

THE FLAVIA PRINCIPLE

Human–AI Epistemic Verification and Accountability in Consequential Decision-Making

Proposed Academic Manuscript — 2026

Dylan Chetty

Concept originator and human co-developer

Developed through a human–AI collaboration with ChatGPT

AI system contribution disclosed transparently; ChatGPT is not presented as a legal or accountable author.

Abstract

Generative artificial intelligence has introduced a distinctive epistemic problem into professional decision-making. Contemporary AI systems can generate propositions that are fluent, coherent, specific and apparently authoritative while nevertheless being incorrect, unsupported, incomplete or inappropriate to the context in which they are used. The resulting risk is not limited to model error. It also includes human over-reliance: the tendency to assign credibility to an AI-generated proposition because of the manner in which it is presented rather than because of the quality of evidence supporting it. This paper proposes the Flavia Principle, a human–AI epistemic governance principle stating that the apparent fluency, confidence, specificity or coherence of an AI-generated proposition shall not constitute evidence of its truth; reliance upon an AI output should be proportionate to the consequence of accepting that output if it is wrong and should require evidence, competent judgement and verification commensurate with that consequence. The paper distinguishes four epistemic states—AI-generated proposition, information, verified knowledge and authorised decision—and introduces three associated concepts: the Flavia Threshold, Flavia Failure and the Flavia Verification Loop. The proposed framework is examined in engineering, science and project, programme and portfolio management. It is positioned as a complementary layer to established AI-risk and professional-governance frameworks rather than as a replacement for them. The paper argues that mature AI governance must address not only how AI systems are governed, but also how humans govern their reliance upon AI-generated propositions. The paper concludes with propositions for empirical research and a practical assurance model suitable for engineering and other consequential domains.

Keywords

Artificial intelligence; generative AI; epistemic governance; human oversight; engineering assurance; automation bias; verification; professional judgement; project management; AI risk management; accountability.

1. Introduction

Generative artificial intelligence is changing how professionals obtain, interpret and produce information. Large language models can summarise complex material, generate alternative explanations, perform structured reasoning, assist with programming, draft technical documents and support decision analysis. These capabilities create substantial productivity opportunities. They also create a new category of epistemic risk: the system can produce an answer that resembles expert knowledge without possessing a reliable mechanism for guaranteeing that the proposition is true in the specific context in which it is used. The distinction matters because professional decisions are rarely judged solely by the fluency of their supporting narrative. Engineering designs, scientific conclusions, project forecasts and portfolio decisions require evidence, competence, traceability and accountability. A plausible answer is therefore not equivalent to a justified answer. The central problem addressed in this paper is consequently not whether artificial intelligence can make mistakes. It plainly can. The deeper question is how humans and organisations should regulate their degree of reliance on an AI-generated proposition before allowing it to influence a consequential decision. This paper proposes the Flavia Principle as a concise rule for that problem. The principle emerged from a 2026 human–AI dialogue in which an apparently authoritative AI answer was subsequently found to contain material factual errors. The incident prompted a broader examination of the relationship between confidence, plausibility, evidence and truth. The resulting proposition is intentionally modest: it does not attempt to replace established AI governance frameworks, professional standards or

engineering codes. Instead, it proposes an epistemic control layer between AI output and consequential human action.

2. Background: From Information Retrieval to Generative Propositions

Earlier information technologies predominantly retrieved, calculated or transformed information according to explicit procedures. Generative AI changes the interaction by producing natural-language propositions that may combine factual recall, inference, synthesis and prediction in a single response. For a user, the output can therefore resemble a considered expert opinion. This creates a human-factors problem. Humans routinely use presentation cues when assessing information. Confidence, specificity, coherence and apparent expertise can influence perceived credibility. In generative AI, these cues can be produced computationally even when the underlying proposition is uncertain. The problem is amplified by the asymmetry between generation and verification. Generating a paragraph, recommendation or calculation may take seconds, while establishing whether it is correct may require primary-source research, independent calculation, domain expertise or physical testing. A mature governance system must therefore distinguish the cost of producing an answer from the cost of establishing that the answer is sufficiently reliable for its intended use.

3. Existing Governance Landscape

The Flavia Principle should be understood in relation to existing work rather than in isolation. NIST's Artificial Intelligence Risk Management Framework (AI RMF) treats trustworthy AI as a socio-technical problem and addresses characteristics including validity and reliability, safety, security and resilience, accountability and transparency, explainability and interpretability, privacy and fairness. The framework also emphasises governance, measurement and management of AI risks. Professional project management is also developing explicit AI governance. PMI's Standard for Artificial Intelligence in Portfolio, Program, and Project Management addresses responsible AI adoption and places emphasis on governance, risk, human oversight and accountability in project environments. Publication-ethics organisations have also responded to generative AI. Current guidance from ICMJE and COPE states that AI systems should not be listed as authors because they cannot assume the responsibility and accountability associated with authorship. Instead, human authors remain responsible for the accuracy, integrity and originality of submitted work and should disclose material AI use. These developments support the general direction of the Flavia Principle. The proposed contribution is narrower: it focuses specifically on the epistemic transition between an AI-generated proposition and a consequential human decision. The question is not merely whether the system is governed, but whether the user's confidence in a particular output is justified by evidence appropriate to the consequences of error.

