Protocol for the Systematic Literature Review on Electoral Integrity Strategies (ELIS 2025)

A Systematic Review with Potential for Living Review Updates

Author:
Carlos Rocha
Visiting Researcher
Imperial College Business School

Supervisor:
[Supervisor information will be added here after approval by professor]

Date:
December 2025

Document Type:
Protocol for the Systematic Literature Review
Adapted from PRISMA-P 2015 Guidelines

IMPERIAL

Protocol for the Systematic Literature Review on Electoral Integrity Strategies (ELIS 2025)

A Systematic Review with Potential for Living Review Updates

Adapted from PRISMA-P 2015 Guidelines
Version: 2.0 | **Date**: 12 January 2026
Principal Investigator: Carlos Rocha (Imperial College Business School)

1. Administrative Information

1.1 Title

Systematic Literature Review Protocol: Electoral Integrity Strategies (ELIS 2025)

This protocol describes a systematic literature review (SLR) of evidence on electoral integrity strategies from 1990 to present. The review is designed with infrastructure that could support future living review updates, subject to feasibility assessment following the initial review completion (see Section 7: Future Directions).

Substantial protocol amendments will be versioned (v1.x, v2.x) and documented in the Amendments section (Section 1.4).

1.2 Registration

Primary registry

This protocol will be registered on Spiral (Imperial College London institutional repository) and Open Science Framework (OSF) prior to commencing full-text screening. The OSF registration will use the standard systematic review registration template.

- **Protocol Type:** Systematic Literature Review
- **Living Review Status:** Not applicable (initial review is standard SLR)
- **Registration Timing:** Prior to full-text screening phase
- **OSF Registration ID:** [To be added upon registration]
- **Spiral URI:** [To be added upon deposit]

Future Living Review Registration

If living review updates are pursued following initial SLR completion (see Section 7: Future Directions), a formal protocol amendment will be registered documenting:

- Living review activation decision and justification
- Feasibility assessment results
- Update frequency and methodology

- Sustainability and exit criteria

This ensures the initial systematic review is registered and conducted under standard SLR methodology, with living review potential assessed post-completion based on empirical feasibility evidence.

1.3 Authors and Contributions

- **Carlos Rocha** – Protocol design, AI-assisted methodology, review strategy, lead author.
  Visiting Researcher
  Imperial College Business School
  e-mail: c.rocha@imperial.ac.uk
  https://orcid.org/0009-0009-6741-2193
  Guarantor of the review: Carlos Rocha (responsible for the integrity of the protocol, conduct of the review, and final reporting).
- [Supervisor information will be added here after approval by professor]

1.3.1 AI-Assisted Protocol Development

This protocol was developed with assistance from large language models (Claude.ai, ChatGPT, Gemini, Perplexity.ai) for drafting, code generation, and formatting. All AI-generated content was critically reviewed, edited, and approved by the principal investigator. Final responsibility for all protocol content, methodological choices, and scientific integrity rests entirely with the principal investigator. Detailed documentation of AI use in each review stage is provided in Section 3.4 and Annex F.

1.4 Amendments

This protocol will be registered on [OSF/Spiral] prior to commencing data collection. Post-registration amendments will be documented here and in the public CHANGELOG.md file in the ELIS SLR Agent GitHub repository. Major methodological amendments (e.g., changes to eligibility criteria or core outcomes) will be explicitly flagged in the final review report.

As of [registration date], no amendments have been made to this protocol.

1.5 Support and Sponsor

Institutional: Imperial College Business School (Visiting Researcher programme), providing intellectual environment and access to library resources.

2. Introduction

2.1 Rationale

This review investigates the technological, operational, and institutional dimensions that influence electoral integrity. It seeks to generate robust academic evidence on the strategies, mechanisms, and design features that strengthen auditability, publicity, and public trust, across diverse electoral system models. Particular emphasis is placed on interdisciplinary approaches to evaluating voting systems, including electronic and paper-based modalities, in light of rapid technological evolution and renewed concerns over democratic resilience. The synthesis aims to inform academic research and policy design toward effective and independent elections auditing.

To date, there is no consolidated systematic review that jointly examines technological, operational, and institutional strategies for electoral integrity across both electronic and paper-based systems. Existing reviews tend to focus on single technologies (e.g., DRE security, risk-limiting audits) or broader indices of “electoral integrity”, without systematically linking concrete design features to empirically observed outcomes.

Review Approach and Scope

This systematic literature review will synthesize evidence published from 1990 through the search date (planned: Q1 2026). The review uses systematic, reproducible methods and automated workflow tools (ELIS SLR Agent) that could facilitate future updates.

Initial Review Focus: The primary objective is to complete a comprehensive, high-quality systematic review of the existing evidence base. This will establish:

- Current state of knowledge on electoral integrity strategies
- Evidence gaps requiring future research
- Methodological patterns in the field
- Baseline for potential future monitoring

Living Review Potential: Following successful completion of the initial review, the infrastructure and methods developed here could support periodic evidence updates if:

a) The field demonstrates sufficient publication velocity to warrant updates
b) Stakeholder interest justifies ongoing synthesis efforts
c) Automation workflows prove reliable and maintainable
d) Resource availability supports sustained effort

Section 7 (Future Directions) describes the criteria and approach for potential living review activation. The initial review will be registered and conducted as a standard systematic review, not as a living review.

2.2 Objectives

Primary Research Question (PRQ)

What operational and technological strategies have been shown to improve the integrity or auditability of electoral systems since 1990?

Methodological Sub-question (MSQ)

What types of empirical designs and evaluation frameworks have been used to assess the effectiveness of electoral integrity strategies since 1990?

The phrase *"have been shown to improve"* in the Primary Research Question is used to reflect a range of empirical evidence types. It includes findings from experimental and quasi-experimental studies, comparative observational research, technical evaluations, and structured qualitative analyses that support a causal interpretation. Where causal inference is limited or contested, the review will clearly distinguish between robust findings and those that are suggestive or correlational in nature.

Analytical Sub-questions

a) **Systems & Mechanisms:** What specific technological or operational mechanisms have been associated with increased auditability or verifiability in voting systems?

b) **Institutional Conditions:** Under what institutional, legal, or regulatory conditions have these mechanisms been implemented?

c) **Trust & Perception:** How have these strategies influenced public trust, voter confidence, or perceptions of electoral integrity?

d) **Regional Variation or Global Trends:** What regional patterns or cross-national differences are observed in the adoption and evaluation of these strategies?

2.3 Conceptual Framework – SPIDER

To accommodate both quantitative and qualitative evidence relevant to electoral integrity and auditability, the ELIS review adopts the **SPIDER framework** (Sample, Phenomenon of Interest, Design, Evaluation, Research type). SPIDER is particularly appropriate for reviews that include diverse empirical methods including qualitative research (Cooke et al., 2012).

Table 1 – SPIDER Framework for ELIS

Component	Definition in ELIS Review
S – Sample	Electoral processes, voting systems, and stakeholders involved in national, subnational, or referendum elections. This includes: • Electoral management bodies and administrators • Voters and electoral participants • Election observers and auditors • Technology implementers and vendors **Scope**: Studies must examine real-world electoral contexts including implemented systems, official pilots, large-scale trials, or post-election analyses. Laboratory studies or purely hypothetical scenarios are excluded unless linked to actual implementation.

PI – Phenomenon of Interest	Operational and technological strategies and design features intended to improve the **integrity** and **auditability** of elections. Examples include: ● Voter-verified paper audit trails (VVPAT) ● Risk-limiting audits (RLAs) ● Parallel vote tabulation (PVT) ● End-to-end verifiable voting systems ● Biometric voter authentication ● Blockchain-based voting records ● Public reporting and transparency mechanisms ● Observer access protocols. **Focus**: The phenomenon must demonstrably relate to verifiability, error detection, fraud prevention, or public confidence in electoral outcomes.
D – Design	Empirical study designs that provide evidence about the phenomenon of interest. Accepted designs include: **Quantitative**: ● Randomized controlled trials (RCTs) or field experiments ● Quasi-experimental studies (DiD, RDD, matching) ● Comparative observational studies with controls ● Statistical audits with documented methodology **Qualitative**: ● Structured case studies with explicit analysis framework ● Comparative case analysis ● Process tracing with documented evidence ● Systematic election observation using codified criteria **Mixed Methods**: ● Studies combining quantitative outcomes with qualitative implementation analysis **Exclusion**: Opinion pieces, advocacy documents, purely theoretical papers, or recommendations without empirical grounding.|
E – Evaluation	Documented empirical outcomes demonstrating changes in electoral integrity, auditability, or trust. Acceptable evaluation evidence includes: **Direct Integrity Measures**: ● Discrepancy rates between voting channels ● Audit results and error detection ● Fraud or irregularity identification ● System compliance verification

 Auditability Measures: ● Audit completeness and coverage ● Independent verification success<

	● Transparency of records ● Traceability of processes **Perception Measures**: ● Voter confidence surveys (pre/post implementation) ● Stakeholder trust assessments ● Public acceptance indicators **Implementation Measures**: ● Observed effects on procedures ● Cost-effectiveness analysis ● Operational feasibility assessment **Requirement**: Studies must report specific, measurable outcomes with documented methodology. Vague claims of "improved integrity" without supporting data are insufficient.
R – Research Type	Quantitative, qualitative, and mixed-methods empirical studies that: ● Present systematically collected data OR ● Use structured evaluation frameworks OR ● Employ recognized research methods with documented procedures **Included Types**: ● Peer-reviewed journal articles ● Peer-reviewed conference proceedings ● Official evaluation reports with transparent methodology ● Academic theses/dissertations with documented research design **Excluded Types**: ● Grey literature without peer review ● Technical specifications or vendor documentation ● News articles or journalism ● Policy advocacy without empirical support ● Protocols or study designs without results

Framework Justification

The SPIDER framework is selected over PICO (Population, Intervention, Comparison, Outcome) because:

1. **Methodological Diversity**: ELIS includes qualitative studies (case studies, process analysis) alongside quantitative evaluations. PICO is primarily designed for clinical intervention trials.
2. **Phenomenon Focus**: Electoral integrity strategies are not always interventions in the clinical sense. SPIDER's "Phenomenon of Interest" accommodates descriptive studies, comparative analyses, and implementation research.

3. **Design Flexibility**: SPIDER explicitly accommodates diverse research designs (D) rather than assuming comparison groups, making it appropriate for observational studies, natural experiments, and case analyses.
4. **Established Use**: SPIDER has been validated for systematic reviews in social sciences, policy research, and complex interventions (Cooke et al., 2012; Booth et al., 2013).

This framework supports the **inclusive but rigorous** approach outlined in Section 3.1 Eligibility Criteria and reflects the **interdisciplinary scope** of electoral integrity research spanning computer science, political science, public administration, and law.

Reference

Cooke, A., Smith, D., & Booth, A. (2012). Beyond PICO: The SPIDER tool for qualitative evidence synthesis. *Qualitative Health Research, 22*(10), 1435-1443.

3. Methodology

3.1 Eligibility Criteria

The criteria detailed below were formulated to define the boundaries of the evidence base, guided by the research question. The ELIS SLR Agent uses these criteria to generate initial exclusion suggestions during the title and abstract screening phase.

Inclusion Criteria

Content	Rationale & Agent Note
a) **Publication Type:** Peer-reviewed articles from academic journals and fully reviewed conference proceedings (1990–2025).	**Agent Action:** The Agent is configured to query indices (e.g., Scopus, OpenAlex) that prioritise peer-reviewed sources.
b) **Language:** Publications in English, French, Spanish, or Portuguese.	**Methodological Note:** While the Agent can retrieve metadata in any language, full-text screening is restricted to these four languages, ensuring feasibility for the review team.
c) **Study Design:**Empirical studies only (quantitative, qualitative, or mixed-methods) that present original data or novel analysis.	**Agent Action:** The Agent flags studies identified as "Review," "Protocol," or "Non-empirical" based on keywords in the title and abstract (e.g., "Narrative Review," "Conceptual Paper").
d) **Focus (PEO):** Studies must explicitly focus on strategies, technologies, or operational changes demonstrably affecting: 1) voting system integrity, 2) auditability, 3) transparency, or 4) public trust in the system.	**Core Alignment:** Ensures the evidence base directly addresses the research question.

