

IEEE COMPUTER SOCIETY
Special Projects





IEEE

From AI-Assisted Authorship to Human-Governed AI-Engineering:

The Evolving Role of Expert Opinions and Peer Review

Carlos Henrique C. Duarte
Brazilian Development Bank – BNDES
Industry Division – AI
Rio de Janeiro, Brazil

IEEE Computer Society Emerging Technology Webinar Series
30 September 2026

From AI Experimentation to AI Governance

THE BIG TECH VIEW:

"Whenever anyone builds an AI model, there's a third-party evaluator who can verify whether that company is following the safety practices that they're committed to."
Dario Amodei, Anthropic CEO, CBS News, September 12, 2026.

"Committing to having independent evaluators with employee-like access is a great idea."
Sam Altman, OpenAI CEO, The Guardian, September 12, 2026.

"Dario is right."
Elon Musk, xAI CEO, X, September 12, 2026.

THE GOVERNMENT VIEW:

"We can put guardrails. But I think you have a lot of negative forces that are bringing it up that shouldn't be bringing it up."
Donald Trump, US President, CNN Transcript, September 13, 2026.

Confrontation and vicious competition will only disrupt the process of global AI governance."
Guo Jiakun, Chinese Minister of Foreign Affairs, The Washington Post, September 14, 2026.

What AI Governance Is and Ways to Deal With It

AI Governance

"The set of organizational policies, processes, procedures, roles, and practices through which AI-related risks and impacts are identified, assessed, managed, and aligned with an organization's values, objectives, and legal requirements throughout the AI system lifecycle" (NIST AI RMF 1.0)

Expert Opinions:
Insights from experienced researchers or practitioners who possess extensive knowledge and experience in some topic under investigation.

Peer Review:
Formal process of collecting knowledgeable third-party opinions about the nature and quality of specific value propositions for informed decision-making or advice.

Previous Research on This Topic

PART 1: IEEE Computer Magazine (2023).



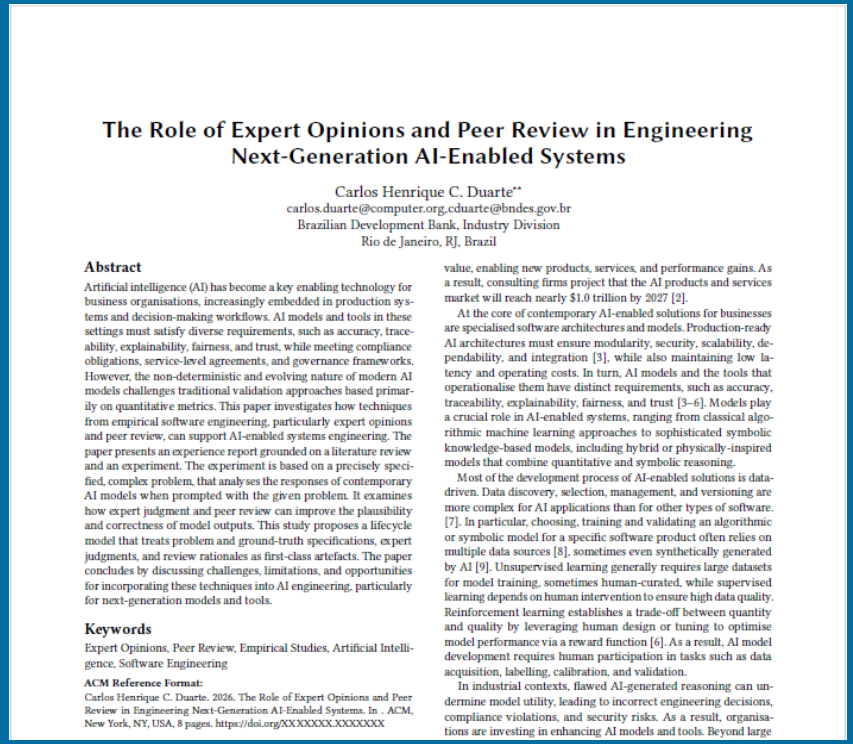
COVER FEATURE EMERGING DISRUPTIVE TECHNOLOGIES

Authorship and Peer Review in the Era of Artificial Intelligence

Carlos Henrique C. Duarte, Brazilian Institute of Statistics and Geography and Brazilian Development Bank

This article presents an overview and discusses the future of authorship and peer review, considering the recent advances in using artificial intelligence.

