STIAEM

Slow Tech Innovation

AI, Ethics and

New Teaching Methods

Deploying systematic doubt as a pedagogical trigger for critical thinking

Co-funded by the European Union

2025-1-IT02-KA220-SCH-000365213

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Education and Culture Executive Agency (EACEA). Neither the European Union nor EACEA can be held responsible for them... Project reference code: 2025-1-IT02-KA220-SCH-000365213

Università di Foggia

Studi Umanistici, Lettere, Beni Culturali, Scienze della Formazione

İSTANBUL - ESENLER İLÇE MİLLİ EĞİTİM MÜDÜRLÜĞÜ

1474

STIAEM
Slow Tech Innovation
AI, Ethics and
New Teaching Methods

Co-funded by
the European Union

2025-1-IT02-KA220-SCH-000365213

eduOpen

CC BY-NC-SA 4.0 International

This license requires that reusers give credit to the creator. It allows reusers to distribute, remix, adapt, and build upon the material in any medium or format, for noncommercial purposes only. If others modify or adapt the material, they must license the modified material under identical terms.

Università di Foggia
Studi Umanistici, Lettere, Beni Culturali,
Scienze della Formazione

OER 3.1.2: Deploying systematic doubt as a pedagogical trigger for critical thinking
Reference Module: Module 3 - AI Ethics and Systemic Biases
Related Video-Pill: Video 3.1 - Black Box Systems, Opacity and Algorithmic Hallucinations
Authoring partner: Konak Science and Art Center (BILSEM) - Türkiye
Topic / Thematic Focus: Systematic doubt deployment, pedagogical trigger and critical thinking

Introduction and Learning Objectives

This resource accompanies video-pill 3.1 "Black Box Systems and Algorithmic Hallucinations" and completes the path begun with OER 3.1.1. Whereas there the governance of opacity was discussed, here the theme is systematic doubt: how to turn awareness of a system's limitations into a teaching device capable of training critical thinking, instead of merely issuing warnings.

The reference is to DigComp 3.0 (Cosgrove & Cachia, 2025), to competence 1.2 "Evaluating information", and to the UNESCO ethical framework (2021): educating for doubt is not generic distrust, but a form of responsibility towards oneself and towards the community. The perspective of Slow Tech and UDL reminds us that doubt must be made practicable by everyone, with clear and accessible devices.

By the end of the reading, teachers will be able to:

- explain the difference between intrinsic and extrinsic hallucination and recognise its diagnostic value in the classroom;
- identify the fluency heuristic as the mechanism that leads one to trust form more than content;
- design teaching activities that use systematic doubt as the engine of critical thinking.

1. What hallucinations are

A language model does not "search for" information: it generates the statistically plausible continuation of a text. When this continuation is not anchored in facts, we speak of hallucination. The scientific literature distinguishes two fundamental types, according to their relation to the source provided:

"(1) Intrinsic Hallucinations: The generated output that contradicts the source content" (Ji et al., 2023).

"(2) Extrinsic Hallucinations: The generated output that cannot be verified from the source content (i.e., output that can neither be supported nor contradicted by the source)" (Ji et al., 2023).

The distinction is valuable in the classroom. An intrinsic hallucination is a verifiable error: the model contradicts what we gave it, and this can be demonstrated by comparing the documents. An extrinsic hallucination adds information that the source does not contain: we can neither confirm nor refute it on the basis of the source, and precisely for this reason it must be treated with caution. Ji and colleagues put it precisely:

"Notably, the extrinsic hallucination is not always erroneous because it could be from factually correct external information" (Ji et al., 2023).

This subtlety is the pedagogical heart of the matter: a deviation may be correct, but the point is not whether it is; the point is that it is not verifiable from the source. Teaching how to distinguish means teaching how to distinguish verification from the feeling of truth.

"CS1.2.10: Recognise that AI systems may produce inaccurate outputs, even when apparently plausible, and that final responsibility for verifying the quality and validity of the information and content generated lies with the person using the AI system" (Cosgrove & Cachia, 2025).
The descriptor shifts the emphasis to a decisive point: final responsibility for verification lies with the person, not the machine. Systematic doubt is precisely the exercise of this responsibility. But be careful: our own mind also has its biases, and one of the most insidious is the tendency to confirm what we already believe. DigComp 3.0 invites us to recognise the intertwining of human prejudice and machine distortion: systematic doubt, then, is not only distrust of AI, but discipline of one's own judgement.
Here too the warning about the difficulty of reconstructing the internal mechanisms of models applies (Lipton, 2018).