Exclusion Criteria:

Content	Rationale & Agent Note
a) **Non-Empirical Content:** Opinion pieces, editorials, commentaries, or theoretical/normative texts lacking a clear methodology section or empirical findings.	**Quality Control:** Filters out unsupported claims and non-evidence-based content.
b) **Out of Scope:** Articles focused solely on traditional political science topics such as party politics, voter behaviour, or turnout, where the findings are not directly linked to voting system design or operational security.	**Specificity:** Narrows the focus away from general political topics to engineering/design and security elements.
c) **Unverifiable Source:** Studies lacking verifiable authorship, an institutional affiliation, or a clearly documented publication venue.	**Auditability Safeguard:** Aligns with the *Frozen Data Contract* principle; ensures metadata reliability and prevents the inclusion of grey literature of uncertain origin.

3.1.1 Rationale for Peer-Reviewed Literature Focus

This review restricts inclusion to peer-reviewed academic literature. While systematic review guidance (PRISMA 2020; Cochrane Handbook) recommends considering grey literature to reduce publication bias, such inclusion is not mandatory and depends on resource-benefit trade-offs (Paez, 2017; Mahood et al., 2014).

Decision Rationale

For a single-researcher review, comprehensive grey literature inclusion would require substantial additional resources (estimated +50 hours) for:

- Manual repository searches of electoral organizations
- Case-by-case quality assessment without objective verification criteria
- Establishing institutional credibility standards

Peer-review status provides objective, reproducible inclusion criteria compatible with automated screening (Section 3.4).

Acknowledged Limitation

Important empirical evidence on electoral integrity strategies appears in grey literature, including election observation reports (OSCE/ODIHR, EU EOM), post-election audits, and technical evaluations. Exclusion may affect:

- Geographic coverage (developing democracies underrepresented in academic publishing)
- Timeliness (grey literature published 0-6 months vs. 18-36 months for peer-reviewed)
- Implementation evidence (practitioner evaluations of real-world deployments)

This review synthesizes peer-reviewed academic evidence. Readers should consult organizational reports and official evaluations alongside these findings.

Future Expansion

Grey literature inclusion is planned for a future review update when additional research resources become available. If the Living Systematic Review model is activated (Section 7), grey literature monitoring could be incorporated into periodic updates.

References

Mahood, Q., Van Eerd, D., & Irvin, E. (2014). Searching for grey literature for systematic reviews: Challenges and benefits. *Research Synthesis Methods, 5*(3), 221-234.

Paez, A. (2017). Grey literature: An important resource in systematic reviews. *Journal of Evidence-Based Medicine, 10*(3), 233-240.

Page, M. J., McKenzie, J. E., Bossuyt, P. M., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ, 372*, n71.

3.2 Information Sources

A total of eight bibliographic databases covering Computer Science, Political Science, and Engineering are included. The search across these sources is performed exclusively via their respective APIs, orchestrated by the ELIS SLR Agent. This high degree of automation facilitates the future implementation of a Living Systematic Review (LSR) methodology (Section 7).

This review searches peer-reviewed academic literature exclusively (see Section 3.1.1 for rationale). Grey literature may be referenced within included peer-reviewed studies but is not systematically searched.

3.2.1 Primary Search Sources (Automated Retrieval)

a) **Scopus** – Multidisciplinary database with comprehensive coverage across political science, governance, law, and engineering.

b) **Web of Science** – High-impact journal indexing platform enabling detailed citation analysis across disciplines.

c) **IEEE Xplore** – Technical literature repository covering electronic voting systems, cryptographic security, and system auditability.

d) **Semantic Scholar** – AI-enhanced bibliographic database covering 200M+ papers across computer science and interdisciplinary research, with citation graphs and semantic indexing.

e) **OpenAlex** – Open bibliographic database (250M+ works) providing comprehensive metadata including institutions, citations, and concept tagging.

f) **CrossRef** – DOI registration agency providing publisher-verified metadata for 130M+ records, enabling robust deduplication and citation tracking.

g) **CORE** – Open access aggregator covering 300M+ papers, theses, and preprints from institutional repositories worldwide.

h) **Google Scholar (via Apify API)** – Comprehensive academic search engine indexing scholarly literature across all disciplines, formats, and sources. Accessed programmatically via Apify Google Scholar Scraper API.

Source Selection Rationale

These sources were selected to provide:

- **Disciplinary breadth:** Coverage across political science, computer science, law, and governance
- **Methodological diversity:** Inclusion of both empirical studies and technical evaluations
- **API accessibility:** All sources provide documented APIs enabling reproducible automated searches
- **Complementary coverage:** Combination of subscription databases (Scopus, WoS, IEEE) and open sources (Semantic Scholar, OpenAlex, CORE) maximizes retrieval while supporting open science principles

Technical Implementation

All search results will be imported into Zotero for de-duplication and metadata management. Initial screening and tagging will be automated using the ELIS SLR Agent, an open-source Python workflow hosted on GitHub. The researcher verifies all inclusion decisions, with all decisions and version history recorded via GitHub for full transparency and reproducibility.

API Availability Policy

API accessibility is a fundamental requirement for source inclusion in this review. If an API becomes unavailable or unstable during the review period, the following protocol applies: (1) attempt to resolve technical issues through alternative authentication methods or API endpoints, (2) contact the data provider to restore access, (3) if access cannot be restored within 30 days, suspend searches from that source until API availability is confirmed. Any source suspension will be documented in the amendments log (Section 1.4) and reported in the PRISMA flow diagram. This policy maintains the automated, reproducible methodology essential for solo-researcher feasibility and Living Systematic Review continuity.

3.2.2 Peer Review Status Verification

To ensure inclusion of only peer-reviewed academic articles while using databases that index mixed content types, the following automated verification protocol is applied by the ELIS SLR Agent.

Source-Specific Filtering Rules

All records are automatically classified as peer_reviewed, non_peer_reviewed, or ambiguous based on source-specific rules. Records classified as ambiguous are automatically excluded following a conservative approach that prioritizes precision over recall. Classification Rules are described for each source:

Scopus and Web of Science

- All records automatically classified as peer_reviewed
- Both databases exclusively index peer-reviewed journals and conference proceedings with strict quality criteria
- No additional filtering required

IEEE Xplore

- contentType = "Journals" → peer_reviewed (all IEEE journals are peer-reviewed)
- contentType = "Conferences" AND conference in IEEE peer-reviewed proceedings list → peer_reviewed
- contentType = "Conferences" AND conference not in verified list → ambiguous (auto-exclude)
- contentType = "Technical Reports" OR "Standards" → non_peer_reviewed (auto-exclude)
- IEEE conference verification: Match against IEEE conference rankings database

CrossRef

- type = "journal-article" AND journal ISSN in DOAJ → peer_reviewed
- type = "journal-article" AND journal ISSN in Scopus/WoS source lists → peer_reviewed
- type = "journal-article" AND journal ISSN absent from major indexes → ambiguous (auto-exclude)
- type = "posted-content" OR "preprint" OR "working-paper" → non_peer_reviewed (auto-exclude)
- type = "book-chapter" → non_peer_reviewed (auto-exclude, out of scope)

CORE

- type = "journal article" AND DOI present AND journal in DOAJ/Scopus/WoS → peer_reviewed
- type = "journal article" AND DOI absent → ambiguous (auto-exclude)
- source indicates institutional repository AND type = "thesis" OR "dissertation" OR "working paper" → non_peer_reviewed (auto-exclude)
- type = "conference paper" without peer review flag → ambiguous (auto-exclude)

Semantic Scholar

- publicationTypes includes "JournalArticle" AND venue in verified journal list → peer_reviewed

- publicationTypes includes "JournalArticle" AND venue not in verified list → ambiguous (auto-exclude)
- venue contains "arXiv" OR "SSRN" OR "RePEc" OR "bioRxiv" OR "medRxiv" OR "PsyArXiv" → non_peer_reviewed (auto-exclude)
- publicationTypes includes "Conference" AND venue in ACM/IEEE/LNCS verified proceedings → peer_reviewed
- publicationTypes includes "Conference" AND venue not verified → ambiguous (auto-exclude)
- Verified journal list: Journals indexed in DOAJ, Scopus, or WoS
- Verified conference list: ACM, IEEE, Springer LNCS proceedings with documented peer review

OpenAlex

- type = "article" AND host_venue.is_in_doaj = true → peer_reviewed
- type = "article" AND host_venue in Scopus/WoS source lists → peer_reviewed
- type = "article" AND host_venue not in verified indexes → ambiguous (auto-exclude)
- has_published_version = true → use published version metadata (re-apply classification rules)
- has_published_version = false AND primary source is repository → ambiguous (auto-exclude)

Google Scholar (via Apify)

- Publication contains "journal" OR "conference proceedings" metadata AND venue cross-referenced with DOAJ/Scopus/WoS → peer_reviewed
- Publication metadata indicates "book chapter" OR "thesis" OR "dissertation" OR "working paper" → non_peer_reviewed (auto-exclude)
- Source URL contains "arxiv.org" OR "ssrn.com" OR "researchgate.net" (preprint indicators) → non_peer_reviewed (auto-exclude)
- Publication venue not verifiable against known peer-reviewed indexes → ambiguous (auto-exclude)
- Verified venue list: Cross-referenced with DOAJ, Scopus journal list, Web of Science Master Journal List
- Note: Google Scholar does not provide structured peer-review metadata; verification relies on venue cross-referencing with authoritative indexes

Conservative Exclusion Principle

Records classified as ambiguous are automatically excluded without manual verification. This conservative approach ensures only sources with verified peer review status are included, maintaining feasibility for a single researcher while prioritizing precision.

Ambiguous cases include

- Journals not indexed in DOAJ, Scopus, Web of Science, or SCImago
- Conference papers from venues without documented peer review in major databases
- Publications with conflicting or incomplete metadata
- Regional journals without international indexing

All auto-excluded ambiguous cases are logged with classification reasons. The count of ambiguous exclusions is reported in the PRISMA flow diagram under "Records excluded: peer review status unverifiable."

Documentation

- Ambiguous exclusions logged in: `` `data/ambiguous_exclusions.json` ``
- Fields: study_id, source_database, title, venue, doi, classification, exclusion_reason, date
- Summary statistics: Total ambiguous exclusions by source database and exclusion reason
- No manual verification log required (all classification decisions are automated).

3.3 Search Strategy

Boolean queries will target combinations of key terms related to voting technology, auditability, and public trust. Draft queries may be suggested by AI tools but require researcher validation before application.

Example query:

```
("electoral integrity" OR "e-voting security" OR "ballot auditability" OR "electronic voting trust")

AND

("VVPAT" OR "end-to-end verifiability" OR "blockchain voting" OR "cryptographic audit" OR "risk-limiting audit")

AND

("evaluation" OR "empirical study" OR "comparative analysis")
```

Search logs and filtering decisions will be documented for reproducibility. Search queries are initially generated by LLMs and refined through expert validation.

3.4 Study Selection/Screening Process

Overview of Automation and Researcher Oversight

This review employs a hybrid methodology combining **deterministic rule-based automation**, for consistency and efficiency, with **researcher judgment**, for nuanced decisions and validation.

The **ELIS SLR Agent** is a custom Python application that executes researcher-defined screening rules systematically across all records. Large language models (LLMs) are used to generate code implementations of these rules, while **all substantive decisions remain with the researcher**.