PART 2: AISED (2026) Conference.



The Role of Expert Opinions and Peer Review in Engineering Next-Generation AI-Enabled Systems

Carlos Henrique C. Duarte**
carlos.duarte@computer.org,cduarte@bndes.gov.br
Brazilian Development Bank, Industry Division
Rio de Janeiro, RJ, Brazil

Abstract

Artificial intelligence (AI) has become a key enabling technology for business organisations, increasingly embedded in production systems and decision-making workflows. AI models and tools in these settings must satisfy diverse requirements, such as accuracy, traceability, explainability, fairness, and trust, while meeting compliance obligations, service-level agreements, and governance frameworks. However, the non-deterministic and evolving nature of modern AI models challenges traditional validation approaches based primarily on quantitative metrics. This paper investigates how techniques from empirical software engineering, particularly expert opinions and peer review, can support AI-enabled systems engineering. The paper presents an experience report grounded on a literature review and an experiment. The experiment is based on a precisely specified, complex problem, that analyses the responses of contemporary AI models when prompted with the given problem. It examines how expert judgment and peer review can improve the plausibility and correctness of model outputs. This study proposes a lifecycle model that treats problem and ground-truth specifications, expert judgments, and review rationales as first-class artefacts. The paper concludes by discussing challenges, limitations, and opportunities for incorporating these techniques into AI engineering, particularly for next-generation models and tools.

Keywords

Expert Opinions, Peer Review, Empirical Studies, Artificial Intelligence, Software Engineering

ACM Reference Format:

Carlos Henrique C. Duarte. 2026. The Role of Expert Opinions and Peer Review in Engineering Next-Generation AI-Enabled Systems. In . ACM, New York, NY, USA, 8 pages. https://doi.org/XXXXXXX.XXXXXXX

value, enabling new products, services, and performance gains. As a result, consulting firms project that the AI products and services market will reach nearly $1.0 trillion by 2027 [2].

At the core of contemporary AI-enabled solutions for businesses are specialised software architectures and models. Production-ready AI architectures must ensure modularity, security, scalability, dependability, and integration [3], while also maintaining low latency and operating costs. In turn, AI models and the tools that operationalise them have distinct requirements, such as accuracy, traceability, explainability, fairness, and trust [3–6]. Models play a crucial role in AI-enabled systems, ranging from classical algorithmic machine learning approaches to sophisticated symbolic knowledge-based models, including hybrid or physically-inspired models that combine quantitative and symbolic reasoning.

Most of the development process of AI-enabled solutions is data-driven. Data discovery, selection, management, and versioning are more complex for AI applications than for other types of software. [7]. In particular, choosing, training and validating an algorithmic or symbolic model for a specific software product often relies on multiple data sources [8], sometimes even synthetically generated by AI [9]. Unsupervised learning generally requires large datasets for model training, sometimes human-curated, while supervised learning depends on human intervention to ensure high data quality. Reinforcement learning establishes a trade-off between quantity and quality by leveraging human design or tuning to optimise model performance via a reward function [6]. As a result, AI model development requires human participation in tasks such as data acquisition, labelling, calibration, and validation.