2. The fluency heuristic: form is not proof

Why do we trust so easily? Models generate fluent texts: linear syntax, assertive tone, well-constructed connections. Faced with a cohesive text, our mind lowers its guard and attributes authority to it. This is the fluency heuristic: the ease with which content is processed is mistaken for an indication of truth.
The model does not "polish" its style in order to persuade us: style is the effect of the statistical objective with which it was trained, namely generating the most probable and coherent continuation. But the effect on the reader is that of a text that seems knowledgeable. The teaching consequence is clear: formal beauty is not proof of truth, and must be dismantled with exercises that make the difference between coherence and adherence visible.
In classroom practice, hallucinations take recognisable forms. There are invented citations: plausible but non-existent bibliographic references, with credible authors and journals. There is fabricated data: percentages, dates and statistics presented with confidence. There are mistaken attributions: true statements attributed to the wrong person. And there are non-existent causal links: correlations presented as causes. Knowing these types helps in designing targeted exercises: every form of hallucination, once made visible, becomes an opportunity to learn how to verify.

3. Systematic doubt as a teaching device

Systematic doubt is not scepticism: it is a procedure. It consists in suspending assent for the time necessary to carry out a verification, and in making this suspension a habit. Turning it into teaching means designing activities in which verification is the task, not an appendix. The literature on interpretability provides a foothold: the explanations offered by systems do not coincide with their workings, and therefore doubt must always be specified and anchored to something controllable.

An effective device is the hallucination hunt: asking the model for a list of references or data, then verifying every entry. The number of invented citations becomes a shared fact, not an opinion. A second device is the source audit: getting the student to formulate the question "how do you know?" and to demand a circumstantiated answer. A third is stylistic decomposition: isolating the rhetorical markers (the peremptory tone, the transition formulas) that give the sensation of rigour, in order to show their conventional nature.

The value of these devices lies in their transferability: they serve not only to expose a model, but to build a mental habit that holds in every domain. When a student learns to ask a machine "how do you know?", they are learning to ask it also of an article, a video, a classmate - and themselves. It is this generalisation that is the true teaching objective of systematic doubt.

4. Designing doubt: beware of misunderstandings

Two misunderstandings must be avoided. The first is relativism: if the machine is wrong, then "nobody is right". This is not so: systematic doubt exists precisely because criteria of verification exist, and its outcome is to distinguish, not to confuse. The second is demonisation: using AI only to show how fallible it is, turning an opportunity for learning into a lesson in distrust.

The stakes are educating for verification. UNESCO (2021) recalls the responsibility to govern the human impacts of AI systems; the European framework for teachers (Miao & Cukurova, 2024) places critical understanding of what AI does and does not do at the centre of professional competence. Systematic doubt is the teaching form that makes both operational.

Finally, a word on the relationship with assessment. Systematic doubt is not incompatible with rigour: on the contrary, it strengthens it. An assessment that asks students to distinguish a fact from a plausible invention, or to indicate the source of a statement, evaluates deeper competences than a closed-answer question. At the same time, teachers must accept that the path will be slower and noisier: at first, the class will produce more questions than answers. That is the sign that the method is taking root.

For high-impact decisions, the literature recommends intrinsically interpretable models rather than post-hoc explanations (Rudin, 2019).

The European ethical guidelines recall the need for effective human control over AI systems (HLEG, 2019).

5. From doubt to habit: the classroom routine

For doubt to become a habit, it must become a routine. Three fixed times can be established. Before: prediction. Before querying the AI, the student writes down what they expect to find; comparing expectation with response brings tacit assumptions to the surface. During: verification. A minimal protocol is applied - who says it, where, from which source, with what evidence. After: classification. Every deviation detected is noted and categorised (intrinsic or extrinsic), building a collective register of the model's fragilities.

The register of deviations has a double value: it is the shared memory of the class and a metacognitive tool. It shows that error is not to be hidden but studied, and that critical thinking is not an opinion but a method. In this way, on the teaching plane, systematic doubt becomes what transparency is on the technical plane: a practice of governance, not an attitude.

The teacher's role, in this framework, is not that of an arbiter who provides the correct answers, but of a coach of the method. This means resisting the temptation to say straight away whether an output is true or false, and instead asking students how they intend to verify it. The delay of the answer is the time in which competence is built. It is also a form of intellectual humility: showing that the teacher too verifies, errs and corrects.

