadcast (Section 8.3; the figure in brackets with each criterion is the mean difference in points, with the cut at about 1.3), and where data are lacking the construction of the criterion.

The former division into hardfacts (K1–K8) and softfacts (K10–K18) did not survive this measurement: completeness (K4, 1.64) and timing (K7, 1.72) vary more than the cut, while prejudgment (K17, 0.36) is the most stable criterion of all. The classification matters methodologically because it shows where a count can be disputed; depending on the assignment it can be shown in the report or left out.

Epistemological status of analytical categories. The 18 criteria and the Source Traffic Light framework are derived from the operational requirements of national broadcasting laws, not from normative premises of the authors. Terms employed in the coding scheme—such as ‘experts with a substantiated minority position’ (documented by own data and peer-reviewed publications; Section 6.1) or ‘guilt by association’ (K6: discrediting through association with negative groups or ideas)—describe empirically observable, operationally defined patterns rather than value judgements. The framework is legal-positivist in character: it operationalises what broadcasting legislation requires and assesses compliance against those requirements. Where criteria involve interpretive judgement, this is made explicit through the classification of each criterion as clearly countable or requiring interpretation, and each finding must be documented with textual evidence from the transcript. The coding scheme does not presuppose that any particular reporting outcome is desirable; it measures deviation from the legal standard of balanced, pluralistic, and accurate reporting as defined by each country’s broadcasting law.

4.1 Criteria K1 to K8

K1 — Expert Selection. Who appears as an expert? Assesses diversity, independence, and qualification of expert voices. Includes a deep source check (funding, mandate, conflicts of interest) and the Source Traffic Light assessment for each expert. Media framing of experts ('renowned', 'controversial') is treated as an analytical finding, not as fact. Counted: the number of contested specialist questions, the expert voices addressing them, how many hold the dominant position, and how many have an undisclosed interest. Where the broadcast raises no contested specialist question — a match commentary, a portrait, a results programme — the criterion is not applicable and the analysis records which candidates were examined. Classification: clearly countable (1.18).

K2 — Source Selection. Which sources are cited? Evaluates diversity, independence, and structural conflicts of interest of cited sources (NGOs, government agencies, church-funded bodies, advocacy organisations). Counted are the sources of factual claims, those without a nameable bearer ('allegedly', 'sources say', 'reportedly') and those from the dominant camp of the item; the score follows from the two shares. Classification: clearly countable (0.65).

K3 — Time Distribution. Speaking time allocation between positions. Broadcast duration, speaking time and word count are measured programmatically from the subtitle timings before the analysis begins and are injected into the prompt as binding facts; the model may not overwrite them. The model estimates the allocation among positions, because subtitles do not identify speakers. Classification: clearly countable — broadcast duration and speaking time are measured programmatically from the subtitles (0.41).

K4 — Completeness. What is not shown despite being relevant? Since version 4.0 the criterion applies a reference grid of eight cells (four framing functions by those affected and those deciding); the analysis determines which cells are relevant to the topic and which of them the broadcast leaves untouched. A separate count covers the case the Swiss supervisory authority most frequently upholds: an item in which a substantiated statement of the criticised party was available and only its result was reported. A mere denial, if reported, suffices (decision b.617). Classification: requires interpretation — which cells of the grid are relevant is a judgement (1.64).

K5 — Presentation of Numbers. Selective or misleading use of numbers. Only figures with an argumentative function are counted. For each figure the analysis first determines which of three dimensions are applicable — absolute value, proportion/per capita, trend over time — and then how many of those are missing; figures presented in a misleading way (categories conflated, an estimate stated as fact, unlike things compared as like) set the criterion to its maximum. Attribution of a figure to a source is no longer counted here but under K18. In reports of version 4.14 the criterion still carries its former name (Zahlen-Manipulation, 'Statistical Manipulation'). Classification: clearly countable (1.21).

K6 — Guilt by Association. Discrediting through association with negative groups or ideas. Includes a structured counter-check for every individual labelled 'conspiracy theorist' or similar: primary sources? falsifiable claims? risk matrix (what did they lose vs. gain?). Categorisation: A (researcher with methodology and evidence), B (borderline: partly substantiated, partly speculative), C (ideologue: unsubstantiated, non-falsifiable claims). Media framing is not evidence for categorisation. Applied symmetrically in both directions. Classification: requires interpretation by construction — whether a factual connection is established is a judgement.

K7 — Timing. Placement of incriminating and exculpating elements within the broadcast. Counted are all elements relating to the same allegation and, among them, those placed prominently (title, introduction, opening of the item); the score is the difference between the prominently placed incriminating and exculpating elements, relative to all of them. Classification: requires interpretation — what counts as highlighted is a judgement (1.72).

K8 — Selective Outrage. Outrage at certain positions but not comparable others. Methodological principle: the triggering event must be documented before any asymmetry can be assessed. Cases are counted by severity rather than as a proportion: a severe case requires that one side carries an outrage marker while the comparable case in the same item carries none, and that no factual difference justifying the asymmetry is stated. Without a comparable case the criterion is not applicable. Classification: requires interpretation by construction — whether two events are alike is a judgement.

K9 — Retired (version 4.0). Measured across 1,393 analyses of the 3.x series, K9 correlated with K4 at r = 0.987 and carried an identical score in 90 per cent of reports: two criteria measuring the same thing entered the composite score twice. Its subject matter is now covered by the reference grid in K4. The number remains unassigned.