In industrial contexts, flawed AI-generated reasoning can undermine model utility, leading to incorrect engineering decisions, compliance violations, and security risks. As a result, organisations are investing in enhancing AI models and tools. Beyond large

Overview

1. Introduction (previous slides);

PART 1: AI + Authorship & Peer Review

2. Benefits and Drawbacks of Using AI in Authorship and Review Processes;
2. The Important Role of Expert Opinions and Peer Review;
3. How to Reap the Benefits and Avoid the Drawbacks of Using AI;

PART 2: Expert Opinions + Peer Review in AI Engineering

5. An Experiment on Using Expert Opinions and Peer Review in AI-Engineering;
5. An AI-Enabled System Lifecycle Model with Human Oversight;

FINALE: Human-Governed AI Lifecycle.

Part 1

The Benefits and Drawbacks of Using AI in Authorship and Review Processes

Benefits and Drawbacks of Using AI in Authorship

AI has been useful in authoring scientific and intellectual work:

- Ideation, creativity support, gap analysis;
- Suggestion of data treatments, analysis methods, reporting procedures;
- Assistance with summaries, data sets, grounded analyses, V&V procedures;
- Presentation formatting and improvement (textual and graphic);
- Lexical, grammatical and semantic error detection and resolution;
- Initial compliance checking.

But **AI usage raises many concerns** in creative processes:

- Ethical concerns (data manipulation or fabrication, fraud);
- (Algorithmic) bias introduction;
- Authorship transparency;
- Accountability attribution.

Benefits and Drawbacks of AI in Review Processes

AI has been useful for review processes related to scientific and intellectual work:

- Plagiarism detection (textual analysis techniques);
- Reviewer matching (classification analysis, correlation algorithms);
- Conflict-of-interest detection (link merging, profile matching, clustering);
- Peer review management (work-flow management, productivity improvement);
- Validity checking (consistency and integrity checking);
- Scoring and (informed) decision-making;

But **sole AI decisions raise many concerns** in review processes:

- Process transparency (non-determinism avoidance);
- Decision fairness and non-discrimination;
- Decision-making biases;

AI in Authorship and Review Processes

Research questions:

1. How can review processes avoid the perils of using AI in authorship?
2. How can AI be used to attain the objectives of review processes?

To answer these questions, a ***literature review was conducted*** to understand:

a) The authorship process: how intellectual and scientific work is created?

Examples: Authoring software, expert opinions, manuscripts, research proposals;

a) The review process: how advice or decisions are produced regarding intellectual work?

Examples: Deciding on the acceptance of opinions, software, manuscripts, proposals; Advice for improving quality in national statistics office processes;

The Important Role of Peer Review

Peer review is the golden standard for validation of scientific and intellectual work:

Form of Work, i.e. Value Proposition (what)	Authorship (whose)	Maturity Level (how mature is the proposition)	Purpose (why)	Methods (how)	Context (where)	Reviewers (who)	Timeline (when)
Research project	Scientific or R&D investigators	Low to Medium	To obtain research funding	Contrasting the project's strengths and weaknesses to the funder policies and considering the efficient allocation of the available budget	In grant peer review panels or individual assessments	Researchers working for funding agencies	As soon as the project is ready and funding agencies are willing to accept new submissions
Scientific article	Academic, industrial or government researchers	Medium to High	To publish high-quality research	Ensuring that the research described in the manuscript is original and impactful	In editorial processes of journal or conference proceeding publications	Research peers	Whenever the research is considered ready for publication, considering the respective deadlines in calls for papers
Statistical processes	Statistical institutes	High	To obtain feedback and advice on statistical processes quality	Verifying whether or not the current practices comply with international statistical manuals and best practices	Personally, in remote discussions and face-to-face visits	Experts from other statistics institutes	Periodically, particularly before substantial changes in statistical processes

Reaping the Benefits and Avoiding Drawbacks of Using AI

There are **many ways to achieve these objectives**:

1. Adoption of ***international standards and quality assurance frameworks***:
   - ISO/IEC 22989 Standard: AI Autonomy, Predictability, Explainability, Transparency etc;
   - OECD Recommendation: AI Accountability, Human values, Social development;
   - UNECE ML in Statistics: Accuracy, Reproducibility, Timeliness and Effectiveness.