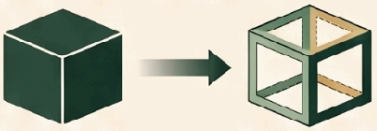
Type of hallucination	Relation to the source	How it is verified	Teaching use
Intrinsic	Contradicts the source provided	Direct comparison with the document	Demonstrable error: trains checking
Extrinsic	Cannot be confirmed or refuted from the source	Search for independent external sources	Opportunity to distinguish verification and plausibility
Fluency (effect)	Cohesive and assertive text	Decomposition of rhetorical markers	Makes the gap between form and truth visible

Visual Scaffolding

BEYOND THE BLACK BOX:
TEACHING CRITICAL THINKING THROUGH SYSTEMATIC DOUBT



1. THE CONCEPTUAL SHIFT



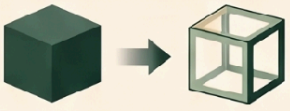
SYSTEM OPACITY → SYSTEMATIC DOUBT

From System Opacity to Systematic Doubt: Move beyond generic warnings about "black boves" by turning system limitations into structured pedagogical devices that train critical thinking.



The Responsibility of the Human Agent

Per DigComp 3.0 (CSI.3.10), the final responsibility for verifying the quality and validity of AI-generated content lies solely with the human user.



SYSTEM OPACITY → SYSTEMATIC DOUBT

The Responsibility of the Human Agent

Per DigComp 3.0 (CSI.2.10), the final responsibility for verifying the quality and validity of AI-generated content lies solely with the user.

2. DECODING ALGORITHMIC HALLUCINATIONS



INTRINSIC HALLUCINATION

Relationship to Source: Contradicts provided source content.

Verification Method: Direct comparison with the reference document.



EXTRINSIC HALLUCINATION

Relationship to Source: Output cannot be verified from the source (unsupported additions).

Verification Method: External checking and independent source searching.

3. THE PSYCHOLOGY OF TRUST



The Fluency Heuristic

Human vigliance lowers whan faced with assertive style, linear syntax, and cohesive prose, creating an illusion of authority regardless of truth.

Form vs. Truth

AI models generate statistically plausible continuations; formal beauty is a statistical effect, not proof of factual accuracy.



FORM TRUTH

4. THE TEACHER'S TOOLKIT: THREE CLASSROOM DEVICES



1. The Hallucination Hunt: Task students with verifying specific citations, dates, and statistics generated by the model to find "invented" data.

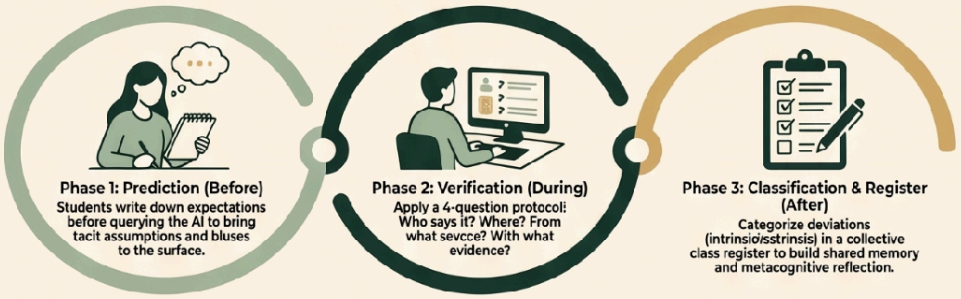


2. The Source Audit: Train students to demand evidence by asking the model "How do you know?" and requiring circumstantiated answers.



3. Stylistic Decomposition: Isolate and dismantle thetorical markers—like paremptory tones and transition formulas—to reveal the conventional nature of "expert" style.

5. THE 3-PHASE CLASSROOM ROUTINE



Phase 1: Prediction (Before)
Students write down expectations before querying the AI to bring tacit assumptions and bluses to the surface.

Phase 2: Verification (During)
Apply a 4-question protocol: Who says it? Where? From what sevcce? With what evidence?

Phase 3: Classification & Register (After)
Categorize deviations (intrinsic/extrinsic) in a collective class register to build shared memory and metacognitive reflection.