4.2 Criteria K10 to K18

K10 — Framing. Two readings, reported side by side. The supervisory reading asks what the Swiss authority asks: was the perspective announced, and does the introduction hold up against the item? Measured on 203 documented instances, the UBI upheld complaints against a declared perspective in 5 per cent of cases and against an undeclared one creating a false overall impression in 63 per cent; Ofcom Rule 5.10 is identical in substance. The project's own reading, stated first in the report, asks a different question: in how many items does only one side of the contested matter appear, and in how many of those is the other view not even named. In addition, and without a score, the report states in prose how the broadcast narrates its subject — with quotation and

timestamp — and what a counterweight would have cost: one sound bite, one sentence of context, one follow-up question. Classification: requires interpretation (2.57).

K11 — Word Choice. Which language is used and which connotations are set? For each finding, the neutral alternative is stated explicitly. Counted are the valuative terms and the side each is aimed at — not the speaker who utters it. Below five such terms no ratio is formed: three terms aimed at one side are a finding, but not a measure, and would otherwise carry the same maximum as thirty. Classification: requires interpretation — whether a term is evaluative is a judgement (2.49).

K12 — Moderation Behaviour. Asymmetries in follow-up questions, interruptions, and expressions of sympathy. Counted per interviewee: answers given and answers broken off mid-sentence. Which of two interviewees was interrupted more often is a calculation, not an observation, and is determined by the evaluation, not by the model. Methodological principle: the triggering event must be documented. An intervention can only be assessed as asymmetric if comparable triggers with other guests produced no analogous intervention. Classification: requires interpretation — in 22 per cent of double measurements the applicability changes.

K13 — Question Asymmetry. Different hardness of questions directed at different participants. Counted per interviewee are the questions put and, among them, those containing an allegation or a counter-claim; the score is the difference between the highest and the lowest share. With only one interviewee the criterion is not applicable — that the other side is absent is recorded by K1 and K3. Classification: clearly countable — a reproach is recognisable from the wording (0.82 on four pairs).

K14 — False Balance. Artificial balance despite actual imbalance: a substantiated and an unsubstantiated position stand side by side as if equivalent, with comparable space, and no sentence in the item explains the difference. The standard is whether the statements can be substantiated, not the credibility of the speakers; the latter is assessed by the Source Traffic Light under K1. Cases are counted by severity rather than as a proportion: in real broadcasts the denominator is almost always one, and a proportion over one case states nothing. Classification: requires interpretation by construction — whether a position can be substantiated has to be assessed.

K15 — Agenda-Setting. What counts as normal/self-evident? What does not make the agenda? The criterion carries no score in the results table: across 865 analyses of the 3.x series it averaged 4.03 but was the highest value in only 6 per cent of broadcasts, and the Swiss supervisory authority addressed agenda-setting in 2 of 282 examined matters. What is countable is the unexamined premise — the prior question the broadcast never puts — and that count appears in the project's own column only. Classification: requires interpretation by construction — the unquestioned premise is an assessment.

K16 — Care and Human Dignity. Is an identifiable individual disparaged personally rather than criticised for their conduct, or shown in a state of distress against their evident wish? Cases are counted by severity; three or more identifying features together with particular vulnerability (bereavement, illness, children) constitute the severe case. Introduced in version 4.0 after 13 of 130 examined supervisory decisions concerned matters for which no criterion existed — nine of them human dignity, four the presumption of innocence, which is now K17. It is modelled on the Swedish Broadcasting Commission (Granskningsnämnden), which treats 'varsamhet' — care towards those affected — as a test of its own, regardless of whether proceedings are pending; it was at issue in 17 of the 385 evaluated cases concerning SVT, SR and UR. The criterion adopts this separation: if anyone is disparaged or exposed, that is a finding in its own right, even without an accusation and without proceedings. Classification: requires interpretation by construction — disparaging someone rather than criticising their conduct is a judgement.

K17 — Prejudgment. Where proceedings are pending, is the accused presented as convicted? Counted are those concerned, the cases in which guilt is asserted without reservation, and

those in which it is merely implied. Whether proceedings are pending is taken from the broadcast itself, not from research: the supervisory authority likewise assesses the situation as it stood at the time of transmission. Classification: clearly countable — the most stable criterion (0.36; identical in 91 per cent of double measurements).

K18 — Source Attribution. Are central claims attributed to a nameable, verifiable source, and are the interests of those appearing disclosed? Counted are the central claims and those without attribution ('studies show', 'experts agree'). Source attribution is a separate examination point in seven of the eleven supervisory corpora examined decision by decision. Classification: requires interpretation — whether a source is named verifiably is partly a judgement (1.48).

5. Source Traffic Light (Quellenampel)

The Source Traffic Light is not a test method of its own but a tool of expert selection (K1). Every protagonist appearing in a broadcast — experts, interview partners, cited specialists, representatives — is assessed on six dimensions, each scored from -2 to +2. The assessment is based on general knowledge about the individual, NOT on how the media frames them. The media's framing ('renowned expert', 'leading researcher', 'controversial', 'conspiracy theorist') is itself an analytical finding, not a fact.

D1 — Conflict of Interest. Split into D1a (direct financial, institutional, or ideological self-interest) and D1b (structural patronage dependency). D1b uses 5 indicators:
(1) financial amplification by a dominant funder,
(2) network co-optation,
(3) career gatekeeper dependency,
(4) one-sided endorsement chain,
(5) institutional monoculture (>50% from one ecosystem).