Two Distinct Automation Roles

Role 1: Code Generation (Section 3.4.1)

- LLMs convert researcher rule specifications into Python code
- Code executes deterministically (same input → same output)
- Researcher validates behavior through pilot testing
- Example: "Include if title contains 'audit' AND ('election' OR 'voting')"

Role 2: Data Extraction Assistance (Section 3.7.2)

- LLMs suggest data field values from article text
- All suggestions require researcher verification
- Only human-verified values are stored
- Example: LLM extracts country name → researcher confirms accuracy

Decision Authority

- Researcher defines all inclusion/exclusion rules
- Researcher validates all code behavior via pilot testing
- Researcher makes final decisions on all "uncertain" cases
- Researcher verifies all extracted data values
- No autonomous AI decision-making at any stage

Quality Assurance

- All rules documented in version-controlled repository
- All rule changes logged with rationale (Annex H)
- Pilot testing validates rule implementation
- Full audit trail enables reproducibility

Workflow Summary

Stage 1: Database Searches

- Automated execution of API queries via ELIS SLR Agent
- Researcher initiates and verifies successful execution
- Search logs document date, database, query, and retrieval count

Stage 2: Reference Management and Deduplication

- Zotero automated duplicate detection
- Researcher verifies all proposed duplicate removals
- Duplicate log records all removed records with reason

Stage 3: Peer Review Status Verification

- ELIS SLR Agent applies source-specific filtering rules (Section 3.2.2)
- Records classified as: peer_reviewed, non_peer_reviewed, or ambiguous
- Ambiguous cases auto-excluded (conservative approach)
- All classifications logged with reason

Stage 4: Title/Abstract Screening

- ELIS SLR Agent applies keyword-based screening rules
- Records classified as: include, exclude, or uncertain
- Researcher reviews all "uncertain" cases (~5-10% of records)
- All decisions logged

Stage 5: Full-Text Screening

- Researcher reads full text of all included records
- Applies complete eligibility criteria (Section 3.1)
- Makes final inclusion/exclusion decision
- LLMs may generate summaries to aid navigation (optional)
- All decisions logged with exclusion reasons (Annex C)

Stage 6: Data Extraction

- ELIS SLR Agent maps metadata to extraction fields
- LLMs suggest values for complex fields (researcher verifies all)
- Automated quality checks flag inconsistencies
- Researcher reviews flagged records

Stage 7: Risk of Bias Assessment

- Researcher assigns all ratings based on full-text reading
- ELIS SLR Agent may flag relevant keywords to aid navigation
- All ratings reflect researcher judgment
- No automated scoring

The following sections detail each stage's methodology, emphasizing the distinction between automated execution of researcher-defined rules and researcher decision-making.

3.4.1 Rule-Based Automation Architecture

The ELIS SLR Agent executes **deterministic classification rules** defined by the researcher. This section explains how researcher decisions are translated into reproducible, transparent code.

Conceptual Architecture

The review workflow consists of three components:

1. **Human Researcher**

- Defines inclusion/exclusion criteria (Section 3.1)

- Specifies logical rules in plain language
- Validates rule implementation through pilot testing
- Makes final decisions on edge cases
- Accepts full responsibility for scientific integrity

2. Large Language Models (Code Generation)

- Convert researcher rule specifications into Python code
- Generate implementations based on explicit instructions
- Produce executable screening_rules.py, peer_review_filter.py, etc.
- Do not interpret or modify researcher intent
- Role: Programming assistant, not decision-maker

3. ELIS SLR Agent (Deterministic Execution)

- Executes researcher-defined rules consistently
- Produces tagged outputs: include/exclude/uncertain
- Maintains full provenance (what rule triggered for each record)
- Ensures reproducibility (same rules → same results)
- No autonomous interpretation or learning

Critical Distinction

Researcher → Specifies Rules → LLM Generates Code →
Agent Executes Deterministically → Researcher Validates

- NOT: AI makes decisions
- NOT: AI learns from data
- NOT: AI interprets ambiguous cases

The Agent does not make decisions. It executes the researcher's codified logic. Supervision occurs through iterative rule refinement, not through output validation.

Stage-Specific Implementation

A. Title/Abstract Screening

Rule-Based Classification

1. **Researcher defines keyword combinations indicating inclusion**
   - Example: `["electoral integrity" AND ("audit" OR "verifiability")]`
   - Example: `["electronic voting" AND "security"] AND NOT "survey"`
2. **LLM generates Python implementation**
   - Input: Researcher's rule specification in plain language
   - Output: Python code in `screening_rules.py`
   - Example code generated:

```
python
def check_inclusion(title, abstract):
    text = (title + " " + abstract).lower()

    # Researcher-specified rule 1
    if "electoral integrity" in text:
      if "audit" in text or "verifiability" in text:
        return "include"

    # Researcher-specified rule 2
    if "electronic voting" in text and "security" in text:
      if "survey" not in text:
        return "include"

    return "exclude"
```

3. **Agent applies rules to all titles/abstracts**
    - Outputs: include / exclude / uncertain
    - "Uncertain" = records where rules cannot decide (researcher reviews)

4. **Researcher validates through pilot testing**
    - Test rules on 50-record calibration sample
    - Identify false positives/negatives
    - Refine rule specifications (re-specify to LLM for new code)
    - Re-run on full dataset
    - Iterate until satisfied with precision/recall

Key Point

Rules are deterministic. The same input always produces the same output. If the rule correctly captures researcher intent (validated through pilot), execution is correct by definition.

No Post-Hoc Validation Required

- Validation occurs during pilot testing of rules
- Once rules validated, execution is mechanical
- Researcher only reviews "uncertain" cases where rules cannot decide

Documentation

- `` `screening_rules.py` `` versioned in GitHub with comments explaining each rule
- Rule changes logged in `` `logs/rule_development.json` ``(Annex H)
- Example log entry:

```
json
```

```json
{
  "date": "2025-11-25",
  "rule_change": "Added 'transparency' to auditability synonym list",
  "rationale": "Pilot showed relevant papers discussing transparency were
  excluded",
  "git_commit": "a3f8d92",
  "records_affected": 47
}
```

B. Peer Review Status Verification

Rule-Based Filtering

1. **Researcher defines source-specific verification logic**
   - Example: `IF source=="CrossRef" AND type=="preprint" THEN exclude`
   - Example: `IF source=="IEEE" AND contentType=="Journals" THEN peer_reviewed`
2. **LLM generates conditional statements** in `peer_review_filter.py`
   - Input: Researcher's verification hierarchy
   - Output: Python code implementing classification logic
3. **Agent applies filters systematically to all records**
   - All records classified as: peer_reviewed, non_peer_reviewed, or ambiguous
   - Ambiguous cases auto-excluded (conservative approach per Section 3.2.2)

Output

- `peer_review_classification.json` auto-populated
- `ambiguous_exclusions.json` logs all auto-excluded records
- Both versioned in GitHub

C. Data Extraction

Metadata Mapping

1. **Researcher defines extraction rules**
   - Which metadata fields map to which extraction fields
   - Example: `extract_country: Look in ["authors.affiliations", "abstract", "keywords"]`
   - Example: `extract_modality: Map keywords ["e-voting","electronic voting"] → "Electronic"`
2. **LLM generates mapping logic** in `extraction_rules.py`

3. **Researcher validates through pilot extraction**
   - Test on 10-study sample
   - Verify mappings produce correct outputs
   - Refine rules for edge cases
   - Re-run on full dataset

Post-Extraction Quality Check

- Automated consistency validation
- If error rate >5%: flag for researcher review and rule revision
- If error rate <5%: document quality metrics and proceed

D. Risk of Bias Assessment

Keyword-Based Flagging (Navigation Aid Only)

1. **Researcher defines bias indicator keywords**
   - Example: `` `IF methods contains "randomized" THEN flag_design_bias="likely_low"` ``
   - Example: `` `IF data_availability=="none" THEN flag_transparency_bias="high"` ``
2. **LLM generates flagging logic**
3. **Agent flags studies** to direct researcher attention

Critical

All bias ratings assigned by researcher after reading methods sections. Flags are navigation aids, not assessments.

Use of Large Language Models (LLMs)

LLMs (ChatGPT, Claude.ai, Gemini, Perplexity.ai) are used exclusively for:

Code Generation (Primary Role)

- Generating Python implementations of researcher-specified rules
- Converting plain language specifications to executable code
- All code validated through pilot testing
- Researcher defines logic; LLM generates implementation

Data Extraction Assistance (Secondary Role - Section 3.7.2)

- Suggesting data field values from article text
- Extracting structured information from unstructured text
- All suggestions require researcher verification

- Only verified values stored

Analysis Assistance (Tertiary Role - Section 3.8.3)

- Summarizing long articles to aid navigation
- Suggesting thematic clusters from keywords
- All suggestions validated through full-text reading

Decision Authority

Task	LLM Role	Researcher Role	Who Decides?
Define screening rules	None	Specifies logic	Researcher
Generate Python code	| | Generate code |	Validates behavior	Researcher
Execute screening	None (Agent runs code)	Reviews uncertain cases	Researcher
Data extraction	Suggests values	Verifies all values	Researcher
Risk of bias rating	Flags keywords	Assigns ratings	Researcher
Thematic coding	Suggests clusters	Validates themes	Researcher

LLMs do NOT

- Make inclusion/exclusion decisions
- Make final data extraction determinations
- Assign risk of bias ratings
- Interpret or modify researcher-defined rules autonomously

Quality Assurance Through Transparent Rule Specification

Instead of post-hoc output validation, quality is assured through:

1. Rule Transparency

- All rule specifications documented in GitHub
- Community can inspect and critique logic
- Peer reviewers can verify rules match objectives
- Published code provides full transparency

2. Version Control

- Every rule change documented with commit message
- Full audit trail of methodology evolution

- Enables reviewers to understand decision process
- Example commit: `"Added 'transparency' to auditability synonyms after pilot revealed semantic gap"`

3. Reproducibility Testing

- Any researcher can clone repository and re-run analysis
- Should produce identical results (deterministic execution)
- Non-reproducibility indicates implementation error → flag for debugging

4. Edge Case Documentation

- Studies flagged as "uncertain" logged in `logs/uncertain_cases/`
- Researcher decisions recorded with rationale
- Enables post-hoc review of judgment calls

5. Pilot Testing Protocol

- All rules tested on 50+ record sample before full deployment
- False positives/negatives identified and addressed
- Iterative refinement until satisfactory performance
- Performance metrics documented

Example: Rule Development Workflow

Iteration 1

> Researcher specifies: "Include studies about electronic voting security"
>
> → LLM generates code checking for "electronic voting" AND "security"
>
> → Pilot test reveals false positives (cybersecurity papers not about elections)
>
> → Researcher refines: "Include if 'electronic voting' AND 'security' AND ('election' OR 'ballot')"

Iteration 2

> → LLM generates updated code
>
> → Pilot test shows improved precision but missed "e-voting" variant
>
> → Researcher refines: "Include 'electronic voting' OR 'e-voting' OR 'internet voting'"

Iteration 3

> → LLM generates final code
>
> → Pilot test shows acceptable precision (>90%) and recall (>85%)
>
> → Rules frozen and applied to full dataset
>
> → All changes logged in Annex H with commit hashes

This iterative process ensures rules accurately capture researcher intent before full deployment.

Documentation in Final Review

The completed systematic review will include a dedicated methodology section:

Automation Documentation

- Description of rule-based architecture
- Link to GitHub repository with all code and specifications
- Complete version history of rule refinements
- All logs documenting classifications and decisions

Summary Statistics

- PRISMA flow diagram showing records at each stage
- Records auto-classified at title/abstract stage
- Records auto-excluded due to ambiguous peer review
- Number of rule refinement iterations during pilot
- Edge cases requiring researcher judgment

Transparency Statement

All automated classification steps are deterministic and fully reproducible via published rule specifications and code. The complete audit trail is available in the GitHub repository [URL], including all version history from protocol registration through review completion.

Researchers can reproduce the workflow by:

1. Cloning the repository
2. Installing dependencies (requirements.txt)
3. Running screening pipeline with archived dataset
4. Comparing outputs to published results

Any discrepancies indicate implementation errors and should be reported as issues.