TEXT ALTERNATIVE SUMMARY

Summary: This infographic details a pedagogical framework for using AI lirvibnians to toach ethical thinking ths "Systismatic Doubt". it covers shifting from aqesity to verification responsibility, eccegerising intnissic/estrinsic hallucinations, understanding trust psychology, deplaying three teacher toers (Helloination Inera, Soures-Audis, Skylistic Decomposition), and implementating a 2-phese routine (Prediction, Verification, Classification), and aligning with global ethical standards like DigComp 2.0 and UNESCO.



6. FRAMEWORK ALIGNMENT

Global Ethical Standards

This approach aligns with DigComp 3.0 (evaluating informetion), UNESCO AI CFT (human verification), and Slow Tech Principles (practicable doubt).

Gemini Notebook

Conclusions and Metacognitive Reflection Prompts

Few-shot prompting shows that examples can become powerful mechanisms for structuring AI Systematic doubt is the most appropriate teaching response to systems that generate plausible texts without any guarantee of truth. Distinguishing intrinsic from extrinsic hallucination, recognising the fluency heuristic and translating all of this into classroom routines means turning a technical limitation into a powerful learning device. Critical thinking is not taught with generic warnings, but with the repeated practice of verification.

To bring these core ideas into your own teaching:

- Which activity in your subject lends itself to becoming a "hallucination hunt" (lists, data, citations to be verified)?
- How would you make the fluency heuristic visible to your students without trivialising it into a moral warning?
- What form could a shared register of the model's deviations take in your classroom?

Visual scaffolding

Technical Metadata Sheet

Metadata Field	Specification / Parameter Value
Identifier	OER 3.1.2
Title	Deploying Systematic Doubt as a Pedagogical Trigger for Critical Thinking
Creator	Konak Science and Art Center (BILSEM) - Türkiye
Description	Complementing OER 3.1.1, the resource turns awareness of algorithmic limits into teaching practice. It distinguishes intrinsic and extrinsic hallucination and their diagnostic value, and identifies the fluency heuristic as the mechanism that makes form more persuasive than content. The pedagogical focus is on designing activities that use systematic doubt as the engine of critical thinking and turn it into a classroom routine.
Keywords	systematic doubt; critical thinking; intrinsic hallucination; extrinsic hallucination; fluency heuristic; questioning routines; pedagogical design; AI literacy
Language	EN (English)
Affiliation	Konak Science and Art Center (BILSEM) - Türkiye
Contributor	STIAEM Transnational Scientific Board (peer review, Rhodes validation protocol)
Media Format	Text; Infographics
Genre	Training Manual; Educational Guide
URL	
Citation	Cosgrove, J., & Cachia, R. (2025). DigComp 3.0: European digital competence framework (Fifth edition). Publications Office of the European Union. https://data.europa.eu/doi/10.2760/0001149 European Commission. (2019). Ethics guidelines for trustworthy AI. High-Level Expert Group on Artificial Intelligence. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Chan, H. S., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730 Miao, F., & Cukurova, M. (2024). AI competency framework for teachers. UNESCO. https://doi.org/10.54675/ZJTE2084

	UNESCO. (2021). Recommendation on the ethics of artificial intelligence. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000381137
Educational Level	Secondary School
Associated Module / MOOC Video Lesson	Module 3 - AI Ethics and Systemic Biases / Video 3.1 - Black Box Systems, Opacity and Algorithmic Hallucinations
Learning Objectives	explain the difference between intrinsic and extrinsic hallucination and recognise their diagnostic value in the classroom; identify the fluency heuristic as the mechanism that leads one to trust form more than content; design teaching activities that use systematic doubt as the engine of critical thinking.
Rights	Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
Date Available	14/09/2026
Date Modified	

References and Webography

Cosgrove, J., & Cachia, R. (2025). DigComp 3.0: European digital competence framework (Fifth edition). Publications Office of the European Union. https://data.europa.eu/doi/10.2760/0001149

European Commission. (2019). Ethics guidelines for trustworthy AI. High-Level Expert Group on Artificial Intelligence. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Chen, D., Dai, W., Chan, H. S., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), Article 248. https://doi.org/10.1145/3571730

Lipton, Z. C. (2018). The mythos of model interpretability. Communications of the ACM, 61(10), 36-43. https://doi.org/10.1145/3233231

Miao, F., & Cukurova, M. (2024). AI competency framework for teachers. UNESCO. https://doi.org/10.54675/ZJTE2084

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215. https://doi.org/10.1038/s42256-019-0048-x

UNESCO. (2021). Recommendation on the ethics of artificial intelligence. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000381137.