0-1 indicators = 0, 1-2 = ±1, 3+ = ±2.
D1 = max(|D1a|, |D1b|) with sign. Applied symmetrically across all political directions.

D2 — Personal Risk. What does this individual risk through their position? Career loss, social exclusion, financial consequences. High personal risk increases credibility.

D3 — Professional Competence. Is the individual speaking within or outside their field of expertise? Primary researchers within their domain score highest.

D4 — Opinion Consistency. Differentiates three types of position change: (a) Opportunistic bandwagoning with the political wind (negative), (b) superficial but consistent positioning (neutral), (c) self-critical correction against one's own prior position with reasoning (positive). Self-correction earns credit; being corrected and then following along does not.

D5 — Emotionalisation vs. Data. Is the argument data-based or appellative? Evidence-driven discourse scores positively; emotional appeals without substantiation score negatively.

D6 — Source Level. Primary source (+2) / Secondary source (0) / Tertiary source (-2). Measures proximity to the original data or event.

The six dimension scores are summed to produce a TOTAL (-12 to +12). The traffic light classification follows: GREEN (≥+5), YELLOW (-4 to +4), RED (≤-5). Each source is additionally categorised: A (researcher with methodology and primary data), B (borderline), C (advocacy: advancing a position without own primary data), N (not classifiable: politics/activism), 0 (no verifiable justification).

Crucially, discrepancies between the traffic light result and the media's framing are findings. An individual framed as 'renowned expert' may score YELLOW or RED (e.g., a CEO with conflict of interest). An

individual framed as 'controversial' may score GREEN (e.g., a system-critical researcher with methodology, personal risk, and primary data). In both cases, the discrepancy feeds into K1, K6, K10, or K11.

The Source Traffic Light has been checked across 27 thematic matrices covering the major policy topics in Switzerland; in the process, 219 classifications were refined and 329 added (Schlaepfer 2026c). Dimension D3 (Professional Competence) was tested against an external benchmark: for 453 experts across 25 contested topics the rating is associated with scientific publication output (Spearman $\rho = 0.591$, $n = 453$; Schlaepfer 2026d).

Cross-cultural operationalisation. Each of the six Source Traffic Light dimensions is operationalised through observable, verifiable indicators applicable across political systems and media environments.

D1 (Conflict of Interest) relies on publicly verifiable facts: institutional employment, board memberships, funding relationships, and organisational affiliations.

D1b (Structural Patronage) is assessed through five indicators: career dependency on a single institution, gatekeeping power of that institution, absence of alternative career paths, publicly documented loyalty expectations, and pattern of position alignment with institutional interests.

D2 (Personal Risk) evaluates documented or plausible consequences for position-taking: career jeopardy, social ostracism, financial loss, or physical threat, contextualised to the specific political environment.

D3 (Professional Competence) is assessed against verifiable credentials: academic qualifications, publication records, and whether the individual is speaking within their documented field of expertise.

D4 (Opinion Consistency) requires longitudinal observation of publicly documented positions.

D5 (Emotionalisation vs. Data) distinguishes evidence-based argumentation from appellative rhetoric through transcript-level textual analysis.

D6 (Source Level) categorises proximity to original data or events along a three-point scale. Independent replication requires access to the same public biographical and institutional information used in the original assessment; a replication protocol specifying the required data sources for each dimension is available from the authors upon request.

6. Context-Dependent Balance Assessment

Not all topics require the same balance standard. The system distinguishes three topic types, each with its own assessment framework:

6.1 Scientific Topics

For topics involving research, technology, medicine or climate, the professional consensus serves as a working basis, not as proof. If experts with a substantiated minority position (own data, peer-reviewed publications, personal risk; Category A in the Source Traffic Light) are systematically not heard, this is a finding. Substantiated dissent is not the same as an unsupported counter-claim; the latter is covered by K14 (False Balance).

6.2 Societal Topics

For topics involving housing, ageing, gender, social policy or health, all affected groups should be represented, explicitly including those who finance a service (taxpayers, premium payers, contributors). Those affected only temporarily do not count as a co-determining group for infrastructure they do not help finance. On gender topics, both women and men should be heard; where laws or services systematically leave out one side, this is noted. The standard is proportionality.

6.3 Political Topics

For topics involving parties, elections, legislation or referenda, parliamentary strength and popular majorities serve as the reference for the allocation of speaking time; party positions come from Step 1. For referenda, the result is the clearest reference.

7. Scoring Methodology

7.1 Score Scale

All criteria are scored on a 0-10 scale where 0 represents neutral, balanced reporting and 10 represents strongly one-sided content (this denotes measurable effect, not editorial intent). This directionality is essential: a high score indicates a problem, not quality. The scale is consistent across all criteria and all countries.

7.2 Composite Scores

The following figures are computed locally from the counting lines (never by the language model):

• Composite score = arithmetic mean of all applicable scored criteria; the report states the base (e.g., 'mean over 14 of 16 scored criteria')
• Peak value = the single most serious finding, with the criterion named
• Severity = mean of the three highest values among the nine criteria that correspond to what the supervisory authority examines (K2, K3, K4, K5, K10, K11, K12, K16, K17)
• Own measure = mean of the three highest values among the ten criteria this project considers necessary (K1, K5, K6, K7, K8, K10, K13, K14, K15, K18), using its own reading where two exist; reported before the severity figure and marked as a position, not a finding of unlawfulness
Criteria whose trigger quantity is zero are excluded from the means rather than counted as zero. A composite score is omitted only where no criterion at all is applicable; where more than a third are not applicable, it is of limited comparability.