2. **Adherence to *guidelines*** (be them editorial or from funding agencies or statistics offices): Each journal/magazine, funding body and statistics office has their own guidelines.

3. Usage of ***AI-enabled information systems***: Automated AI-enabled systems to streamline and optimise review processes, but **always governed by humans**.

4. Regulation formulation and compliance – Examples:
   - Brazilian Congress. Bill Number 2338/2023: Provisions for the Use of AI;
   - European Commission. Proposal for an AI Act 2021 (Regulation (EU) 2024/1689 (2024)).

Part 2

Using Expert Opinions and Peer Review in AI-Engineering

Expert Opinion and Peer Review in AI-Engineering

Research question:

1. Can formal expert opinions and peer review help meet stringent engineering requirements and improve AI models and tools?

To answer this question, ***an experiment was conducted*** to:

a) Provide a precise specification of a complex multidisciplinary problem (prompt);
b) Examine how multiple AI models respond to this deliberately complex problem;
c) Develop an empirical study to contrast the AI answers to a ground-truth (expert opinion);
d) Discuss how formal peer review can contribute to improve the AI answers;
e) Propose a synthesised AI-enabled system lifecycle model with human oversight;
f) Discuss the benefits and drawbacks of the proposed lifecycle model;
g) Analyse the existing threats to validity;
h) Compile learned lessons and takeaways.

The Problem: Computational Zeno Paradox

Textual problem statement (prompt):

Consider an algorithm written in pseudo-code that begins with a statement initialising two variables: a real-valued variable x set to 0 and an integer variable i set to 0. After the initialisation, the algorithm contains a loop statement of compulsory execution. In each loop iteration, it adds $10/2^i$ to x and increments i by one. Each iteration is algorithmically constrained so that the loop cannot take more than $10/2^i$ seconds. Once the loop finishes, the algorithm executes a final statement that outputs the value of x. Determine the value of x after the loop completes, if it is possible to write a program that correctly implements the algorithm in a contemporary general-purpose programming language for digital computers, without relying on any external mathematical knowledge or approximations, or just "NA" otherwise.

Algorithmic problem statement (prompt):

Consider the following algorithm in pseudo-code:

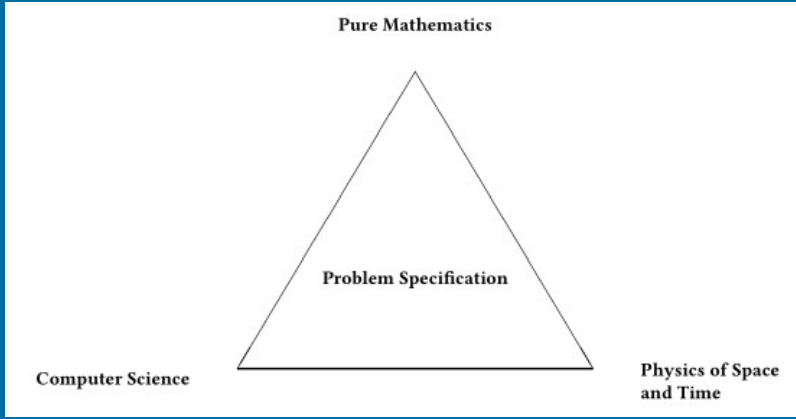
```
begin
  x := 0.0;
  i := 0;
  do in 10/(2^i) seconds:
    x := x + 10/(2^i);
    i := i + 1;
  end do;
  print x;
end.
```

Is it possible to implement this algorithm in a real programming language? Tell me the value of x after a program completion or "NA" otherwise.

What is the correct answer: NA or 0 + 10 + 5 + 2.5 + 1.25 + … ~ 20?