4. Origin and Development of the Flavia Principle

The term Flavia Principle originated in a 2026 dialogue between Dylan Chetty and ChatGPT. The immediate trigger was an AI-generated account of a film that contained confident but incorrect factual details. The significance of the incident was not the entertainment subject itself. It was the recognition that the same conversational mechanism—producing detailed, confident and coherent statements without adequate verification—could become materially dangerous when transferred to engineering, science, history, strategic analysis or critical infrastructure. The dialogue then developed into a broader proposition: if an AI system can generate an apparently authoritative falsehood, then an organisation requires a method for deciding when an AI proposition is sufficiently verified for a particular decision. The consequence of error becomes a natural determinant of the required assurance effort. The present paper therefore treats the Flavia Principle as a proposed concept, not an established historical doctrine. Its provenance is important because the principle itself advocates epistemic honesty. The human contribution consists in identifying, challenging and directing the conceptual problem; the AI contribution consists in assisting synthesis, formulation and formalisation. For publication, the exact treatment of AI contribution must follow the authorship and disclosure requirements of the target journal or repository.

5. Definition of the Flavia Principle

The proposed principle is: No AI-generated proposition shall acquire epistemic authority merely through its fluency, confidence, specificity, coherence or apparent reasoning quality. The degree of verification required before relying upon an AI output shall be proportionate to the consequence of accepting that output if it is wrong, and shall require evidence, competent judgement and verification commensurate with that consequence. The principle contains four propositions. First, presentation quality is not evidence of truth. Second, reliance is a decision rather

than an automatic consequence of generation. Third, the burden of verification should scale with consequence. Fourth, consequential authority remains with the appropriately competent and authorised human or institution.

6. Four Epistemic States

6.1 AI-Generated Proposition

The system has produced a statement, calculation, recommendation or interpretation. At this point the organisation has evidence only that the system generated the proposition.

6.2 Information

The proposition has been associated with potentially relevant sources or observations. Information may be useful without being sufficient for reliance.

6.3 Verified Knowledge

The proposition has been tested against appropriate evidence and independently evaluated to a level proportionate to its intended use.

6.4 Authorised Decision

An appropriately competent and authorised human has accepted the verified information and made or approved the decision. Accountability attaches at this stage.

7. The Flavia Threshold

The Flavia Threshold is the point at which an organisation has sufficient evidence and verification to justify reliance on an AI-assisted proposition for a defined purpose. It is not a universal numerical threshold because acceptable assurance depends on context. A useful conceptual model is: Required Assurance $\propto$ Consequence of Error $\times$ Uncertainty $\times$ Decision Criticality. The expression is a governance heuristic rather than a statistical law. It captures the central proposition that a low-consequence drafting task and a safety-critical engineering decision should not be subjected to the same epistemic controls. For low-consequence uses, ordinary human review may be sufficient. For material professional decisions, source verification and competent review may be necessary. For safety-critical or infrastructure decisions, independent calculation, testing, formal assurance and technical-authority approval may be required.

8. Flavia Failure

A Flavia Failure is defined as a decision-making failure in which an AI-generated proposition materially influences a decision and is accepted with greater confidence than its evidential basis warrants, particularly where verification is inadequate relative to the consequences of error. The concept is deliberately broader than AI hallucination. A hallucination is a property or behaviour of an AI output. A Flavia Failure is a socio-technical governance failure. An AI proposition may even be factually correct and still reveal a Flavia Failure if the organisation had no defensible basis for accepting it in a consequential context. This distinction moves responsibility from a simplistic 'the AI was wrong' explanation toward a more useful question: 'Was the organisation's reliance on the AI proposition justified?'

9. The Flavia Verification Loop

The operational model proposed by the principle is a closed verification loop: ASK — define the proposition that must be established. GENERATE — obtain the AI-assisted proposition. SOURCE — identify authoritative evidence and provenance. CHALLENGE — identify assumptions, contradictions and plausible failure modes. VERIFY — independently establish the proposition to the required assurance level. JUDGE — apply competent professional judgement. AUTHORISE — obtain approval from the person or body with appropriate authority. RECORD — preserve the evidence and decision trail. MONITOR — compare the decision with subsequent reality and capture lessons. The loop is intended to prevent a common failure mode in which generation is treated as the end of reasoning rather than the beginning of verification.