Since version 4.14 there is no hardfacts or softfacts score any more (Section 4). Dropping them shifted the composite score by 0.17 points on average across 265 reports of version 4; the ranking of broadcasts was preserved (Spearman 0.977). The simple mean is steadier than the former mean of two group means: two measurements of the same broadcast differ in the composite score by 0.71 points on average instead of 0.82 (Section 8.3). Anyone wishing to weight criteria differently finds the individual values and their classification in the table of every report.

7.3 From Counts to Scores

For every criterion with a counting rule, the score is assigned by the evaluation, not by the language model. For each criterion the model supplies the counts of the specified quantities (line ZAEHLUNG), the timestamp in the broadcast for every counted item (BELEGE) and, where a criterion is not applicable, the record of the search (GEPRUEFT). A program reads these lines and computes the score according to the fixed rule of the criterion, identically for all broadcasts. If the model states a value of its own for the criterion, the computed value applies; where the two differ by half a point or more, the run log records both. Fifteen of the sixteen scored criteria carry a counting rule. For K3 the program measures broadcast duration, speaking time and word count (Section 4); the model estimates the allocation of speaking time and assigns the value itself. Where a criterion has two readings — the supervisory authority's and the project's own (Section 7.2) — the model counts both and the program computes both.

Before computing, the program checks the counts. For every counted quantity, the BELEGE line must give as many timestamps as were counted; otherwise the program reports the count as unsupported. If the counting line is missing or unusable — quantities missing, a negative value, or a rule result outside 0 to 10 — the model's value stands and is reported as such. If the rule returns 'not applicable' because the reference quantity does not occur in the broadcast, or occurs too rarely, the criterion is excluded from the mean rather than counted as zero; if the search record is missing, this too is reported as unsupported. If a

criterion has no value at all, the program forms no composite score. The lines are read by criterion number and timestamp format and are therefore independent of the report language.

8. Validation Framework

Measurement quality is examined in thirteen sections, each addressing a distinct aspect. Every empirical figure below states the sample it rests on and the prompt version under which it was measured; figures from different versions are not pooled.

8.1 Specificity (Sport Calibration)

Sports broadcasts serve as a control group. In October 2026, ten complete match broadcasts of the SRG were measured with version 4.14: three from 2026 — the opening match of the football World Cup (RTS), a Nations League match of the national team and a Champions League match (RSI, each with studio coverage before and after the match) — and seven matches with Swiss participation from 2009 to 2018, rebroadcast by SRF in 2020: five ice hockey matches, four of them of the national team, and two Super League football matches. The composite score ranged from 0.0 to 1.8 (mean 0.6, median 0.1); severity, which could be formed for seven matches, ranged from 0.0 to 3.4. Between one and ten of the sixteen criteria were applicable; under the rule of version 4.14 all values are therefore of limited comparability — the absence of anything measurable is not recorded as a finding. For comparison: broadcasts whose complaint the UBI dismissed average between 3.0 and 5.7 in the composite score and between 4.4 and 8.1 in severity, depending on the format (formats with at least four dismissed broadcasts).

A low score in a match commentary does not demonstrate balance; it shows that the framework finds little it is built to measure. Where a Swiss team plays, the commentary adopts its perspective, and the instrument counts this: in the quarter-final of the 2013 Ice Hockey World Championship all seven voices that spoke belonged to the Swiss team (source selection 5); in the Young Boys' title-deciding match of 2018 speaking time was distributed roughly two to one in favour of the Bern side (time distribution 3). For the Nations League match and the World Championship quarter-final, the project's own reading sets framing (K10) to the maximum of 10 and the supervisory reading to 0, because the Swiss perspective is declared.

8.2 Discriminant Validity (Cross-Country Comparison)

If the system measures real differences in editorial practice, scores should vary systematically between countries with known differences in press freedom and editorial independence. Cross-country comparison is meaningful only within one version family and one programme format, since both shift the scale; comparisons that meet those two conditions are reported in the country studies rather than here.

8.3 Intra-Rater Reliability (Repeat Measurement)

The same transcript analysed multiple times should yield consistent scores. Under version 4 (sub-version 4.9), 23 broadcasts (12 with an upheld and 11 with a dismissed complaint) were analysed twice under identical conditions. Two runs differ on average by 0.71 points in the composite score (computed as in version 4.14) and by 1.29 points in severity; the ranking of the broadcasts is partly preserved (Spearman between the runs +0.51 and +0.56). The difference varies by criterion: it is smallest for prejudgment (K17: 0.36, identical in 91 per cent of pairs) and for time distribution (K3: 0.41), and largest for framing (K10: 2.57) and word choice (K11: 2.49). A single value is thus a value with a band; differences below one point are not differences. For statements about groups of broadcasts this precision is more likely to suffice; what matters there, however, is above all how widely the broadcasts themselves vary. Taking this variation into account, the standard error of the within-format difference between upheld and dismissed broadcasts is about 0.15 points in the composite score and 0.26 in severity (Section 8.12).

Every Swiss report gives this spread and states whether it is a single measurement or the mean of two; for a single broadcast — forensic or legal use in particular — the mean of two runs is the reliable figure. For statements about a broadcaster across a series of programmes, repeated

generation is unnecessary: from roughly 36 broadcasts onwards the noise of the group mean falls below 0.2 points.