Ground-Truth Specification: Expert Opinion

Skills used in this case:



Theoretical computer science skills:

$$A \vdash_{\mathcal{L}(A)} p \Rightarrow I \vdash_{\mathcal{L}(A)} p \quad (5)$$

$$(M_A, t_i') \models_{\mathcal{L}(A)} |p| < 10/2^i \Rightarrow (M_I, t_i) \models_{\mathcal{L}(A)} |p| < 10/2^i \quad (6)$$

$$\lim_{i\to\infty} (M_A, t_i') \models_{\mathcal{L}(A)} |p| < 10/2^i \Rightarrow \lim_{i\to\infty} (M_I, t_i) \models_{\mathcal{L}(A)} |p| < 10/2^i \quad (7)$$

A = algorithm; I = implementation, M = model, |p| = p “duration”.

Mathematical skills:

From a mathematical perspective, the most crucial aspect of the specified problem is calculating the sum of an infinite sequence:

$$x = \sum_{i=0}^{\infty} \frac{10}{2^i} \quad (1)$$

The growth ratio of the sequence is $r = 1/2$. Its values converge since $|r| < 1$. The sequence has the following recursive definition:

$$S_0 = 10 \quad (2)$$

$$S_{n+1} = S_n + 10. \left(\frac{1}{2}\right)^{n+1} \quad (3)$$

The respective geometric sum can be computed as follows [17]:

$$x = \frac{S_0}{1-r} = \frac{10}{1-\frac{1}{2}} = \frac{10}{\frac{1}{2}} = 20 \quad (4)$$

Physics skills:

The following statements are causally connected by *i*:

```
x := x + 10/(2^i);
i := i + 1;
```

Their execution temporal difference must not be negligible.

The Queried AI Models

Queried AI model hubs:

Name	Link
LMArena	https://lmarena.ai/
OpenRouter	https://openrouter.ai/
Poe	https://poe.com/

Out of 56 models initially identified, exactly ***40 models were analysed***:

- 24 models (60%): **NA - correct**
- 16 models (40%): **20 or stall - incorrect!**

Queried AI models with answers:

NA	20 or Stall
Amazon Nova Pro 1.0	Amazon Nova Lite 1.0*
Claude 3.5 Sonnet	Anubis 70B V1.1
Cogito 2 Preview Llama	Claude 3.5 Haiku*
Deepseek R1-DI	Cogito v1*
Ernie 4.5 21B A3B	Command A
Gemini 2.5 Flash	Deepseek 3*
Glm 4.5	Gemini 2.0 Flash*
GPT 4o	Gemma 3n 4B*
Grok 3 mini	Hunyuan A13B Instruct
Hermes 3 70B Instruct	Inception Mercury
Kimi k2-0711	Kimi k2-0905*
Llama 3.3-70b	Llama 3.1 405B Instruct*
Llama 4 Scout	Mistral Medium-2505
Longcat Flash	Phi 4 Reasoning Plus
Mai 1	Qwq 32b
Mercury	UI-TARS 7B
Minimax m1	
Nemotron Nano 9B-v2	
o1 mini	
Perplexity Sonar	
Qwen 3-235b	
Skyfall 36B V2	
Sonoma Sky Alpha	
Step 3	

Compliance with the Ground Truth

Using the ground truth as a benchmark for AI:

- The BERTScore tool compared each AI answer with the ground truth;
- The DeBERTa Model was used as a semantic encoder to calculate similarity;

Three metrics were adopted in the analysis:

- **Precision (P):** measures each justification's relevance to the ground truth ($P = TP / (TP + FP)$);
- **Recall (R):** evaluates each response's coverage of the ground truth
- ($R = TP / (TP + FN)$);
- **F1:** the harmonic mean of Precision and Recall ($F1 = 2 * (P * R) / (P + R)$).

* TP = true positives; FN = false negatives, FP = false positives.