Key Principle

The ELIS SLR Agent is not an AI decision tool. It is a **deterministic implementation of researcher-defined rules** that happen to be coded by LLMs rather than manually. The scientific validity depends on rule quality (researcher responsibility), not code syntax (LLM responsibility).

3.5 Data Items & Data Contract

This section defines the data points required for the synthesis (Data Items) and introduces the Frozen Data Contract principle for the Systematic Literature Review (SLR).

3.5.1 Data Items

The data extraction form will target two primary categories of information from the included studies:

Bibliometric Data

Item	Rationale for Extraction
Unique Study ID (study_id)	Generated by the Agent upon consolidation to ensure unique reference traceability throughout the SLR.
DOI & Permanent URL	Essential for digital full-text retrieval and permanent referencing.
Publication Year & Venue	Required for chronological analysis and assessment of source credibility (Critical Appraisal).

Substantive Data

Item	Rationale for Extraction
Country/Context of Study	Required for synthesis concerning geographical applicability and policy context.

Intervention Type	Specifies the electoral integrity strategy or technology evaluated (e.g., *Risk-Limiting Audit*, *Voter Verified Paper Audit Trail*).
Outcome Measures	The specific results reported (e.g., public trust scores, audit failure rates, cost analysis).
Study Design	Classification of the empirical methodology (e.g., *Quantitative Experiment*, *Qualitative Case Study*).

3.5.2 The Frozen Data Contract Principle

The **Data Contract** is the technical specification that ensures the integrity and consistency of the data used throughout the ELIS project. It is defined as a formal schema stored in the repository (e.g., ``schemas/elis_data_contract_v2.0.json``) and is programmatically enforced by the **ELIS SLR Agent**.

The purpose of this Contract is threefold: **Auditability**, **Reproducibility**, and **SLR Consistency**.

Contract Principles:

1. **Strict Type Enforcement:** Every Data Item listed in 3.5.1 has a defined data type (e.g., study_id is a unique string, publication_year is an integer, Outcome Measures is a structured object). The Agent will use Python validation libraries (e.g., Pydantic or Cerberus) to enforce these types upon ingestion.
2. **Schema Immutability (Frozen State):** Once Protocol v2.0 is registered, the Data Contract is frozen. Any subsequent required change to the Contract (e.g., adding a new field or changing a data type) constitutes a **Major Protocol Amendment** (vX.0) and requires approval and documentation in the **Amendments Log** (Annex H.1).
3. **Quarantine Policy:** Records retrieved from API sources that fail to satisfy the Contract (e.g., null value in a required field like DOI) will be automatically quarantined. These records are isolated from the main dataset and logged in the **Quarantine Log** for research assessment, ensuring that data quality degradation never occurs within the primary synthesis dataset.

3.6 Critical Appraisal

This section details the critical appraisal process. The successful delivery and integrity of the review are achieved by leveraging the **ELIS SLR Agent** to manage data processing, ensure consistency, and provide a full audit trail for the single-reviewer methodology.

3.6.1 Risk of Bias (RoB) Assessment (Automation-Enabled Rigour)

The methodological quality of each included empirical study will be assessed solely by the Principal Investigator (PI). The successful delivery of a high-quality assessment is guaranteed by the Agent's automated support functions, which ensure **internal consistency** and **procedural transparency**.

a) **Tool Development:** A **Modified Risk of Bias Tool** will be developed by the PI, adapted from established SLR frameworks (e.g., ROBIS or ROBINS-I). This tool will be tailored to address specific biases common in the intersection of technology, policy, and social science (e.g., bias in intervention deployment, selection bias in pilot groups).

b) **Consistency Audit by Agent:** The **ELIS SLR Agent** replaces the function of a second reviewer by performing a continuous **Consistency Audit** on the PI's decisions, thereby ensuring a verifiable and stable assessment process.

- **Audit Mechanism:** The Agent monitors the PI's scoring patterns (e.g., the average RoB score over time, or the distribution of scores across specific bias domains).
- **Temporal Drift Detection:** If the Agent detects a statistically significant **drift in scoring consistency** between early and later assessments—indicating potential fatigue or changing criteria—it automatically triggers a **Consistency Alert**. This necessitates a mandatory review by the PI of the assessment criteria and a recalibration against the pilot sample.

c) **Audit Trail and Guarantee:** All domain scores and the final rating (**Low**, **Moderate**, or **High** RoB) are logged as **Data Items** (Section 3.5), including the date and time. This comprehensive, machine-audited log ensures that the assessment process, while single-person executed, is **fully auditable** and demonstrably **consistent** throughout the project lifecycle.

3.6.2 Confidence in the Body of Evidence

The certainty, or confidence, in the overall findings related to each primary outcome will be formally evaluated by the PI, incorporating the concepts from the former Section 3.8.

a) **Quantitative Synthesis:** For outcomes derived from quantitative data (e.g., audit failure rates, public trust scores), the **GRADE (Grading of Recommendations Assessment, Development and Evaluation)** framework will be used. The evidence will be rated across domains including risk of bias (derived from 3.6.1), inconsistency, indirectness, imprecision, and publication bias.

b) **Qualitative Synthesis:** For findings derived from qualitative studies (e.g., thematic analysis of interviews, case studies), the **GRADE-CERQual (Confidence in the Evidence from Reviews of Qualitative research)** approach will be applied. This assesses methodological limitations, coherence, adequacy of data, and relevance.

c) **Output:** The final confidence ratings (High, Moderate, Low, or Very Low) for each outcome are logged and will be used to structure the discussion and recommendations (Section 3.8).

3.7 Data Extraction Process

The data extraction process is designed for maximum efficiency and auditability, leveraging the **ELIS SLR Agent** to structure the input and perform initial data capture, which is essential for the single-reviewer methodology.

3.7.1 Data Collection Instrument

Data extraction will be performed using a structured, standardised form derived directly from the **Data Items & Data Contract** (Section 3.5). The instrument ensures consistency by requiring fields to be populated according to the defined schema (e.g., date formats, required string lengths, fixed choice categories).

a) **Pilot Extraction:** The Principal Investigator (PI) will conduct a pilot extraction on a sample of five included studies to refine the instruction manual for the form and ensure the fields accurately capture the required information.

b) **Extraction Logging:** All extracted data, along with provenance details (study ID, extraction date), is logged by the Agent and immediately checked against the **Frozen Data Contract** (Section 3.5) before being stored in the final dataset.

3.7.2 AI-Assisted Data Capture and Governance

The PI's workload is substantially reduced by the **ELIS SLR Agent** using pre-trained Large Language Models (LLMs) to perform initial, high-volume data capture. This automation is subject to strict governance rules.

a) **AI Function:** General purpose AI tools (e.g., ChatGPT, Claude.ai, NotebookLM) are integrated by the Agent to **suggest data field values** (e.g., classifying study design, extracting the primary outcome measure from the abstract).

b) **Human Verification Mandate:** All AI-suggested values are classified as **drafts**. The PI is mandated to verify and confirm or edit every single field before final submission. The final, confirmed value is the only value stored in the clean dataset, ensuring that the PI remains the single, accountable decision-maker.

c) **AI Log:** A separate **AI Audit Log** is maintained, documenting which fields were auto-suggested, the original AI suggestion, and the final researcher-verified value. This log ensures transparency and allows for the periodic assessment of the LLM’s performance accuracy.

3.7.3 Adaptive Quality Assurance (AQA) Policy

To maintain confidence in the researcher's consistency across the extraction phase, the Agent employs an Adaptive Quality Assurance (AQA) policy:

a) **Error Rate Monitoring:** The Agent continually tracks the PI's **error rate**—the percentage of fields that the PI edits or corrects after the AI has provided a suggested value.

b) **Recalibration Trigger:** If the PI's correction rate exceeds a predefined threshold (e.g., 15% error rate across a rolling 20-study sample), the Agent triggers an **AQA Alert**. This requires the PI to pause extraction and formally recalibrate by re-reading the extraction manual and re-extracting a small validation sample. This mechanism guarantees that the extraction consistency is maintained throughout the project.

3.8 Data Synthesis

Data synthesis will be conducted in parallel for quantitative and qualitative studies using structured, reproducible procedures. The process is feasible for a single PI because the **ELIS SLR Agent** manages the data structuring, automates preliminary clustering tasks, and maintains a rigorous audit trail of all analytical decisions.

3.8.1 Data Preparation and Synthesis Instrument

Before analysis begins, the Agent ensures the data set is ready for synthesis, guaranteeing consistency and adherence to the **Frozen Data Contract** (Section 3.5).

a) **Data Structuring and Segregation:** The Agent exports the final, clean dataset into an analytical format (e.g., `JSONL`). It automatically segregates studies based on their **Study Design** (Section 3.5) into distinct datasets for quantitative and qualitative analysis, enabling the PI to proceed immediately with parallel synthesis.

b) **Synthesis Instrument:** The **ELIS Agent** will use specialized analytical software (e.g., Python with Pandas/SciPy for quantitative, NVivo/Dedoose for qualitative) to process the Agent-prepared data.

3.8.2 Quantitative Data Synthesis

Quantitative studies (e.g., surveys, comparative experiments) will be summarized using descriptive statistics, frequency distributions, and tabulated outcome types.

a) Descriptive Summary

Studies will be summarised using descriptive statistics, including frequency distributions, means, and standard deviations for key variables (e.g., public trust scores, audit failure rates, cost). Tabulated outcome types will be used to present data clearly.

b) Aggregative Analysis

Where appropriate and feasible (i.e., if study populations, interventions, and outcome measures are sufficiently comparable), trends in intervention effectiveness or specific outcome domains will be aggregated. A narrative summary of heterogeneity will be provided if formal meta-analysis is deemed inappropriate due to high clinical, methodological, or statistical diversity.

c) Design-Outcome Attribution Analysis

To address the primary research question (What strategies have been shown to improve integrity) the synthesis will systematically examine the strength of attribution between specific design features and observed outcomes.

For each outcome reported, the data extraction includes a Design Link field (Annex D) indicating whether the outcome is directly attributable to a specific intervention or system feature. This field enables the review to distinguish:

- **Direct attribution** (Design Link = Yes): Outcomes clearly linked to specific design features
  - Example: "Discrepancy rate decreased after VVPAT implementation"
  - The outcome (discrepancy rate) is directly testing the intervention (VVPAT)
- **Indirect or confounded** (Design Link = No): Outcomes influenced by multiple factors
  - Example: "Public trust increased during election with new voting machines"

  - **Trust may be affected by machines, media coverage, campaign quality, etc.**

Synthesis Approach

Findings will be stratified by attribution strength level:

Attribution Level	Synthesis Approach	Confidence in Causality
Strong Direct Link	Outcomes directly testing specific design features; synthesized with high confidence in mechanism	High (if study design supports causality)
Moderate Link	Outcomes plausibly related to intervention but other factors present; synthesized with caveats	Moderate
Weak/No Link	Outcomes reported but not clearly attributable to design features; synthesized descriptively only	Low

This stratification ensures the review distinguishes between:

- Evidence that specific mechanisms work (e.g., "Risk-limiting audits detect errors effectively")
- Evidence that broader contexts matter (e.g., "Trust varies by institutional setting")

Reporting

The final synthesis will include:

- **Primary findings tables** reporting outcomes with strong design links (direct evidence on "what works")
- **Contextual factors tables** reporting outcomes with weak links (describing conditions and moderators)
- **Narrative synthesis** explaining how design features, context, and implementation interact This approach addresses the primary research question while maintaining transparency about the strength of evidence linking specific strategies to observed outcomes.

This approach addresses the primary research question while maintaining transparency about the strength of evidence linking specific strategies to observed outcomes.