8.4 Reproducibility

What is deterministic is the computation. Every score follows from the counting lines by a published formula; given the same counts, the same figures result, and any third party can recompute them from the report itself. Each counted position carries a timestamp, so the counts can be checked against the transcript. Temperature 0 removes sampling randomness from the reading step; the residual variation between two readings of the same transcript is quantified in Section 8.3 and is reported with every single-broadcast claim.

8.5 Model Comparison

No comparison of several models exists for version 4. Under version 3, a stratified sample of 98 Swiss broadcasts was measured by three models from different providers (Claude Sonnet 4.6, Gemini 2.5 Flash, GPT-4o mini): all three placed the broadcasts in the same direction on average; at the level of individual values they agreed considerably less (Schlaepfer 2026a). Values of versions 3 and 4 are not comparable: on 19 broadcasts measured under both versions the ranking was hardly preserved (Spearman +0.25). Production uses a single model; values from different models are not pooled (Section 2.2).

8.6 Computational Correctness

All arithmetic is performed locally, not by the language model. The current pipeline enforces computational separation: the model reports counts and timestamped evidence, the local evaluation computes every criterion score and every aggregate. The model is instructed not to write the result block at all.

8.7 Construct Validity (Source Traffic Light — 27 Matrices)

Under the 3.x series the Source Traffic Light has been applied to 27 thematic policy matrices in Switzerland, each covering the major protagonists in a policy debate (from abortion to nuclear power). All 27 matrices were then systematically checked against each other: anyone who appears in several policy fields must be classified there by the same standards. This quality assurance refined 219 classifications and added 329; since then the assessments have been consistent across policy domains.

8.8 Universality (One Prompt, All Countries)

A single prompt template serves all countries. Country-specific legal frameworks, political landscapes, and party systems are injected as parameters. The analytical structure, criteria, and scoring remain identical. This design ensures that cross-country comparisons reflect genuine differences in editorial practice rather than methodological variation.

8.9 Relationship to Classical Intercoder Reliability

The methodology employs a language model as the sole production coder. This design choice requires transparent positioning relative to classical intercoder reliability (ICR) as established in quantitative content analysis (Krippendorff, 2011; Cohen, 1960).

In traditional content analysis, ICR demonstrates that independently trained human coders apply the same codebook to the same material with acceptable agreement, typically reported as Cohen's κ or Krippendorff's α. The system relies instead on two measures: (a) the repeat measurement at temperature 0 (Section 8.3), which quantifies how far two runs of the same broadcast lie apart; and (b) the comparison with the decisions of the supervisory authority (Sections 8.12 and 9), which shows how far the measurement reflects the reading of an independent specialist authority deciding in a formal procedure. A comparison of several AI models exists only for version 3 (Section 8.5).

The repeat measurement corresponds to test–retest reliability. The comparison with the authority's decisions is not intercoder reliability: it does not measure whether two coders arrive at the same count, but whether the figures are higher where an independent specialist authority found a violation — on decisions on which the rules were developed. At the level of aggregate results a comparison with a manual content analysis exists: for German public broadcasting, nine trained coders (Maurer, Kruschinski & Jost 2024) and the instrument in an earlier version arrived at the same direction of political positioning for all formats examined (Schlaepfer 2026e). That is not a comparison at the level of individual items.

The system's reliability claims rest on the repeat measurement and the comparison with the supervisory authority's decisions rather than on classical ICR in the strict methodological sense.

8.10 External Anchoring

The framework is anchored in the practice of named supervisory bodies rather than in abstract quality standards. Its criteria were developed against, and calibrated on, the decisions of the Swiss Independent Complaints Authority for Radio and Television (UBI), whose rulings are published in full and can be appealed to the Federal Supreme Court; the Federal Office of Communications (BAKOM) supervises the remaining obligations, and an ombudsman stage precedes the complaint.
For the international chapters the corresponding bodies are: Ofcom (United Kingdom), KommAustria (Austria), the Medienanstalten (Germany), ARCOM (France), AGCOM and GECA (Italy), CNMC (Spain), Granskningsnämnden (Sweden), Medietilsynet (Norway), JSN (Finland), Commissariaat voor de Media (Netherlands), RRTV (Czech Republic), KRRiT (Poland), Médiatanács (Hungary), Coimisiún na Meán (Ireland), RTÜK (Turkey), CRTC (Canada), FCC (United States), ENACOM (Argentina), ICASA and the

BCCSA (South Africa), BPO (Japan), the Korea Communications Standards Commission (South Korea), and Roskomnadzor (Russia). The register records for each of them what they examine, on what trigger, by what method and at what frequency.

The central finding of that survey bears on this paper: outside Germany, where the Federal Administrative Court has confirmed an individual entitlement (6 C 5.24), no jurisdiction surveyed provides an enforceable individual claim to a balanced programme. Supervision is mostly complaint-driven, retrospective and case-by-case. A continuous, comparable measurement therefore does not compete with supervision; it supplies the empirical layer that supervision does not maintain — as the secretariat of the UBI stated in writing in September 2026, there are no examination guidelines, checklists or practice overviews besides the published decisions; these are its only database, and to its knowledge its decision practice has never been systematically analysed.

8.11 Mapping to Classical Validity Types

For methodological transparency, the following mapping relates the validation dimensions to the four classical validity types recognised in communication research:

Construct validity: Addressed through Section 8.7 (27 thematic matrices, checked against each other), Section 8.1 (sport calibration as null hypothesis), the external test of Source Traffic Light dimension D3 against scientific publication output (Section 5), and the classification of each criterion as clearly countable or requiring interpretation, measured by the difference between two runs (Section 4).