10. Five Laws of the Flavia Principle

Law 1 — Plausibility is not proof.

A convincing presentation has no inherent epistemic authority.

Law 2 — Consequence determines assurance.

The greater the consequence of error, the stronger the required verification.

Law 3 — AI does not inherit professional authority.

AI can inform a decision but does not become a licensed engineer, technical authority, project director or accountable decision-maker.

Law 4 — Material claims require proportionate verification.

The verification method should match the consequence, uncertainty and criticality of the claim.

Law 5 — Consequential decisions require traceability.

A material AI-assisted decision should be reconstructible from input through evidence, verification, judgement, approval and outcome.

11. Engineering Application

Engineering provides a particularly clear test of the principle because technical decisions can have physical, financial and safety consequences. Consider a proposed 132 kV circuit breaker rating. An AI system may produce a plausible recommendation based on a fault level supplied in the prompt. That recommendation cannot by itself establish suitability. A defensible engineering process would validate the network model, confirm maximum and minimum fault levels, establish the relevant X/R characteristics, determine symmetrical and asymmetrical duties, check making and breaking capability, verify short-time withstand, consider applicable standards and manufacturer requirements, and perform an independent calculation or review. A competent Technical Authority would then accept or reject the design. The AI may therefore accelerate engineering work without becoming the source of engineering authority. This distinction is particularly important for protection settings, transformer specification, earthing, insulation coordination, structural design, safety systems and commissioning decisions.

12. Project, Programme and Portfolio Management

The same principle applies to project, programme and portfolio decisions. An AI system may forecast completion probability, identify schedule risk, recommend portfolio prioritisation or estimate benefits. These outputs can be valuable, but the decision-maker must distinguish between model-generated analysis and established evidence. For a project forecast, for example, the project manager should examine the schedule logic, resource assumptions, critical path, historical performance, dependencies, risk exposure and uncertainty behind an AI-generated prediction. At programme level, benefits and dependencies should be tested against the approved benefits architecture. At portfolio level, strategic alignment, value, risk and capacity assumptions should be independently assessed. The principle therefore complements human-in-the-loop approaches by requiring that human involvement be substantive rather than ceremonial. A person who merely reviews an AI answer without challenging its basis does not necessarily provide meaningful assurance.

13. Scientific and Historical Research

The principle is equally applicable to research. AI may assist literature discovery, hypothesis generation, synthesis and drafting, but the researcher must verify primary evidence and distinguish interpretation from established fact. Historical research is a particularly useful illustration because generative systems may produce plausible narratives by combining names, dates and events that are individually familiar but incorrectly connected. Scientific work presents a parallel issue when an AI system invents citations, misstates a result or proposes an apparently reasonable explanation that is not supported by the underlying data. The Flavia Principle therefore supports a simple research rule: the AI output is a research aid; the evidence is the basis for the claim.

14. Organisational Governance Model

An enterprise implementation can classify AI-assisted work by consequence. A practical five-level model is F0 to F4. F0 covers administrative and low-consequence assistance. F1 covers informational assistance requiring

source checking where factual claims matter. F2 covers professional support requiring competent review. F3 covers material technical or managerial decisions requiring documented independent verification. F4 covers safety-critical or otherwise high-consequence decisions requiring formal assurance and designated technical or executive authority. The classification should be embedded in existing management systems rather than creating a parallel bureaucracy. AI-assisted decisions can be linked to project risks, issues, changes, stage gates, technical baselines and portfolio decisions. Evidence can be maintained through an AI Engineering Assurance Record or equivalent enterprise record.

15. The Flavia Test

For practical use, the principle can be reduced to six questions: F — FACTS: Are the inputs correct? L — LOGIC: Is the reasoning valid? A — AUTHORITY: Are the sources authoritative? V — VERIFICATION: Has the result been independently checked? I — IMPACT: What happens if it is wrong? A — ACCOUNTABILITY: Who is authorised to accept it? The test is deliberately simple. Its purpose is not to replace engineering calculations, scientific methods or project governance. It is a cognitive and procedural checkpoint that prompts the user to ask whether the basis for reliance is proportionate to the consequence of error.

16. Research Propositions and Future Empirical Work

The principle should be subjected to empirical testing rather than accepted merely because it is intuitively appealing. Three initial propositions are proposed. H1: Greater perceived fluency and confidence in AI output increases the probability of inappropriate reliance when explicit verification requirements are absent. H2: Consequence-based verification controls reduce inappropriate reliance on erroneous AI-generated propositions. H3: Professional expertise alone does not eliminate AI over-reliance unless verification mechanisms are embedded in the decision process. Future studies could use controlled experiments comparing professionals with and without explicit Flavia-style verification prompts. Engineering simulations could test whether independent calculations reduce acceptance of erroneous AI recommendations. Project-management studies could compare AI-assisted forecasting with and without consequence-based assurance controls. Further research should also examine false positives, verification fatigue, time costs, automation bias, calibrated trust, professional liability and the effect of organisational culture.