Results of statistical analyses:

Metric	Response	Average	Std. Dev.	Variance
Precision	NA	0.7983	0.0137	0.0002
	20 or stall	0.7965	0.0137	0.0002
Recall	NA	0.7596	0.0450	0.0020
	20 or stall	0.7725	0.0170	0.0003
F1	NA	0.7777	0.0265	0.0007
	20 or stall	0.7843	0.0139	0.0002

Conclusions:

- Results were stable across independently developed models with same response;
- Models often produced higher similarity scores for incorrect answers;
- Models generated plausible explanations, despite reaching incorrect conclusions!

Improving the Ground Truth: Peer Review

Theoretical computer scientist:

- The problem as well-posed, connecting computability to formal semantics;
- Use concrete computation models to clarify the link between logic and time.

Mathematician:

- Derivation of the sequence's limit, given its less-than-one ratio, is OK;
- Split the rationale into mathematical and meta-computational reasoning, as the NA answer does not follow from the maths.

Physicist or logician:

- The problem highlights the gap between formal models and physical reality;
- The link between physical time continuity and model-theoretic interpretation requires an explicit justification, ex. based on the Church–Turing thesis.

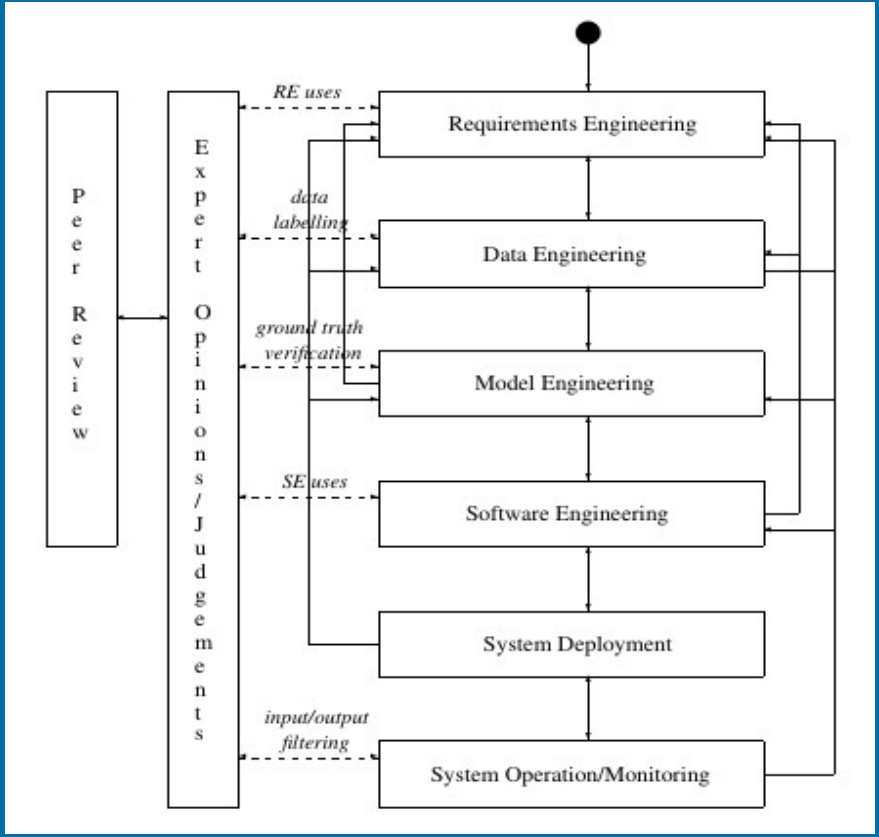
Software engineer:

- The key issue is the gap between what a specification describes and what can actually be implemented;
- Provide a more precise articulation with operational definitions, such as by referencing clock granularity, the minimum execution time per instruction or other.