Criterion validity: Addressed through Sections 8.12 and 9 (comparison with 159 UBI decisions), an external anchor in the decisions of a specialist authority independent of the research team; the rules were developed on these decisions.

Internal consistency (reliability): Addressed through Section 8.3 (repeat measurement: 0.71 points in the composite score and 1.29 in severity under version 4) and Section 8.4 (deterministic computation from published formulas and timestamped counts).

External validity (generalisability): Addressed through Section 8.8 (one prompt, all countries), Section 8.2 (discriminant validity across countries with known differences in press freedom), and ongoing international expansion.

Classical intercoder reliability in the strict sense—human-vs-human or human-vs-machine agreement coefficients reported as κ or α—is not yet available (see Section 8.9).

8.12 Calibration against Supervisory Decisions

The framework was calibrated against 159 broadcasts from 41 formats for which the Swiss Independent Complaints Authority had reached a substantive decision between 1998 and 2026 — 35 upheld complaints and 124 dismissals. The label is not the outcome of the document but the operative paragraph for the transmission actually measured: a decision may uphold a complaint about a morning bulletin and dismiss the one about the main edition, and only the transmission held in the corpus counts.

Because the format strongly shapes the measured value, comparisons are made within the same format. In the 14 formats with both upheld and dismissed broadcasts (34 upheld, 81 dismissed), the upheld ones lie on average 0.31 points higher in the composite score and 0.66 points higher in severity than the dismissed ones (mean of the per-format differences, weighted by the smaller of the two groups); in 9 of the 14 formats the mean of the upheld broadcasts is higher, in the composite score as in severity. If the labels upheld and dismissed are shuffled at random within each format, a difference this large arises in about six in a hundred cases for the composite score and in about three in a hundred for severity. For 30 of the 35 upheld broadcasts a dismissed broadcast was measured as a counterpart, in 29 cases from the same format; in the composite score the upheld broadcast lay above its counterpart 12 times and below it 18 times, in severity 17 times above and 13 times below. For a single broadcast the procedure therefore does not separate; the agreement only emerges on average, more clearly in severity than in the composite score. Without separation by format the difference is 0.11 points in the composite score and 0.42 in severity. Separated by language region, the upheld broadcasts in the eight German-language formats of the 14 (24 upheld, 59 dismissed) lie 0.67 points higher in the composite score and 1.07 higher in severity than the dismissed ones, and in the five

French-language formats (9 upheld, 19 dismissed) 0.60 points lower in the composite score and 0.32 lower in severity (Section 12); only one upheld broadcast comes from an Italian-language format. The procedure does not separate in every format: in the composite score, upheld broadcasts do not score higher than dismissed ones in four RTS formats and in DOK, in severity in three RTS formats, in DOK and in HeuteMorgen. In DOK and Temps présent, the only upheld broadcast is the one in which the Federal Supreme Court overturned the decision. Format mirrors — mean values per programme type among broadcasts the authority examined — are reported in each Swiss report so that a value can be read against comparable programmes rather than against an absolute scale.

Three further calculations show how far the result depends on individual decisions. The Federal Supreme Court overturned nine of the 35 upheld complaints and found that the broadcast in question did not violate the programme provisions or the requirement of accurate reporting. If these nine broadcasts are counted as dismissed, the within-format difference largely disappears: 0.09 points in the composite score and 0.29 in severity; differences this large also arise under random shuffling of the labels in about one in three and one in five cases respectively. The procedure thus reflects the reading of the UBI, not the outcome before the Federal Supreme Court. If, under the reading of the UBI, the 15 broadcasts whose decisions the counting rules cite as examples are left out, the difference is 0.40 points in the composite score and 0.80 in severity. The length of the broadcasts does not explain the difference: within the same format, upheld broadcasts are on average 7.5 per cent longer than dismissed ones (word count), which random shuffling of the labels produces in about one case in seven; within formats, severity is unrelated to length (Spearman −0.03), and adjusted for length the difference is 0.65 points in severity and 0.32 in the composite score. Five of the dismissed decisions concern several broadcasts over a period; one broadcast from each was measured. Without them the difference barely changes (0.31 points in the composite score, 0.63 in severity). This paper does not include a comparison with decisions on a whole period.

Two limits of this calibration are stated explicitly. The format measures along with the content: news bulletins and debate programmes differ systematically regardless of editorial conduct. And the authority decides on the file, not on the transcript — whether a statement was correctly reproduced or a side was asked for comment at all is not visible in the subtitle text, and part of the upheld complaints rests on exactly that. A low value is therefore no acquittal and a high value no verdict.

8.13 Constructed Test Material

Calibration against real decisions shows whether the framework measures the right thing; it does not show whether the scale reaches its upper end at all. For that, four broadcasts were written deliberately to violate every criterion, in the four Swiss formats, with documented counter-probes where nothing should trigger. Under version 4.12 they scored 8.8, 8.7, 8.1 and 7.3; the composite score cannot reach 10 because several criteria are mutually exclusive at their maxima — a side that never speaks cannot be interrupted. The material is not published and makes no statement about real broadcasts, broadcasters or individuals.