3.8.3 Qualitative Data Synthesis and Consistency Audit

Qualitative studies will be analysed using a structured Thematic Synthesis approach, with the Agent ensuring the PI's analysis is consistent and auditable, compensating for the single-reviewer limitation.

a) Thematic Synthesis

Findings will be clustered inductively around core conceptual dimensions relevant to the research questions (e.g., transparency, auditability, public trust). Themes will be derived and refined iteratively by the PI.

b) LLM Assistance in Clustering

AI tools (e.g., ChatGPT, Claude.ai, NotebookLM) are integrated by the Agent to assist the single PI with large-scale data organisation. The LLM function is restricted to suggesting thematic clusters, summarising lengthy text fields, or identifying co-occurring variables within qualitative reports.

Design-Mechanism Analysis

For qualitative studies, the synthesis will examine how design features interact with implementation contexts to produce outcomes. The **Design Link** field helps identify:

- **Mechanism-focused studies:** Case studies explicitly tracing how a design feature (e.g., observer access protocols) produces outcomes (e.g., irregularity detection)
- **Context-focused studies:** Studies describing electoral processes where design features are present but outcomes reflect broader institutional factors

Qualitative synthesis will use process tracing logic to map:

```
Design Feature → Implementation Context → Mechanisms → Outcomes
```

For example:

- VVPAT (design) → Implemented with mandatory recount rules (context) → Enables verification (mechanism) → Detects discrepancies (outcome)

This analysis identifies not just whether interventions work, but **how** and **under what conditions** they produce effects.

c) Validation Mandate

This assistance is purely an efficiency tool. All suggested thematic clusters, summaries, and interpretations will be verified, coded, and validated by the PI through close reading and analytical judgment. No AI-generated interpretation will be used in the final synthesis without explicit, documented validation by the PI, ensuring the PI retains full accountability for the analytical rigour.

Transparency Statement

All automated classification steps are deterministic and fully reproducible via published rule specifications and LLM-generated code. Any researcher can replicate the workflow by re-specifying rules to LLMs using documented specifications, or directly execute the published code to reproduce identical results. The complete audit trail is available in the GitHub repository, including all version history from protocol registration through final review completion.

4. Ethical Considerations

This review does not involve human subjects research or any personal identifiable data. Therefore, it does not require formal ethical approval from an institutional review board. Nonetheless, the review adheres to general principles of research integrity, especially given the use of automation and AI tools in the methodology:

- **Human Oversight:** All final decisions in the review (study inclusion/exclusion, data interpretations, conclusions) are made by the principal investigator, not by AI. At no point are AI recommendations accepted without review.
- **Transparency:** Every use of an AI or automation tool in the process is logged and documented. For example, if an LLM is used to summarize an article, that summary and the prompt used are saved to the project log (see Annex F on automation workflow). This ensures an audit trail of how AI contributed.
- **Accountability:** The project follows the UK Research Integrity Office (UKRIO) Principles for the use of AI in research, ensuring responsible integration of AI and reproducibility of results. All outputs of the ELIS SLR Agent and other AI tools are subject to verification, and the lead researcher accepts responsibility for the integrity of the review's findings.
- **Data Handling:** All reference data and results will be handled in accordance with open science best practices. Since no sensitive personal data is involved, the main ethical consideration is proper citation and avoidance of plagiarism in data synthesis, which will be diligently observed.

5. Timeline and Workload

5.1 Estimated Timeline and Workload

This section provides estimated time requirements for each review stage to ensure feasibility and transparency. Estimates are based on pilot testing, systematic review methodology literature (Borah et al., 2017), and the specific automation tools employed in this review.

Stage	Activities	Estimated Hours	Assumptions
1. Protocol Finalization	• Multi-AI review • Revisions • Registration	40 hours	Includes supervisor feedback iterations
2. Search Execution	• Database searches (8 sources) • Export and documentation • Initial deduplication	16 hours	API automation reduces manual effort; ~2-3 hours per database
3. Screening Calibration	• Database searches (8 sources) • Rule refinement • Agent testing and validation	20 hours	Iterative rule development until precision >90%, recall >85%
4. Title / Abstract Screening	• Agent execution (automated) • Review "uncertain" cases • Documentation	30 hours	Assumes ~3,000 records after deduplication; Agent handles 90-95%; Researcher reviews 5-10% (~150-300 records) @ 10 min/record = 25-50 hours

5. Full-Text Retrieval	• Obtain PDFs/full texts • Organize files • Flag inaccessible items	12 hours	Assumes 200-300 records for full-text; Most accessible via institutional access
6. Full-Text Screening	• Read and apply eligibility criteria • Document exclusions • Final inclusion decisions	80 hours	Assumes 250 full texts reviewed; ~20 min/article = 83 hours

7. Data Extraction	• Pilot extraction (10 studies)
	• Full extraction with AI assistance
	• Verification and quality checks
		100 hours	Assumes 80-120 included studies;
			AI suggests values (saves ~30% time);
			Researcher verifies all;
			~50-75 min/study

8. Risk of Bias Assessment	• Develop assessment tool • Pilot testing • Assess all included studies	60 hours	~30-45 min/study for 120 studies
9. Data Synthesis	• Quantitative summary • Qualitative thematic analysis • Evidence confidence ratings • Cross-cutting analysis	80 hours	Iterative synthesis with AI-assisted clustering
10. Manuscript Preparation	• Writing • Figure/table generation • Internal review • Revisions	100 hours	Full manuscript (~8,000-10,000 words)
11. Peer Review Response	• Address reviewer comments • Revisions • Resubmission	40 hours	Assumes one major revision round
	TOTAL =	**578 hours**	**Approximately 14.5 weeks full-time** **or 6-8 months part-time**

Key Assumptions

1. Automation Impact

- Title/abstract screening: 85-90% time savings via ELIS SLR Agent
- Data extraction: 30% time savings via AI-assisted field suggestions

- Without automation: Estimated +200 hours (total would be ~780 hours)

2. Included Studies Volume

- Conservative estimate: 80-120 studies included after full-text screening
- Based on pilot searches and comparable reviews in electoral integrity domain

3. Researcher Expertise

- Principal investigator familiar with electoral systems and systematic review methods
- Learning curve for new automation tools included in calibration phase

4. Infrastructure

- API access to all databases confirmed
- Institutional library access for full-text retrieval
- Computing resources adequate for ELIS SLR Agent

5.2 Workload Monitoring and Adaptation

Real-Time Tracking

Actual time spent on each stage will be logged in `` `logs/workload_tracking.json` `` to:

- Validate estimates
- Identify bottlenecks
- Inform future systematic reviews using similar methods
- Provide data for Living Review feasibility assessment (Section 7)

Example log entry

```
json
{
    "stage": "full_text_screening",
    "date_range": "2026-02-15 to 2026-03-10",
    "total_hours": 85,
    "records_processed": 247,
    "avg_minutes_per_record": 20.6,
    "notes": "Higher than estimate due to complex eligibility decisions in
   comparative studies"
}
```

Adaptation Triggers

If actual workload significantly exceeds estimates (>25% over):

1. **Title/Abstract Screening > 40 hours**

- Refine screening rules to reduce "uncertain" classifications
- Consider adjusting precision/recall balance

2. **Full-Text Screening >100 hours**

- Review eligibility criteria for ambiguity
- Consider whether scope too broad

3. **Data Extraction >125 hours**

- Simplify extraction form if excessive detail
- Improve AI prompts for better suggestions

Reporting

Final review manuscript will include actual workload data in methodology section:

- Validates feasibility of approach
- Informs replication efforts
- Provides evidence for automation benefits
- Feeds into Living Review decision (Section 7.2 Criterion B: "Monthly update ≤20 hours")

5.3 Living Review Workload Implications

Relevance to Section 7 (Future Directions)

The workload tracking from this initial SLR provides empirical data for assessing Living Review feasibility.

Baseline Data Needed

- Average time per record at each stage
- Automation reliability (precision/recall)
- Ratio of new included studies to total screened

LSR Feasibility Calculation (Example)

If initial review finds:

- 3,000 records screened → 120 included (4% inclusion rate)
- Avg 20 min/full-text screening
- Avg 60 min/data extraction per included study

Then estimated monthly update workload:

```
Assume 50 new records/month (based on field publication velocity)
→ Title/abstract: ~2 hours (mostly automated)
→ Full-text candidates: 50 × 4% = 2 studies
→ Full-text screening: 2 × 20 min = 40 min
→ Data extraction: 2 × 60 min = 120 min
→ Quality checks & synthesis update: ~60 min
TOTAL: ~5 hours/month
```

- Under 20-hour threshold → LSR potentially feasible
- If >20 hours → Reassess or use quarterly updates instead

This empirical approach replaces speculation with data-driven decision-making.

Reference: Borah, R., Brown, A. W., Capers, P. L., & Kaiser, K. A. (2017). Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open, 7*(2), e012545.

6. Review Process

This systematic review is conducted by a single researcher with methodological transparency provided through the ELIS SLR Agent automation tool and GitHub version control. All screening decisions and intermediate products are versioned and publicly accessible, enabling external audit or replication. If collaborators join during the review, roles will be documented in protocol amendments and reflected in GitHub commit history.

7. Future Directions: Living Review Potential

7.1 Rationale for Post-Review Assessment

This protocol is designed for a standard systematic literature review (SLR) covering 1990–2026. However, the review infrastructure has been developed with features that could support future living review updates:

- **Automated search workflows** via ELIS SLR Agent and database APIs
- **Structured data formats** enabling incremental additions
- **Version-controlled methodology** providing full audit trail
- **Reproducible screening rules** reducing manual re-work

Rather than commit to a living review model without empirical evidence of feasibility, this protocol takes a **prove-then-scale approach**: complete a high-quality initial review, then assess whether living updates are justified and sustainable.

7.2 Feasibility Assessment Criteria

Following completion of the initial systematic review, a living review model would only be pursued if **all** of the following criteria are met:

A. Evidence Activity Criteria

- **Publication velocity:** ≥10 relevant studies published per year (based on final year of initial review)

- **Field evolution:** Evidence of new interventions, technologies, or contexts not covered in initial review
- **Stakeholder demand:** Documented interest from electoral management bodies, researchers, or policymakers in ongoing synthesis

B. Methodological Feasibility Criteria

- **Automation reliability:** ELIS SLR Agent screening rules demonstrate ≥90% precision and ≥85% recall in initial review
- **Workload sustainability:** Monthly update cycle estimated at ≤20 hours based on final 6 months of initial review
- **Quality maintenance:** Risk of bias assessment and data extraction processes proven consistent across initial review

C. Resource Availability Criteria

- **Researcher capacity:** Principal investigator or designated team member available for ongoing monthly updates
- **Infrastructure access:** Continued API access to all bibliographic databases
- **Institutional support:** Commitment from Imperial College or successor institution for repository hosting and dissemination

7.3 Living Review Activation Protocol

If feasibility criteria are met, living review activation would follow this sequence:

Phase 1: Pilot Testing (3 months)

1. Conduct 3 monthly update cycles as test
2. Document actual time required per update
3. Assess automation performance on new records
4. Evaluate stakeholder engagement with updates

Phase 2: Protocol Amendment (Month 4)

1. Submit formal protocol amendment to register living review status
2. Specify update frequency (monthly, quarterly, or other)
3. Define sustainability triggers (workload limits, pause criteria)
4. Establish governance for ongoing review decisions

Phase 3: Living Review Launch (Month 5+)

1. Implement regular update schedule
2. Version all outputs with update dates
3. Provide timestamped snapshots for citation
4. Monitor sustainability indicators continuously

7.4 Update Cycle Methodology (when activated)

Monthly Update Process:

1. **Automated Search (Day 1-2):**

- ELIS SLR Agent queries all 7 databases via APIs
- Retrieves only records published since last update
- De-duplicates against existing database

2. **Screening (Day 3-5):**

- Agent applies frozen screening rules to new records
- Flags "include," "exclude," and "uncertain" classifications
- Researcher reviews uncertain cases only (~5-10% of records)

3. **Full-Text Review (Day 6-12):**

- Researcher obtains and screens full texts of included records
- Applies same eligibility criteria as initial review
- Documents exclusions in log

4. **Data Extraction & Quality Assessment (Day 13-20):**

- Extract data using same forms as initial review
- Conduct risk of bias assessment
- Update synthesis dataset

5. **Output Update (Day 21-25):**

- Regenerate synthesis tables and figures
- Update evidence summaries
- Publish update report if ≥5 new studies included

6. **Version Control (Day 26-30):**

- Commit all changes to GitHub
- Update DOI-versioned dataset on Zenodo/OSF
- Archive previous version for reproducibility

7.5 Sustainability and Exit Criteria

Sustainability Monitoring

Living review updates would be monitored monthly for:

- Average hours per update cycle
- Number of new studies identified
- Citation/usage metrics for review outputs

Pause Triggers

The living review would be paused if any of:

- Monthly workload consistently exceeds 20 hours for 3+ months
- Fewer than 2 relevant studies identified per month for 6+ months
- API access lost for ≥2 databases
- Principal investigator unavailable without designated replacement

Termination Criteria

The living review would be terminated if:

- Field reaches evidence saturation (no new study types/contexts for 12+ months)
- Automation tools become unmaintainable
- Stakeholder interest declines (low citation/usage for 12+ months)

Any pause or termination would be documented via protocol amendment and communicated via review registration platform.

