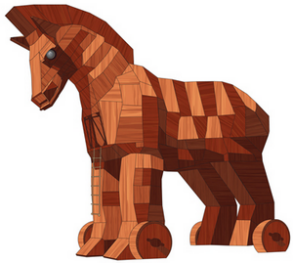
CISPA
HELMHOLTZ CENTER FOR INFORMATION SECURITY



Inspectable AI for Science: A Research Object Approach to Generative AI Governance

MetaCRiSP@IEEE S&P

Ruta Binkyte I May 2026

Supported by DFG, German Research Foundation, under grant agreement No. 568154038

Collaboration between CISPA, CSIRO, DATA61 & Macquarie

Sharif Abuaddba
DATA61,CSIRO

Chamikara Mahawaga Arachchige
Data61, CSIRO

Ming Ding
Data61,
CSIRO

Natasha Fernandes
Macquarie University

Mario Fritz
CISPA, Helmholtz Center for Information Security

Contributions:

★ Framework for governing generative AI as inspectable research infrastructure;

★Operationalization of the framework;

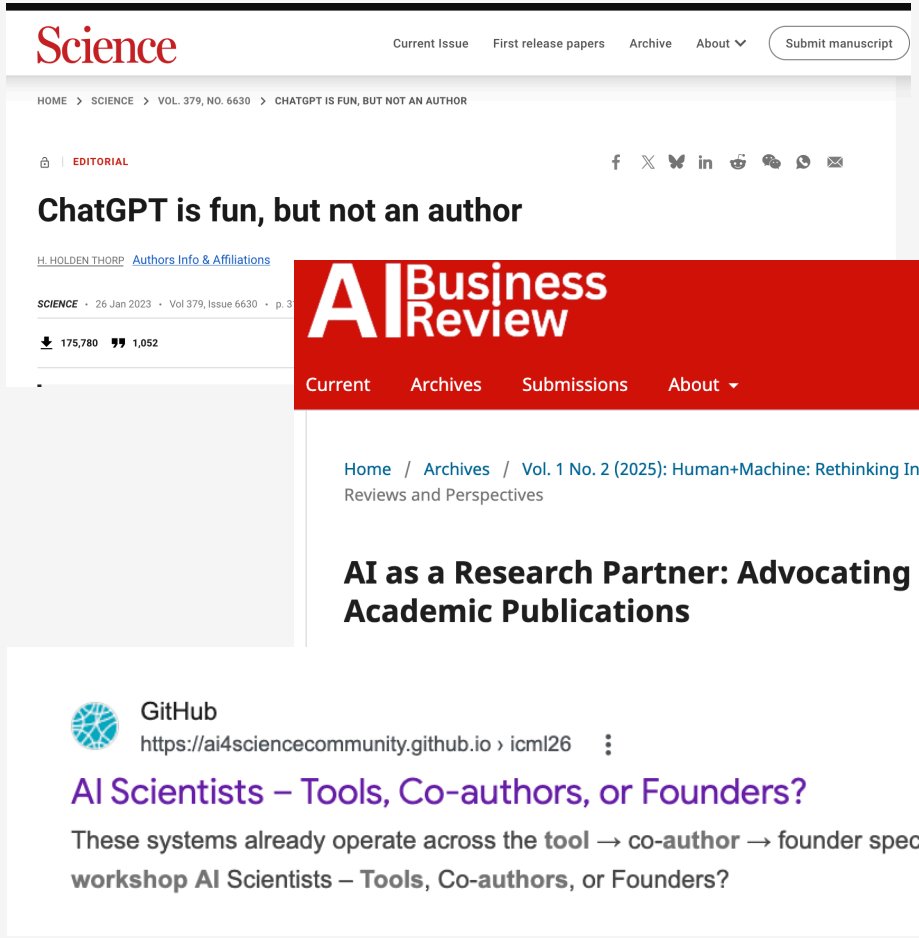
★ Feasibility demo;

★ Limitations and call for action.

Binkyte, R., Abuaddba, S., Mahawaga, C., Ding, M., Fernandes, N., & Fritz, M. (2026). Inspectable AI for Science: A Research Object Approach to Generative AI Governance. *MetaCRiSP@IEEE S&P*.

Background

- The **use of AI in scientific research** is disrupting existing scientific pipelines and trust
- The **practical** solutions are often limited to voluntarily disclosures
- The **theoretical** discussions focus on who should be considered an author in human-AI co-creation



Binkyte, R., Abuaddba, S., Mahawaga, C., Ding, M., Fernandes, N., & Fritz, M. (2026). Inspectable AI for Science: A Research Object Approach to Generative AI Governance. *MetaCRiSP@IEEE S&P*.