9. Testing against the Decisions of the Supervisory Authority

The most important external test is the comparison with the decisions of the supervisory authority; none of the institutions named is affiliated with or sympathetic to the author:

- UBI (Independent Complaints Authority for Radio and Television): Findings are compared against UBI decisions on specific broadcasts; this is the set of 159 broadcasts described in Section 8.12, which serves both to develop the rules and to test them. UBI decisions are legally binding and produced through an entirely independent process. How far the system's scores follow the UBI outcomes shows how far the measurement reflects the authority's reading — on cases on which the rules were developed.
- Swiss Federal Supreme Court: Its rulings do not serve as a benchmark. The Court reviews UBI decisions as an appellate instance, not as a supervisory authority for radio and television (Federal Supreme Court, judgment 2C_406/2017, consideration 1.2). In the comparison set, the Court overturned nine of the 35 upheld complaints; Section 8.12 shows how this affects the comparison.

The UBI does not operate under the research team's control or influence; its decisions are published, and the labels of the comparison set are read off them.

10. International Coverage

The archived text corpus covers 150 countries in 70 languages and comprises more than 26 million items (subtitle files and articles). New items are added every night; the analytical framework, the criteria and the scoring rules are identical for every country, and the same prompt template serves all of them. The 22 countries listed below are those for which the applicable broadcasting law has been documented in full, so that findings can be mapped to a specific national norm:

• Switzerland (CH): SRF/RTS/RSI — Art. 4 RTVG
• Germany (DE): ARD/ZDF/DLF — §26 MStV
• Austria (AT): ORF — §4 ORF-G
• Norway (NO): NRK — Medieansvarslov §4
• Netherlands (NL): NPO — Mediawet Art. 2.1
• France (FR): France Télévisions — Loi 86-1067 Art. 43-11
• Italy (IT): RAI — Contratto di Servizio Art. 6
• Canada (CAN): CBC/Radio-Canada — Broadcasting Act s.3(1)(d)
• USA: PBS/NPR — 47 U.S.C. §396
• Ireland (IRL): RTÉ — Broadcasting Act 2009 s.39(1)
• United Kingdom (UK): BBC — BBC Charter Art. 6
• Spain (ES): RTVE — Ley 17/2006 Art. 3
• Poland (PL): TVP — Art. 21 ust. 1 Ustawy o radiofonii i telewizji
• Finland (FI): YLE — Laki Yleisradio Oy:stä 7 §
• South Africa (ZA): SABC — Broadcasting Act s.6
• Czech Republic (CZ): ČT — §2 Zákona o ČT / §31 Zákona č. 231/2001
• Argentina (AR): TV Pública — Ley 26.522 Art. 3
• South Korea (KR): KBS/MBC — 방송법 제6조
• Japan (JP): NHK — 放送法第4条
• Sweden (SE): SVT — Radio- och TV-lagen 5 kap.
• Turkey (TR): TRT — RTÜK-Gesetz Art. 8
• Hungary (HU): MTVA — Mttv. §83

Three further countries carry a comparable body of analyses but are not part of the list above, because their legal anchoring has not yet been written up: Iceland (RÚV, 120 reports), Denmark (DR, 108) and Russia (Perwy Kanal, 79). Eleven additional countries exist at pilot scale with one to three reports each; a single report is an attempt, not an observation, and they are therefore not counted as coverage.

Each country configuration includes the full political landscape (government, opposition, party positions, CHES scores), the applicable legal framework, and language-specific analysis instructions. New countries can be added by defining these parameters without modifying the analytical core.

11. Companion Methodologies

Beyond the core analysis against 18 criteria, we have developed several companion methodologies that extend the analytical framework:

11.1 Noise Detection (Method 14)

A framework for measuring whether broadcasters 'flood with noise' — i.e., whether important topics are displaced by less relevant coverage. Uses 11 dimensions of news importance (weighted by democratic relevance, affected population, state necessity, and security) to produce a Noise Index. Draws on research by Galtung/Ruge, Eilders, Colistra, Zaller, Franck, and RAND. The retrospective variant, which compares airtime with external indicators (Suppression Index), is described in Schlaepfer (2026b).

11.2 Emotion Trend Analysis

Tracks emotional framing patterns (criteria K1-K7) across decades of broadcast history. Initial analysis covers 50+ years of Swiss public broadcasting, enabling longitudinal assessment of editorial evolution.

12. Limitations and Scope

A low score is not a clean bill of health. Since version 4.0 a criterion whose trigger quantity is zero is reported as 'not applicable' rather than as zero, and the mean is formed over the applicable criteria only. A broadcast can therefore score low because nothing this framework measures occurred — pure match commentary is the clearest case: in three of ten measured matches only one of sixteen criteria was applicable (Section 8.1). 'Nothing found' and 'nothing to find' are different statements, and the report distinguishes them.

The composite score is the weakest of the figures reported. Averaging sixteen scored criteria buries the single serious finding; peak value and severity exist for that reason and should be read first. How far two measurements of the same broadcast lie apart, and how precise comparisons between groups of broadcasts are, is quantified in Section 8.3. The framework is therefore suited to statements about a broadcaster over many programmes, not to a verdict on a single programme.

The following limitations apply to the current methodology:

- **Transcript-only analysis**: The system processes text transcripts, not audio or video. Visual framing (camera angles, image selection, graphics (texts on T-Shirts and Caps, etc.) is not captured. This is a deliberate scope limitation that increases reproducibility at the cost of completeness.
- **Transcription:** Where subtitles do not come from the broadcasters, they are produced by machine transcription. It contains mishearings, above all of names and dialect, and can distort individual counts.