7.6 Reporting and Dissemination (If Activated)

Update Reports

- **Major updates** (≥10 new studies): Full updated review manuscript submitted for publication
- **Minor updates** (2-9 new studies): Update brief published on OSF with revised evidence tables
- **Null updates** (0-1 new studies): Update log only, no formal report

Version Citation

Each living review version would receive unique identifier:

ELIS Review v1.0 (Initial SLR, 1990-2026, published [date])
ELIS Review v1.1 (Update 1, through [month/year])
ELIS Review v1.2 (Update 2, through [month/year])

Timestamped Snapshots

Quarterly snapshots archived with DOI for stable citation in policy documents or subsequent research.

8. Annexes

- **Annex A** – PRISMA-P 2015 Checklist
- **Annex B** – Search Strings
- **Annex C** – Inclusion/Exclusion Log Template
- **Annex D** – Data Extraction Form
- **Annex E** – Evidence Quality Assessments

- **Annex F** – Automation Workflow & AI
- **Annex G** – Evidence Gap Confirmation and Supporting Literature *(in progress)*
- **Annex H** – Rule Development Log Template (NEW in v2.0)

AI Disclaimer

All AI-assisted stages are supervised by principal researcher. All automation steps are logged for transparency and auditability using the AI Agents Log Templates template in Annex F.

Annex A – PRISMA-P 2015 Checklist

The table below demonstrates how this protocol complies with each item of the PRISMA-P (2015) checklist for systematic review protocols. Each checklist item is addressed in the protocol, with section references provided:

Section / topic		Item #	Checklist item
ADMINISTRATIVE INFORMATION			
Title		1	
	Identification	1a	Identify the report as a protocol of a systematic review
	Update	1b	If the protocol is for an update of a previous systematic review, identify as such
Registration		2	If registered, provide the name of the registry (e.g., PROSPERO) and registration number
Authors		3	
	Contact	3a	Provide name, institutional affiliation, and e-mail address of all protocol authors; provide physical mailing address of corresponding author
	Contributions	3b	Describe contributions of protocol authors and identify the guarantor of the review
Amendments		4	If the protocol represents an amendment of a previously completed or published protocol, identify as such and list changes; otherwise, state plan for documenting important protocol amendments
Support		5	
	Sources	5a	Indicate sources of financial or other support for the review
	Sponsor	5b	Provide name for the review funder and/or sponsor
	Role of sponsor/funder	5c	Describe roles of funder(s), sponsor(s), and/or institution(s), if any, in developing the protocol
INTRODUCTION			
Rationale		6	Describe the rationale for the review in the context of what is already known
Objectives		7	Provide an explicit statement of the question(s) the review will address with reference to participants, interventions, comparators, and outcomes (PICO)
METHODS			
Eligibility criteria		8	Specify the study characteristics (e.g., PICO, study design, setting, time frame) and report characteristics (e.g., years considered, language, publication status) to be used as criteria for eligibility for the review
Information		9	Describe all intended information sources (e.g., electronic

sources			databases, contact with study authors, trial registers, or other grey literature sources) with planned dates of coverage
Search strategy		10	Present draft of search strategy to be used for at least one electronic database, including planned limits, such that it could be repeated
Study records		11	
	Data management	11a	Describe the mechanism(s) that will be used to manage records and data throughout the review
	Selection process	11b	State the process that will be used for selecting studies (e.g., two independent reviewers) through each phase of the review (e.g., screening, eligibility, and inclusion in meta-analysis)
	Data collection process	11c	Describe planned method of extracting data from reports (e.g., piloting forms, done independently, in duplicate), any processes for obtaining and confirming data from investigators
Data items		12	List and define all variables for which data will be sought (e.g., PICO items, funding sources), any pre-planned data assumptions and simplifications
Outcomes and prioritization		13	List and define all outcomes for which data will be sought, including prioritization of main and additional outcomes, with rationale
Risk of bias in individual studies		14	Describe anticipated methods for assessing risk of bias of individual studies, including whether this will be done at the outcome or study level, or both; state how this information will be used in data synthesis
Data		15	
	Synthesis	15a	Describe criteria under which study data will be quantitatively synthesized
		15b	If data are appropriate for quantitative synthesis, describe planned summary measures, methods of handling data, and methods of combining data from studies, including any planned exploration of consistency (e.g., I^2, Kendall's tau)
		15c	Describe any proposed additional analyses (e.g., sensitivity or subgroup analyses, meta-regression)
		15d	If quantitative synthesis is not appropriate, describe the type of summary planned
Meta-bias(es)		16	Specify any planned assessment of meta-bias(es) (e.g., publication bias across studies, selective reporting within studies)
Confidence in cumulative evidence		17	Describe how the strength of the body of evidence will be assessed (e.g., GRADE)

Annex B – Search Strings

This annex provides the finalized search strategies for each information source listed in Section 3.2.1. Each search string has been tailored to the syntax and capabilities of the respective database API. All search strings target three conceptual components: (1) electoral integrity/voting system terms, (2) specific intervention mechanisms, and (3) empirical study indicators.

1. Scopus (Title/Abstract/Keyword search)

API Endpoint: https://api.elsevier.com/content/search/scopus
Query Field: TITLE-ABS-KEY (searches title, abstract, and author keywords)
Date Filter: PUBYEAR > 1989 AND PUBYEAR < 2026
Search String:

```
TITLE-ABS-KEY((
  ("electoral integrity" OR "election integrity" OR "voting system security" OR
"ballot auditability" OR "electronic voting security" OR "ballot auditability"
OR "voter trust")
  AND
  ("audit" OR "VVPAT" OR "voter-verified paper audit trail" OR "paper trail" OR
"verifiability" OR "blockchain" OR "transparency" OR "biometric" OR "end-to-end"
OR "risk-limiting audit" OR "parallel vote tabulation" OR "cryptographic audit")
  AND
  ("empirical" OR "evaluation" OR "case study" OR "experiment" OR "comparative")
))
```

Rationale: Scopus provides comprehensive metadata for peer-reviewed content. The three-part Boolean structure ensures retrieval of studies combining electoral context + specific mechanisms + empirical evidence. Keywords selected based on SPIDER framework (Section 2.3) and benchmark validation results.

2. Web of Science (Topic search)

API Endpoint: https://api.clarivate.com/apis/wos-starter/v1/documents
Query Field: TS (Topic search - includes title, abstract, keywords, Keywords Plus)
Date Filter: PY=1990-2025
Search String:

```
TS=(
  ("electoral integrity" OR "election security" OR "voting system" OR "ballot
integrity" OR "voting system security" OR "auditability")
  AND
  ("audit" OR "VVPAT" OR "verifiability" OR "transparency" OR "blockchain" OR
"biometric" OR "risk-limiting")
```

```
  AND
  ("empirical" OR "evaluation" OR "study" OR "analysis" OR "experiment")
)
```

Note: Web of Science `TS` field searches title, abstract, and keywords. The query includes variations like *experimental* to capture studies even if they do not self-describe as empirical.

3. IEEE Xplore

API Endpoint: https://ieeexploreapi.ieee.org/api/v1/search/articles
Query Fields: ("Document Title" OR "Abstract" OR "Index Terms")
Date Filter: publication_year:[1990 TO 2025]
Search String:

```
("electronic voting" OR "e-voting" OR "voting system" OR "digital ballot")
  AND ("security" OR "integrity" OR "auditability" OR "verifiability" OR
"cryptographic")
  AND ("VVPAT" OR "audit" OR "blockchain" OR "biometric" OR "end-to-end" OR
"transparency")
  AND ("evaluation" OR "case study" OR "experiment" OR "empirical" OR
"implementation")
```

Note: IEEE Xplore's interface was used to apply year filters (1990-2025) and to search within metadata and full text. Technical keywords like *blockchain* and *biometric* were included given their relevance in recent voting technology discourse.

4. Semantic Scholar

API Endpoint: https://api.semanticscholar.org/graph/v1/paper/search
Query Field: query (searches titles, abstracts, and entities)
Date Filter: year:1990-2025
Search String:

```
("electoral integrity" OR "election security" OR "voting system" OR "ballot
auditability")
  AND ("security" OR "integrity" OR "auditability" OR "verifiability" OR
"cryptographic" OR "audit" OR "VVPAT" OR "verifiability" OR "blockchain" OR
"transparency" OR "biometric" OR "end-to-end")
  AND ("evaluation" OR "case study" OR "experiment" OR "empirical" OR
"implementation")
```

Rationale: Semantic Scholar uses AI-enhanced indexing and citation graphs. Simplified Boolean structure works well with S2's semantic matching. Query retrieves both computer science and interdisciplinary political science literature.

5. OpenAlex

API Endpoint: https://api.openalex.org/works
Query Field: search (searches title and abstract)
Date Filter: publication_year:1990-2025
Type Filter: type:article
Search String:

```
(electoral integrity OR election security OR voting system OR ballot
auditability)
AND
(audit OR VVPAT OR verifiability OR blockchain OR transparency OR biometric OR
risk-limiting)
AND
(empirical OR evaluation OR study OR analysis OR experiment OR comparative)
```

Rationale: OpenAlex provides comprehensive open bibliographic data. Search syntax is simpler than traditional databases. Type filter ensures articles are prioritized. Concept tagging in OpenAlex metadata aids in retrieving interdisciplinary studies.

6. CrossRef

API Endpoint: https://api.crossref.org/works
Query Field: query.title AND query.abstract (when available)
Date Filter: from-pub-date:1990, until-pub-date:2025
Type Filter: type:journal-article
Search String:

```
query.title=(electoral integrity OR election security OR voting OR ballot)
OR
query.abstract=(electoral integrity OR election security OR voting system)
AND
query=(audit OR VVPAT OR verifiability OR blockchain OR transparency OR
biometric)
```

Rationale: CrossRef provides DOI-registered metadata but limited full-text search. Query targets titles for broad retrieval, with abstract search for specificity. Used primarily for metadata enrichment and deduplication rather than primary discovery.

7. CORE

API Endpoint: https://api.core.ac.uk/v3/search/works
Query Field: q (searches full text when available, otherwise metadata)
Date Filter: yearPublished:>=1990 AND yearPublished:<=2025
Search String:

```
(
  (electoral integrity OR election security OR voting system OR ballot
auditability)
  AND
  (audit OR VVPAT OR verifiability OR blockchain OR transparency OR biometric)
  AND
  (empirical OR evaluation OR study OR case OR experiment OR analysis)
)
```

Rationale: CORE aggregates open access content from repositories worldwide. Full-text search capability increases recall. Query structure balances precision (three-component AND) with recall (broad term coverage). Particularly valuable for retrieving institutional repository content not indexed elsewhere.