The Shortcomings of Existing Solutions

- **Focus on authorship** does not address provenance and auditability;
- **Narrative disclosures** are hard to verify and lack procedural details;
- **AI Detection tools** are error prone and unreliable;
- **Blanket prohibitions** discourage disclosure and legitimate use;
- **S&P Research** requires confidentiality, integrity and auditability.

Binkyte, R., Abuaddba, S., Mahawaga, C., Ding, M., Fernandes, N., & Fritz, M. (2026). Inspectable AI for Science: A Research Object Approach to Generative AI Governance. *MetaCRiSP@IEEE S&P*.

Shifting Discussion to AI Research Object

Research Object (RO) - a shared structured digital artifact from research process (Belhajjame et al, 2012)

- Traditional RO (data sets, software, workflows) → Interaction logs, prompts
- Versioning and configuration → Model versions, Temperature
- Reproducibility → Rerun model with the same context (limited)

- Authorship debates → AI as an artifact
- Binary disclosure → Inbuilt verifiable documentation
- Post hoc detection → Identifying failure points and learning

Binkyte, R., Abuaddba, S., Mahawaga, C., Ding, M., Fernandes, N., & Fritz, M. (2026). Inspectable AI for Science: A Research Object Approach to Generative AI Governance. *MetaCRiSP@IEEE S&P*.

Relevance for S&P Research

★ Confidentiality

✤ Vulnerability disclosures cannot be publicly shared

✓ Tiered, controlled disclosure, Trusted Research Environments (TREs)

★ Integrity

✤ Interaction logs create attack surface for modifications and misrepresentation of the role of AI

✓ Cryptographic hashing of logs

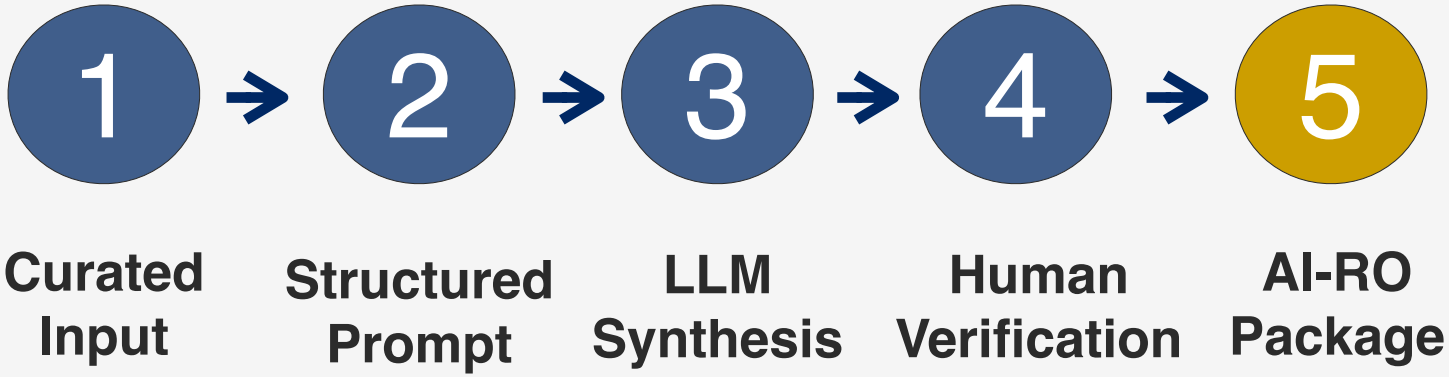
★ Auditability

✤ Binary or generic AI use disclosure is not auditable

✓ Structured record and auditable interface

Binkyte, R., Abuaddba, S., Mahawaga, C., Ding, M., Fernandes, N., & Fritz, M. (2026). Inspectable AI for Science: A Research Object Approach to Generative AI Governance. *MetaCRiSP@IEEE S&P*.

Use-case Example: Literature Review



AI as bounded synthesis tool
Human verification
Interaction provenance

Open ended use
No/minimal verification
Unverifiable use claims

PROMPT TEMPLATE

You are assisting with drafting a RELATED WORK section for a scientific paper on {{TITLE}}.

You will be given:

- a topic
- a contribution statement
- a taxonomy (clusters and assignments)
- structured human notes for each paper (summary, strengths, limitations, relation)

Task: Draft a related work section (target length: {{TARGET_WORDS}} words) that:

- Organizes discussion around the provided clusters
- Uses neutral, scholarly tone
- Highlights similarities and differences with respect to the contribution statement
- Mentions strengths/limitations carefully
- Avoids overclaiming (if evidence is weak, use cautious language)

Hard constraints (must follow):