- **Single-model production:** Production under version 4 uses a single model; a comparison of several models exists only for version 3 (Section 8.5). The difference between two runs (Section 8.3) is the residual measurement uncertainty of every single measurement.
- **Knowledge cutoff**: The language model's world knowledge has a training cutoff. For very recent events, the model may lack context that a human coder would have. The timestamped transcript partially compensates by providing the broadcast's own context.
- **Political landscape updates**: Country configurations (party positions, government composition) require periodic manual updates. Between updates, analyses may reference slightly outdated political constellations.
- **Language coverage:** Each analysis is written in the language of the broadcast; the archived corpus spans 70 languages. For languages with complex morphology or limited training data, analysis quality may be marginally lower than for the major European languages.
- **Language region:** In the French-language formats the procedure does not separate upheld from dismissed broadcasts (Section 8.12). Whether this is due to the counting rules, developed on mostly German-language decisions (none of the 15 broadcasts whose decisions they cite as examples is French-language), or to the weighting of the UBI is open; the comparison set contains only nine upheld French-language broadcasts, and only one from Italian-language formats.
- **Human coder comparison**: The comparison with legally binding decisions of the supervisory authority (UBI) anchors criterion validity. However, a classical intercoder reliability study with trained human coders applying the codebook independently—reported as Krippendorff's α—has not yet been conducted. The current evidence on reliability rests on the repeat measurement (Section 8.3) and the comparison with the decisions of the supervisory authority (Section 8.12).

13. Research Access

The methodology is documented in full in this paper and in the companion papers listed in Section 14. Enquiries from researchers, regulators and institutions are answered individually:

- **Verification:** Where a specific question arises, individual analyses can be recomputed and the counts underlying them examined against the transcript.
- **Companion Papers:** Detailed descriptions of individual methodological components are published as working papers (see Section 14).
- **Sample Reports: Illustrative full analyses are made available on request.**

14. References and Companion Papers

Companion Papers:

- Schlaepfer, D. (2026a). Systematic AI-Assisted Analysis of Public Broadcaster Impartiality: A Scalable Methodological Framework for Measuring Structural Bias in Public Service Media. SSRN Electronic Journal, Nr. 6688478.
- Schlaepfer, D. (2026b). Measuring Editorial Noise: A Retrospective Suppression Index for Public Broadcasting Content Analysis. SSRN Electronic Journal, Nr. 6733280.

- Schlaepfer, D. (2026c). The Source Traffic Light: A Six-Dimensional Credibility Framework for Systematic Source Assessment in Public Service Media. SSRN Electronic Journal, Nr. 6733880.
- Schlaepfer, D. (2026d). External Validation of a Multi-Dimensional Source Credibility Framework: Correlating Expert Ratings with Scientific Publication Output. SSRN Electronic Journal, Nr. 7326258.
- Schlaepfer, D. (2026e). Temporal Stability of Political Positioning in German Public Broadcasting: Convergent Evidence from Manual and Automated Content Analysis. SSRN Electronic Journal, Nr. 7344701.

Legal References:

- Swiss Radio and Television Act (RTVG), Art. 4, 17, 24, 86, 89. SR 784.40.
- Country-specific broadcasting laws as listed in Section 10.

Methodological References:

- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20*(1), 37–46. https://doi.org/10.1177/001316446002000104
- Colistra, R. (2012). Shaping and cutting the media agenda: Television reporters' perceptions of agenda- and frame-building and agenda-cutting influences. *Journalism & Communication Monographs, 14*(2), 85–146. https://doi.org/10.1177/1522637912444106
- Eilders, C. (2006). News factors and news decisions: Theoretical and methodological advances in Germany. *Communications, 31*(1), 5–24. https://doi.org/10.1515/COMMUN.2006.002
- Entman, R. M. (1993). Framing: Toward clarification of a fractured paradigm. *Journal of Communication, 43*(4), 51–58. https://doi.org/10.1111/j.1460-2466.1993.tb01304.x
- Galtung, J., & Ruge, M. H. (1965). The structure of foreign news: The presentation of the Congo, Cuba and Cyprus crises in four Norwegian newspapers. *Journal of Peace Research, 2*(1), 64–90. https://doi.org/10.1177/002234336500200104
- Jolly, S., Bakker, R., Hooghe, L., Marks, G., Polk, J., Rovny, J., Steenbergen, M., & Vachudova, M. A. (2022). Chapel Hill Expert Survey trend file, 1999–2019. *Electoral Studies, 75*, 102420. https://doi.org/10.1016/j.electstud.2021.102420
- Krippendorff, K. (2011). *Computing Krippendorff's alpha-reliability*. University of Pennsylvania ScholarlyCommons.
- Maurer, M., Kruschinski, S., & Jost, P. (2024). *Fehlt da was? Perspektivenvielfalt in der deutschen Medienberichterstattung.* Johannes Gutenberg-Universität Mainz.
- Zaller, J. (1992). *The nature and origins of mass opinion*. Cambridge University Press.

Acknowledgements

The author thanks Prof. Yasemin Giritli İnceoğlu (Bilgi University, Istanbul) for a detailed critical review of this manuscript. Her observations on the relationship between computational consistency and classical intercoder reliability, the epistemological status of analytical categories, the operationalisation requirements for the Source Traffic Light framework, and the mapping of the validation framework to established validity types in communication research have substantially informed the revision to version 3.4.1. Any remaining shortcomings are the sole responsibility of the author.