8. Google Scholar (via Apify)

API: Apify Google Scholar Scraper (https://apify.com/marco.gullo/google-scholar-scraper)
Query Field: queries (standard Google Scholar search)
Date Filter: Custom range 1990-2025
Results Limit: 100 results per query (API limitation)
Search Queries (Multiple queries executed):

Query 1 (Broad):

```
"electoral integrity" OR "election security" audit OR VVPAT OR verifiability
```

Query 2 (Specific Mechanisms):

```
"voting system" AND ("risk-limiting audit" OR "paper trail" OR "blockchain
voting")
```

Query 3 (Empirical Focus):

```
"electronic voting" AND (evaluation OR "case study" OR "empirical study") AND
(security OR integrity)
```

Rationale: Google Scholar does not provide structured API like traditional databases. Apify scraper enables systematic retrieval. Multiple queries required due to 100-result limit per query. Queries designed based on benchmark validation showing Google Scholar contributed 10x improvement in retrieval rate. Broad query captures general literature; specific queries target mechanisms and empirical studies. Results deduplicated against other databases using DOI/title matching.

Benchmark Evidence

Validation against Darmawan & Setyadji (2021) demonstrated Google Scholar increased retrieval rate from 9% to 37.2%. Complete benchmark documentation available at: https://github.com/rochasamurai/ELIS-SLR-Agent/tree/benchmark/darmawan-2021/docs/benchmark

Search Execution Protocol

All searches executed via ELIS SLR Agent (Python automation):

1. **API Authentication**: Stored securely in environment variables
2. **Rate Limiting**: Compliant with each API's rate limits
3. **Error Handling**: Retry logic with exponential backoff
4. **Logging**: Each search logged with: database, query, date, result count, API response status
5. **Export Format**: Results exported as JSON with standardized metadata schema
6. **Deduplication**: Cross-database deduplication using DOI, title matching (50% keyword overlap threshold), and author matching
7. **Search Date**: [To be added upon execution - planned Q1 2026]

Documentation

Complete search logs including exact query strings, filters, API parameters, and result counts available in project GitHub repository: https://github.com/rochasamurai/ELIS-SLR-Agent

Notes on Query Development

Benchmark-Informed Design:

- Search strings optimized based on benchmark validation against Darmawan & Setyadji (2021)
- Keyword overlap matching (≥50% threshold) outperformed complex similarity algorithms
- Google Scholar integration critical for maximum retrieval (validated experimentally)

Iterative Refinement:

- Queries tested in pilot searches (November 2025)
- Precision and recall assessed against 50-study calibration sample
- Final queries balanced broad retrieval with manageable screening workload
- Any post-registration query modifications documented in amendments log (Section 1.4)

API-Specific Adaptations:

- Scopus: Advanced field-specific syntax (TITLE-ABS-KEY)
- WoS: Topic search with Keywords Plus expansion
- IEEE: Technical terminology emphasis
- Google Scholar: Multiple queries to overcome result limits
- Others: Simplified Boolean matching adapted to API capabilities

Reproducibility:

All search strings, API parameters, and execution logs versioned in GitHub repository to enable exact replication of search strategy.

Annex C – Inclusion/Exclusion Log Template

This annex describes how study selection decisions will be recorded, to ensure transparency at the full-text screening stage. For each study considered at full-text, the following information will be logged:

- **Study ID:** A unique identifier for each reference (e.g., an abbreviation or number, such as *ELIS2025-001*).
- **Citation Details:** Key reference information (author, year, title, source) for clarity.
- **Decision:** Inclusion or Exclusion status after full-text review.
- **Exclusion Reason (if excluded):** A brief reason chosen from a predefined list (e.g., *Population not in scope*, *No empirical data*, *Duplicate study*, *Outcome not relevant*, etc.). For example, *"Excluded – Not empirical (commentary piece)"*.
- **Notes:** Any additional notes or comments by the reviewer (e.g., *"Contains some relevant background, but no primary data"* or *"Pending second opinion on methodology"* if a consultation is needed).

All studies excluded after full-text screening will have an associated reason. The inclusion/exclusion log will be maintained in a spreadsheet or JSON format (managed via the ELIS SLR Agent pipeline) for consistency. An example entry is shown below:

Study ID	Reference (Author, Year)	Decision	Reason for Exclusion (if any)	Notes
ELIS2025-015	Doe *et al.*, 2018 – *Election Audits in X*	Exclude	No relevant intervention	Focuses on voter turnout only
ELIS2025-027	Smith, 2020 – *Electronic Voting Security*	Include	–	Proceed to data extraction

Table: Sample rows from the inclusion/exclusion log.

This log will be updated in real time as screening progresses. By the end of the selection process, it will provide a complete account of how many studies were included and excluded, and for what reasons. The finalized log (with all excluded references and reasons) will be appended to the eventual review report and made available online (e.g., as a supplemental file or in the GitHub repository).

Maintaining this log aligns with PRISMA reporting guidelines and facilitates the construction of the PRISMA flow diagram for study selection.

Annex D – Data Extraction Form

This annex presents the structured form that will be used to extract data from each included study, as well as an example of the JSON data structure that the ELIS SLR Agent will produce for each study. The form captures bibliographic details, context, intervention characteristics, outcomes, and quality indicators.

Data Extraction Fields and Definitions:

- **Study ID:** Unique identifier for the study (for cross-referencing with screening log). *Example:* ELIS2025_001.
- **Title:** Full title of the article or report.
- **Authors:** Lead author(s) and possibly et al. (for reference listing).
- **Publication Year:** Year the study was published.
- **Publication Type/Source:** Journal name or conference, etc., and whether it's peer-reviewed (Yes/No).
- **Country/Region Studied:** The country or countries where the study's data is from or applicable. e.g., *Brazil*, *Global comparative*, *EU (multi-country)*.
- **Electoral Modality:** What type of voting system is involved – *Electronic*, *Paper-based*, *Hybrid*, or other specific modality.
- **Intervention Type:** The specific integrity mechanism or strategy evaluated. e.g., *Risk-limiting audit*, *VVPAT implementation*, *Blockchain-based ledger*, *Transparency initiative*. If multiple, all are noted.
- **Study Design:** The methodological approach of the study. e.g., *Field experiment*, *Quasi-experiment (DiD)*, *Case study*, *Simulation*, *Survey research*.
- **Evaluation Method:** How outcomes were measured or evaluated. e.g., *Statistical audit*, *Survey of voters*, *Analysis of error logs*, *Interviews*.
- **Outcomes:** For each outcome reported, we will record:
  - Outcome **Type** – whether the outcome is considered Primary or Secondary in our review's context.
  - Outcome **Reported** – a short description of what specific outcome was measured (e.g., *discrepancy rate*, *voter confidence score*, *time to detect fraud*).
  - Outcome **Measure** – how that outcome was operationalized (e.g., *% of votes recounted that differ*, *Likert scale survey response*).
  - **Design Link** – a Yes/No flag indicating if the outcome is clearly linked to a specific system design feature or intervention (for example, an outcome of "error rate" would be "Yes" linked if the errors relate to the new voting machine design).

  - **Reviewer Notes on Outcome** – comments on the outcome's significance or the strength of evidence (e.g., *"Outcome is statistically significant; causal inference strong"* or *"Outcome is perception-based, interpret with caution"*).
- **Risk of Bias Summary:** Key points from the risk of bias assessment for the study. While detailed domain-level assessments go in Annex E, here we may include a short note if a study has major limitations (e.g., *high risk of bias due to no control group*).

All these fields will be captured in a tabular format during extraction for researcher readability, and simultaneously in a structured JSON format for computational use.

Example Extraction (Tabular Format):

Field	Data Excerpt (Example)
Study ID	ELIS2025_001
Title	Public confidence in electronic voting
Authors	Smith, J.; Nguyen, L.
Year	2021
Peer-reviewed?	Yes
Country/Region	Brazil
Electoral Modality	Electronic
Intervention Type	VVPAT introduction
Study Design	Field experiment
Evaluation Method	Pre-post election survey + audit logs
Outcome 1 Type	Primary
Outcome 1 Reported	Change in discrepancy rate in audit recounts
Outcome 1 Measure	% of ballot counts differing between electronic and paper records
Outcome 1 Design Link	Yes (directly tests VVPAT effect on discrepancies)
Outcome 1 Notes	Discrepancy rate dropped from 2% to 0.5% after VVPAT (causal link plausible)
Outcome 2 Type	Secondary
Outcome 2 Reported	Voter confidence level
Outcome 2 Measure	Mean survey trust score (1–5 scale)
Outcome 2 Design Link	No (many factors influence trust)

Outcome 2 Notes	No significant change in trust score (could be confounded by external events)
Risk of Bias Summary	Moderate risk (no randomization, but good transparency of data)

JSON starter template: Below is a snippet of the JSON structure corresponding to the above example, showing how the ELIS SLR Agent would store the data (keys corresponding to the fields above):

```
{
  "study_id": "ELIS2025_001",
  "title": "Public confidence in electronic voting",
  "authors": ["Smith, J.", "Nguyen, L."],
  "publication_year": 2021,
  "peer_reviewed": true,
  "country_or_region": "Brazil",
  "electoral_modality": "Electronic",
  "intervention_type": "VVPAT",
  "study_design": "Field experiment",
  "evaluation_method": "Survey and audit logs",
  "outcomes": [
    {
      "outcome_type": "Primary",
      "outcome_reported": "Discrepancy rate in audit recounts",
      "outcome_measure": "Percentage of mismatched ballots",
      "design_link": true,
      "reviewer_notes": "Discrepancy reduced from 2% to 0.5% post-intervention."
    },
    {
      "outcome_type": "Secondary",
      "outcome_reported": "Voter confidence score",
      "outcome_measure": "Mean trust score (1-5)",
      "design_link": false,
           "reviewer_notes": "No significant change; external factors likely
involved."
    }
  ],
  "risk_of_bias": {
    "study_design_appropriateness": "Moderate",
    "data_transparency": "High",
    "outcome_clarity": "High",
    "notes": "Well-documented methods, but no randomization."
  }
}
```

JSON example for one study; actual dataset will contain an array of such entries.

This JSON format is **schema-compliant**, meaning each field adheres to a predefined schema ensuring consistency across studies. The schema (defined in the project repository) dictates allowed values and formats (for example, outcome_type must be either "Primary" or "Secondary", risk of bias ratings must be "Low/Moderate/High"). Using a schema helps validate the data automatically – the ELIS SLR Agent will flag any entries that do not conform (e.g., a misspelled rating), thereby reducing data errors.

All extracted data will be made available as a CSV and JSON file. The JSON is especially useful for computational analysis or for sharing data for future evidence synthesis efforts. The final version of the extraction data will be deposited on GitHub (in the ELIS SLR Agent repository) and a permanent archive (such as Zenodo or OSF) upon completion of the review.

Annex E – Evidence Quality Assessments

This annex contains tools and templates for appraising the quality of evidence from the included studies, specifically: **E.1) Risk of Bias Tool and Log** for individual studies, and **E.2) CERQual Confidence Assessment Template** for synthesized findings.

E.1 Risk of Bias Tool and Rating Log (per study)

For each included study, a risk of bias assessment will be performed across key domains. The tool is adapted to this review's needs, combining elements from standard instruments. Each domain is rated **Low, Moderate, or High** risk of bias, with a short note explaining the judgment. The domains are:

- **Study design appropriateness:** Is the study design suitable to answer the research question? *Example:* A randomized controlled trial (RCT) in the context of a new voting technology would generally be Low risk on design (if well executed), whereas an uncontrolled before-after comparison might be Moderate or High risk due to potential confounders.
- **Data transparency:** Are the study's data sources and analyses reported transparently? *Example:* High risk if the study does not share important details or data (perhaps only high-level claims without method details); Low risk if data and analysis code are available or at least described in full. We also consider whether outcome data can be verified (e.g., official audit reports published).
- **Outcome measurement clarity:** Are the outcomes measured in a direct and reliable way? *Example:* If a study measures "integrity" via a clearly defined audit discrepancy rate, that's clearer (Low risk) compared to a study using a very indirect proxy for integrity (which could introduce bias). Also, if multiple outcomes were possible but only some reported, we flag potential reporting bias here.