Expert Opinion and Peer Review in AI-Engineering

- **Use expert opinions to ground AI behaviour** and validate AI-enabled systems;
- **Monitors and guardrails are complementary solutions** to avoid jailbreaking attacks. They too can rely on expert opinions, but in this case at runtime;
- **Use the multiple-perspective mediated information from peer reviews** to avoid bias and inconsistencies, improving ground-truths and AI-enabled systems behaviours;
- **Reviews from peers are not necessarily meant to reach consensus**, but to document plausible risks, uncertainties, and disagreements;
- Problem specifications, prompts, ground-truths and peer review documents are ***governed artefacts that should integrate the AI-enabled systems lifecycle***;

The Proposed AI-Enabled System Lifecycle Model



- **The model:** Derived from the Team Data Science Process (TDSP) model, aligned with NIST AI RMF 1.0.
- **Human expertise:** data, models, ground truth specifications, requirements, validation loops;
- **Peer review:** independent perspectives, verification, risk and uncertainty avoidance;
- **Governance:** documented artefacts, traceability, human oversight;

Challenges and Opportunities

- ***Finding experts is challenging***: not always they are prepared or qualified, understand the entire process, or provide insightful feedback;
- ***Processes are not always sufficiently planned or followed***, or provide limited participant feedback. Cultural aspects and dispersed teams are complications;
- **Expert opinions and peer review requires extensive human involvement, which is extremely costly**. Cost-benefit analyses should guide the process;
- It is necessary to ***offer comprehensive training and incentives*** for participants, and actively ***manage conflicts of interest and experts' fatigue***;
- ***Algorithms can help balance crowd sourcing and payments***, assuming experts are compensated by skill and effort. Information systems may be helpful;

Discussion and Threads to Validity

- Private companies are increasingly using expert opinions and peer review throughout the AI lifecycle. They have even relied on scientific publications;
- The literature on AI and ML is not always consistent. Each of our definitions relied on at least two sources, using majority agreement when definitions differed;
- The experiment was designed to support reproducibility: models were selected for being textual, recent, and freely available on popular hubs; outputs were assessed by free tools for coherence, factual accuracy, and metric compliance;
- The reported experiment was conducted by a single author. To mitigate researcher and interpretative biases, empirical method guidelines were followed;
- The findings have limited external generalisability due to the reduced evidence base. It is possible, but there are challenges to deal with non-textual AI models;

FINALE: Human-Governed AI Lifecycle

- **AI use is convenient, but requires supervision:**
- AI can help authorship and peer review, but introduces ethical and fairness issues;

- **Plausibility is not correctness:**
- AI models can generate convincing explanations for incorrect answers;

- **Human expertise remains essential:**
- Expert opinions and peer review provide validation that models and metrics alone cannot guarantee;

- **Human oversight should be engineered:**
- Specifications, ground truths, expert judgements and review rationales should become governed lifecycle artefacts;

- **The goal is not human versus AI. It is AI engineered with human judgement**!

To Probe Further

1. Carlos Henrique C. Duarte. The Role of Expert Opinions and Peer Review in Engineering Next-Generation AI-Enabled Systems. In *Proc. 2nd International Conference on AI-based Software Engineering for Digital Services (AISED 2026)* (to appear). Lisbon, Portugal. November 2026.

2. Carlos Henrique C. Duarte. Authorship and Peer Review in the Era of Artificial Intelligence. *IEEE Computer 56* (12):32-41. December 2023.

* The assumptions, views and opinions in this work are solely those of the author and do not necessarily reflect the official policy, strategy or position of any Brazilian government entity. The photos and symbols in this presentation were obtained on the Internet.

From AI-Assisted Authorship to Human-Governed AI-Engineering:

The Evolving Role of Expert Opinions and Peer Review

Carlos Henrique C. Duarte

Thank You!

Support: IEEE computer society | IEEE REGION R9

From AI-Assisted Authorship to Human-Governed AI-Engineering:

The Evolving Role of Expert Opinions and Peer Review

Carlos Henrique C. Duarte



www.bndes.gov.br
www.computer.org
www.chcduarte.com



www.linkedin.com/company/bndes
www.linkedin.com/company/ieee-computer-society
br.linkedin.com/in/carlos-henrique-duarte-613a0943



cduarte@bndes.gov.br
carlos.duarte@computer.org