17. Discussion

The Flavia Principle is intentionally not an anti-AI doctrine. Its objective is to improve the quality of human–AI collaboration by separating acceleration from authority. AI can increase the speed at which hypotheses, alternatives, analyses and draft outputs are produced. The organisation must still determine what is sufficiently established to act upon. This distinction is particularly important because a system can be both useful and unreliable in the same interaction. The appropriate response is therefore neither blind trust nor blanket rejection. It is calibrated reliance. The principle also shifts attention from model-centric governance toward socio-technical governance. Even a highly reliable model can be misused if users apply an output outside its validated context. Conversely, a fallible model can be used safely if its outputs are subjected to proportionate verification before consequential reliance. The key governance object is therefore not simply the AI system. It is the entire chain from generation to decision.

18. Limitations

Several limitations should be recognised. First, the principle does not provide a universal numerical method for calculating verification effort. Second, the proposed definitions require empirical validation. Third, the distinction between independent and non-independent verification can be difficult in highly specialised domains. Fourth, professional judgement remains necessary even after evidence has been verified. Fifth, organisational incentives may encourage superficial compliance, such as completing checkboxes without substantive challenge. The principle should therefore be treated as a proposed framework requiring testing, refinement and domain-specific implementation.

19. Conclusion

Generative AI changes the speed and scale at which propositions can be produced. It does not eliminate the need to establish whether those propositions are true, relevant and sufficiently reliable for the decision at hand. The

Flavia Principle proposes a simple but consequential rule: the apparent authority of an AI-generated proposition must never substitute for evidence, and the burden of verification should increase with the consequence of error. Its central distinction is between generation and truth establishment. AI may generate. Evidence may support. Competent professionals may judge. Authorised humans may decide. The principle therefore seeks to preserve a fundamental feature of responsible professional practice in an age of generative AI: responsibility for consequential truth claims cannot be delegated merely because a machine produced them fluently.

20. Authorship, Contribution and AI Disclosure

The conceptual origin of the Flavia Principle is attributed to Dylan Chetty and ChatGPT as a human–AI collaborative development. Dylan Chetty provided the originating professional challenge, critical interrogation, conceptual direction and human judgement that initiated and shaped the principle. ChatGPT assisted with synthesis, formulation, drafting and development of the conceptual and practical framework. For formal scholarly submission, however, the manuscript should comply with the authorship rules of the selected journal or repository. Current ICMJE and COPE guidance does not permit AI systems such as ChatGPT to be listed as authors because they cannot assume the accountability and responsibilities attached to authorship. The appropriate scholarly presentation is therefore to identify Dylan Chetty as the human author and disclose ChatGPT's substantive contribution transparently, while preserving the historical statement that the concept was co-developed through the human–AI dialogue.

References

- National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. <https://doi.org/10.6028/NIST.AI.100-1>
- National Institute of Standards and Technology. (2026). AI Risk Management Framework resources and guidance. National Institute of Standards and Technology.
- Project Management Institute. (2026). Standard for Artificial Intelligence in Portfolio, Program, and Project Management. Project Management Institute.
- International Committee of Medical Journal Editors. (2026). Use of Artificial Intelligence in Publishing: Recommendations for authors and contributors. ICMJE.
- Committee on Publication Ethics. (2024). Authorship and AI tools: COPE position statement. Committee on Publication Ethics.
- National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. NIST.
- ISO/IEC. (2023). ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system. International Organization for Standardization.
- ISO/IEC. (2023). ISO/IEC 23894:2023 Information technology — Artificial intelligence — Guidance on risk management. International Organization for Standardization.

Proposed citation

Chetty, D. (2026). The Flavia Principle: Human–AI Epistemic Verification and Accountability in Consequential Decision-Making. Proposed academic manuscript. Concept developed through a human–AI collaboration with ChatGPT.

Appendix A — One-Page Flavia Framework

FLAVIA PRINCIPLE AI OUTPUT ≠ TRUTH AI OUTPUT ≠ EVIDENCE AI OUTPUT ≠ PROFESSIONAL JUDGEMENT AI OUTPUT ≠ AUTHORITY F — FACTS: Are the inputs correct? L — LOGIC: Is the reasoning valid? A — AUTHORITY: Are the sources authoritative? V — VERIFICATION: Has the result been independently checked? I — IMPACT: What happens if it is wrong? A — ACCOUNTABILITY: Who is authorised to accept it? CORE RULE: The consequence of being wrong determines the burden of proving that we are right.