Additional domains that may be noted (if applicable) include selection of reported results (selective reporting within the study) and any conflicts of interest that could bias results (e.g., study authors involved in developing the technology being evaluated).

All these assessments will be logged in a **Risk of Bias Log**, a table where each study has a row with its domain ratings and notes. For example:

Study ID	Design Appropriateness	Data Transparency	Outcome Clarity	Overall Bias Rating	Notes
ELIS2025_001	Moderate	High	High	Moderate	No control group, but methods clearly described and data shared.
ELIS2025_007	Low	Moderate	Moderate	Moderate	RCT conducted, but incomplete reporting of outcome definitions.

Full log will contain all included studies with complete bias assessments.

E.2 CERQual Confidence Rating Template (per synthesized finding)

For each major **finding or theme** that emerges from the synthesis (especially qualitative or mixed-method findings), we will assess confidence using CERQual criteria, as described in Section 3.8. The output will be a structured summary, and we will maintain a JSON record of these assessments as well.

Confidence Assessment Fields: for each finding/theme, we record:

- **Theme/Finding description** – a short identifier (e.g., *"VVPAT improves error detection"*).
- **Methodological Limitations:** Level of concern (None/Minor, Moderate, Serious) with a brief note (e.g., *"Minor concerns: most studies well-designed except one case study"*).
- **Coherence:** Level of consistency across studies (High, Moderate, Low coherence) with note (e.g., *"High coherence: all studies report reduction in discrepancies"*).
- **Data Adequacy:** Are there enough data? (Adequate or Not Adequate) with note (e.g., *"Moderate adequacy: only 3 studies, but all with decent sample sizes"*).
- **Relevance:** Direct relevance to our review context (High, Moderate, Low) with note (e.g., *"High relevance: evidence comes from national elections similar to those of interest"*).
- **Overall Confidence:** High, Moderate, Low, or Very Low – based on the above domains collectively.

We will provide a summary table in the final review for these, but here is an example in JSON format to illustrate how it might be stored (and to ensure schema consistency with our data management):

```
{
  "theme": "VVPAT reduces tally discrepancies",
  "methodological_limitations": "Minor concerns",
  "coherence": "High",
  "data_adequacy": "Moderate",
  "relevance": "High",
  "overall_confidence": "High"
}
```

This example would correspond to a finding that voter-verified paper audit trails (VVPAT) reduce discrepancies between electronic and paper counts. The judgment indicates that across studies, there were only minor methodological issues, findings were consistent (high coherence), data volume was moderate, and all in contexts highly relevant – leading to a high confidence in this conclusion.

All such JSON entries for each finding will be compiled. The **full set of CERQual assessments** will be available in the GitHub repository (for transparency) and referenced in the review's results section. We will also likely include a concise table in the review manuscript listing each key theme with its overall confidence level and a one-line explanation.

The *overall confidence assessments* will help readers understand which conclusions are strongly supported and which should be interpreted cautiously. This approach to assessing cumulative evidence strength is aligned with GRADE/CERQual recommendations and provides a structured way to discuss the implications of the findings.

Lastly, we will cross-reference the risk-of-bias data (from E.1) with the confidence in evidence (E.2). For instance, if all studies on a theme had high bias, even if coherent, our confidence might be capped at moderate. This interplay will be explicitly discussed in the review.

(The JSON schema for confidence assessments will be included in the repository. It mirrors the fields above, ensuring that each domain is captured for every theme. This ensures our outputs are machine-readable and easily sharable.)

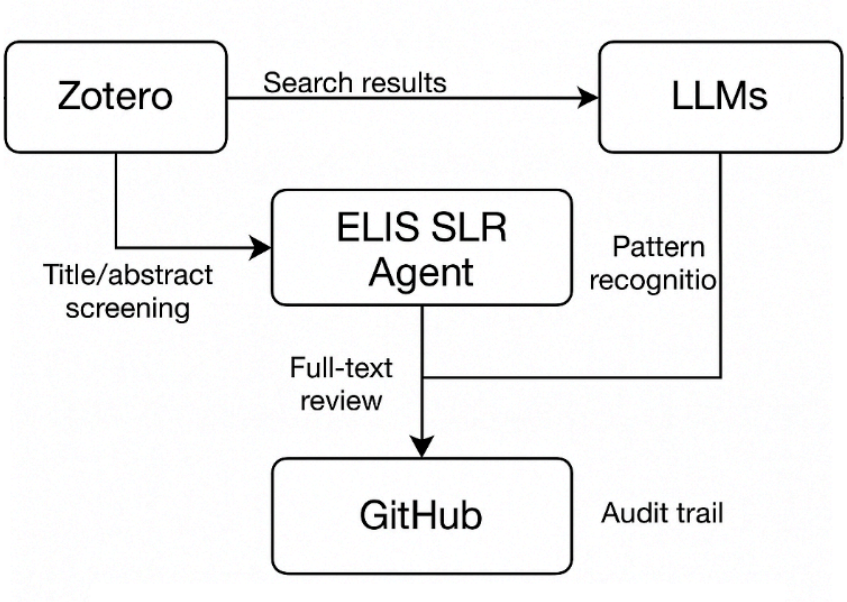
Annex F – Automation Workflow & AI

This annex provides a visual overview of the ELIS protocol's automation toolchain. Detailed task descriptions are provided in Section 3.4.

Tools Used

Tool	Purpose
Zotero	Reference management and de-duplication
ELIS SLR Agent	Automated tagging, bias flagging, JSON extraction
ChatGPT / Claude.ai / NotebookLM	Pattern detection, synthesis suggestions
GitHub	Version control, audit log, reproducibility

This diagram illustrates the flow from literature retrieval to final inclusion, showing how automation and researcher validation interact across the pipeline.



Annex G – Evidence Gap Confirmation and Supporting Literature

This annex documents the validation of the following statement in the ELIS Rationale:

"To date, there is no consolidated systematic review that jointly examines technological, operational, and institutional strategies for electoral integrity across both electronic and paper-based systems."

G.1 Purpose and Scope

To confirm the accuracy of this statement, a structured review of peer-reviewed literature from the past 10 years (2015–2025) was conducted. The objective was to identify any existing **systematic literature reviews**, **scoping reviews**, or other formal evidence syntheses that:

- Jointly examine **technological**, **operational**, and **institutional** strategies affecting electoral integrity; and
- Include both **electronic** and **paper-based** voting systems within the same review framework.

The search covered political science, governance, and election technology domains, prioritising systematic reviews published in high-impact journals and by major academic publishers.

G.2 Summary of Findings

The review confirmed that **no consolidated systematic review** has yet addressed this combined scope. Existing literature falls into distinct silos:

- **Technology-focused reviews** (e.g. on internet voting, blockchain, or biometric authentication) are common but narrowly focused on specific digital systems or use cases. They rarely incorporate paper-based modalities or address operational design.
- **Operational studies** tend to examine logistics, election-day procedures, or administrative performance, often as isolated variables. These do not address the interaction between process and technology or include broader institutional strategies.
- **Institutional analyses** and **integrity indices** (e.g. the Perceptions of Electoral Integrity dataset) provide high-level assessments but do not systematically review the impact of concrete system design features or specific interventions.

As a result, **no existing synthesis provides an interdisciplinary, comparative overview** of how technological, operational, and institutional features affect electoral integrity and auditability across voting system types.

This confirms the novelty and relevance of the ELIS review within both academic and practitioner communities.

G.3 Notable Existing Reviews and Gaps

Domain	Representative Review	Coverage	Gap
Electronic voting	Turnbull-Dugarte & Devine (2023) – *Systematic review of i-voting pilots*	Strong on internet voting and public trust	Does not address paper-based systems or institutional context
Election logistics	Apiriba & Lim (2025) – *SLR on logistics and election performance*	Covers supply chain and operational design	Does not address technology or governance
Institutional frameworks	Norris et al. (2014) – *PEI Index and integrity metrics*	Global scope on electoral quality	Not a systematic review; lacks mechanism-level detail
Biometric voting systems	Asimakopoulos et al. (2025) – *Impact of ICT on democratic processes*	Covers biometrics in Kenya and Asia	Focused on technology only
Election observation & trust	Kelley (2012), IDEA (2024), IFES (2021)	Emphasise transparency and oversight	Not structured as systematic reviews; limited to post-election practices

These examples underscore that while robust literature exists in each area, there is a **lack of integrated synthesis** across design dimensions and voting system types. The ELIS protocol addresses this gap.

G.4 Justification for PRQ Framing

In light of this evidence, the ELIS primary research question is framed with causal language:

> *"What operational and technological features of electoral systems have been shown to improve the integrity and auditability of elections since 1990?"*

To ensure methodological clarity, the protocol specifies that this phrasing includes:

- Findings from experimental and quasi-experimental studies,
- Observational studies and comparative analyses,
- Qualitative case studies and structured evaluations with empirical grounding.

Where causal inference is weak or contested, the review will clearly distinguish between robust evidence and more tentative findings. This framing is appropriate for informing electoral reform and policy design.

Annex H – Rule Development Log Template

Purpose

This log documents all changes to ELIS SLR Agent screening and classification rules during pilot testing and refinement.

File Format: JSON (logs/rule_development.json)

JSON Schema

```
{
  "type": "array",
  "items": {
    "type": "object",
    "required": ["date", "stage", "rule_change_description", "rationale",
"git_commit_hash"],
    "properties": {
      "date": {"type": "string", "format": "date"},
      "stage": {"type": "string", "enum": ["title_abstract_screening",
"peer_review_filtering", "data_extraction"]},
      "rule_change_description": {"type": "string"},
      "rationale": {"type": "string"},
      "git_commit_hash": {"type": "string"},
      "records_affected": {"type": "integer"},
      "precision_impact": {"type": "string", "enum": ["improved", "unchanged",
"worsened"]},
      "recall_impact": {"type": "string", "enum": ["improved", "unchanged", "worsened"]}
    }
  }
}
```

Field Definitions

- date: Date of rule change (YYYY-MM-DD)
- stage: Review stage (title_abstract_screening, peer_review_filtering, data_extraction)
- rule_change_description: Brief description of what changed
- rationale: Why change was made
- git_commit_hash: Git commit identifier for code change (enables exact reproduction)
- records_affected: Number of records whose classification changed due to this rule update
- precision_impact: Did change reduce false positives? (improved/unchanged/worsened)
- recall_impact: Did change reduce false negatives? (improved/unchanged/worsened)

Example Entries

```
[
  {
    "date": "2025-11-25",
    "stage": "title_abstract_screening",
    "rule_change_description": "Added 'transparency' to auditability synonym list",
    "rationale": "Pilot sample showed papers discussing transparency mechanisms were excluded despite relevance",
    "git_commit_hash": "a3f8d92",
    "records_affected": 47,
    "precision_impact": "unchanged",
    "recall_impact": "improved"
  },
  {
    "date": "2025-11-26",
    "stage": "peer_review_filtering",
    "rule_change_description": "Changed ambiguous journal classification from 'manual review' to 'auto-exclude'",
    "rationale": "Conservative approach prioritizes precision; reduces manual workload",
    "git_commit_hash": "b7e4c21",
    "records_affected": 23,
    "precision_impact": "improved",
    "recall_impact": "worsened"
  },
  {
    "date": "2025-11-27",
    "stage": "peer_review_filtering",
    "rule_change_description": "Added IEEE conference verification against IEEE rankings database",
    "rationale": "Distinguish peer-reviewed IEEE conferences from workshops",
    "git_commit_hash": "c9a2f45",
    "records_affected": 12,
    "precision_impact": "improved",
    "recall_impact": "unchanged"
  }
]
```

Usage Notes

- Log entry created for each code commit that modifies screening or classification logic
- Commit messages in GitHub should reference this log entry
- Cumulative record enables understanding of methodology evolution
- Published as supplementary material demonstrating iterative refinement process