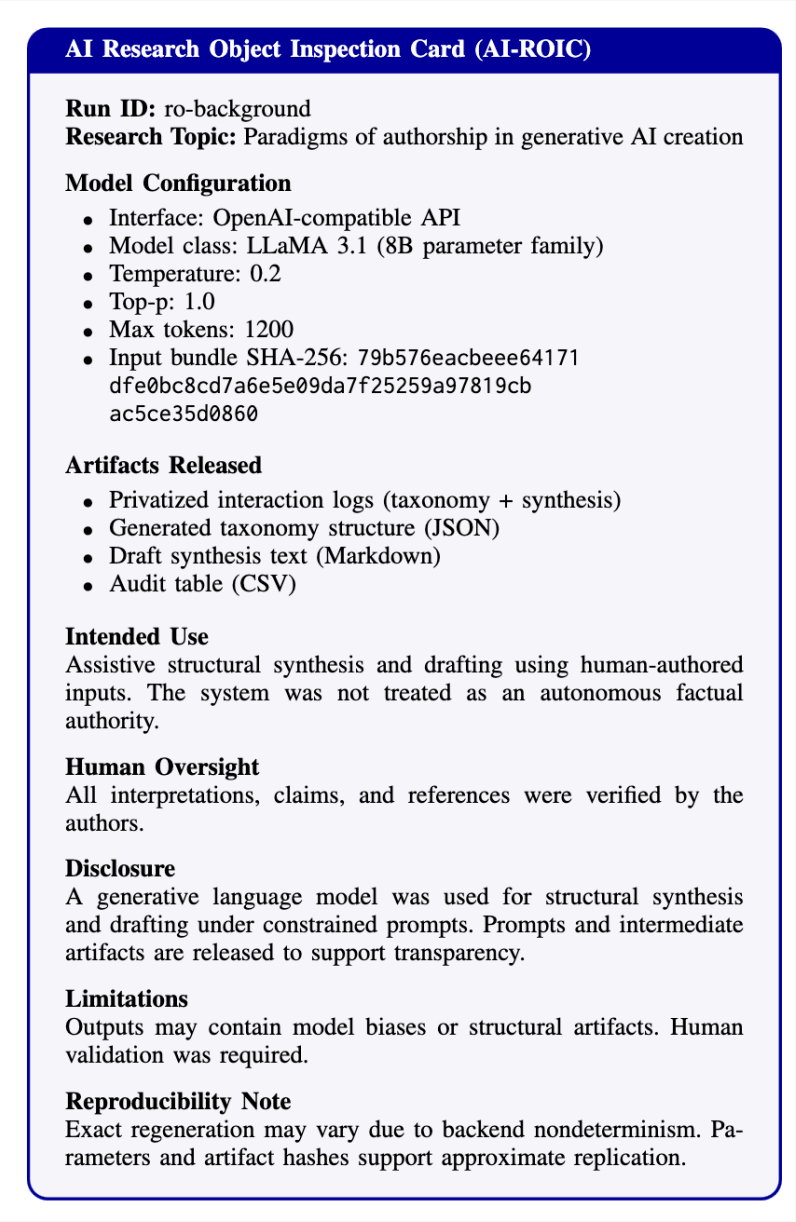
- Use ONLY the provided papers and identifiers. Do NOT invent citations or claims.
- Every comparative claim must be supported by the human notes provided.
- Use citations in the form: ({{citation}}; {{pid}}).
- If a claim is not supported by the notes, mark it as [NEEDS HUMAN CHECK].

Output format: 1) "RELATED WORK (DRAFT)" header 2) Draft text in paragraphs 3) "CLAIM CHECKLIST" bullet list with supporting paper IDs

Binkyte, R., Abuaddba, S., Mahawaga, C., Ding, M., Fernandes, N., & Fritz, M. (2026). Inspectable AI for Science: A Research Object Approach to Generative AI Governance. *MetaCRiSP@IEEE S&P*.

Components of AI-RO

- AI RO package: Prompts and anonymised interaction logs, Generated outputs, Human editing traces
- Machine-readable AI-RO metadata
- AI Research Object Inspection Card



AI Research Object Inspection Card (AI-ROIC)

Run ID: ro-background
Research Topic: Paradigms of authorship in generative AI creation

Model Configuration

- Interface: OpenAI-compatible API
- Model class: LLaMA 3.1 (8B parameter family)
- Temperature: 0.2
- Top-p: 1.0
- Max tokens: 1200
- Input bundle SHA-256: `79b576eacbeee64171 dfe0bc8cd7a6e5e09da7f25259a97819cb ac5ce35d0860`

Artifacts Released

- Privatized interaction logs (taxonomy + synthesis)
- Generated taxonomy structure (JSON)
- Draft synthesis text (Markdown)
- Audit table (CSV)

Intended Use
Assistive structural synthesis and drafting using human-authored inputs. The system was not treated as an autonomous factual authority.

Human Oversight
All interpretations, claims, and references were verified by the authors.

Disclosure
A generative language model was used for structural synthesis and drafting under constrained prompts. Prompts and intermediate artifacts are released to support transparency.

Limitations
Outputs may contain model biases or structural artifacts. Human validation was required.

Reproducibility Note
Exact regeneration may vary due to backend nondeterminism. Parameters and artifact hashes support approximate replication.

Binkyte, R., Abuaddba, S., Mahawaga, C., Ding, M., Fernandes, N., & Fritz, M. (2026). Inspectable AI for Science: A Research Object Approach to Generative AI Governance. *MetaCRiSP@IEEE S&P*.

Challenges & Future Directions

- Incentives for Transparency → Developing protocols for AI use supporting cultural norms that reward transperancy
- Security and Privacy → Developing privacy-aware provenance infrastructure
- Ethical Considerations → Encouraging use of open and responsibly trained models
- Scalability and Scope → Tools and interfaces for researchers and reviewers/auditors

Binkyte, R., Abuaddba, S., Mahawaga, C., Ding, M., Fernandes, N., & Fritz, M. (2026). Inspectable AI for Science: A Research Object Approach to Generative AI Governance. *MetaCRiSP@IEEE S&P*.

Key Take-Aways

1. AI-RO reframe the use of AI as structured, inspectable artefact

2. Legitimacy of AI-assisted scientific workflow depends on procedural transparency and human accountability

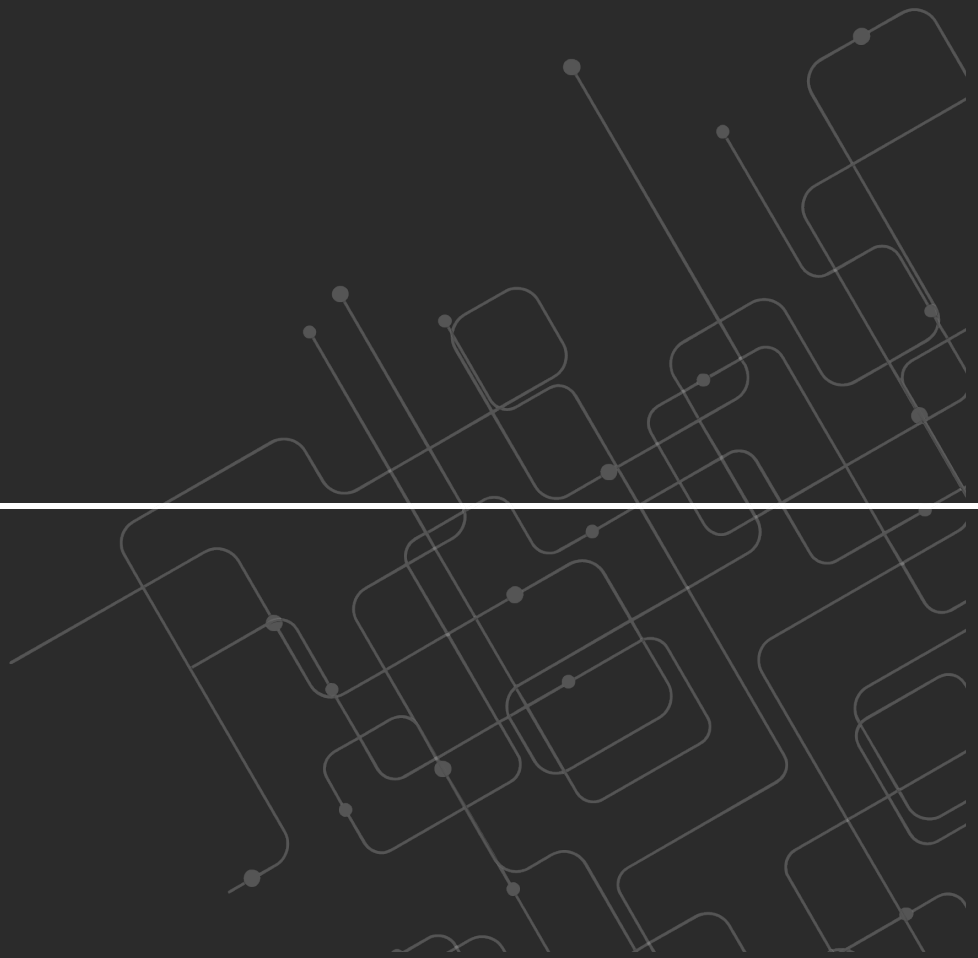
3. AI-RO differ from traditional RO in stochastic and interaction driven behaviour

4. Social and technical developments are necessary for the efficient adoption

Binkyte, R., Abuaddba, S., Mahawaga, C., Ding, M., Fernandes, N., & Fritz, M. (2026). Inspectable AI for Science: A Research Object Approach to Generative AI Governance. *arXiv preprint arXiv:2604.11261*.

Contact

www.rutabinkyte.com
ruta.binkyte@gmail.com
https://www.linkedin.com/in/ruta-binkyte/



Thank You!