STIAEM

Slow Tech Innovation

AI, Ethics and New Teaching Methods

Managing 'Black Box' Systems: Governing Opacity to Safeguard Intellectual Freedom

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

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

İSTANBUL - ESENLER İLÇE MİLLİ EĞİTİM MÜDÜRLÜĞÜ

1474

OER 3.1.1: Managing 'Black Box' Systems: Governing Opacity to Safeguard Intellectual Freedom
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: Few-shot prompting strategies, pre-solved exemplar models and output stability

Introduction and Learning Objectives

This resource accompanies video-pill 3.1 "Black Box Systems and Algorithmic Hallucinations", together with OER 3.1.2, and provides its conceptual in-depth analysis. Its aim is not to teach how to "explain" artificial intelligence systems, but to put teachers in a position to govern their opacity: to recognise where it lies, to distinguish its levels and to translate it into demands and defensible teaching choices. Opacity, as we shall see, is not an inevitable technical destiny, but a relationship that can be safeguarded.

The text connects to DigComp 3.0 (Cosgrove & Cachia, 2025), in particular to competence 1.2 "Evaluating information", and to the perspective of *Slow Tech* and Universal Design for Learning: if we design for everyone's understanding, transparency becomes a teaching requirement, not a technical detail. The reference to the "AI pedagogy" pillar of the UNESCO Framework for teachers (Miao & Cukurova, 2024) completes the picture.

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

- distinguish the three levels of transparency - simulability, decomposability and algorithmic transparency - and not reduce opacity to a single metaphor;
- recognise the difference between a post-hoc explanation and real transparency, using it as a criterion for evaluating tools;
- translate the right to explanation into concrete demands on providers and into teaching choices that protect students' intellectual freedom.

1. The "black box" is not an object: it is a relationship

When we say that an AI system is a "black box", we mean that we cannot reconstruct its internal workings by looking at it from the outside. The expression, however, is ambiguous and can become a shortcut. Zachary Lipton, in the essay that opened the contemporary debate on interpretability, shows that the term "interpretability" does not denote a single concept but gathers distinct ideas that are used without being defined:

"...the term interpretability is ill-defined, and thus claims regarding interpretability of various models may exhibit a quasi-scientific character" (Lipton, 2018).

From this follows an important methodological consequence for schools. Saying that a system is opaque "without specifying with respect to what" does not help to govern it: it is like saying that a door is closed without saying from which side. The teacher's first task is therefore to make the question precise: opaque for whom, with respect to which aspect, with a view to which decision. Not a generic judgement, but a circumstantiated diagnosis.

Why is a system opaque? The reasons are not a mystery, and knowing them helps us not to mythologise them. The first is non-linearity: deep models combine millions or billions of parameters through non-linear functions, and the contribution of each parameter to the

result depends on the others. The second is dimensionality: internal representations live in spaces with hundreds or thousands of dimensions, impossible to visualise directly. The third is dependence on data: the model's behaviour is not written in the code, but emerges from the distribution of the examples on which it was trained. None of these causes is ethically neutral: precisely because opacity is structural, transparency must be demanded as a design choice, not awaited as a natural effect.

2. Three levels of transparency

Lipton proposes distinguishing transparency by levels, each of which answers a different question:

"Informally, transparency is the opposite of opacity or blackbox-ness. It connotes some sense of understanding the mechanism by which the model works. We consider transparency at the level of the entire model (simulatability), at the level of individual components (e.g. parameters) (decomposability), and at the level of the training algorithm (algorithmic transparency)" (Lipton, 2018).

Simulability concerns the possibility of contemplating the entire model at a glance: a simple model "can be held in one's head". Decomposability concerns the possibility of understanding each component or parameter separately. Algorithmic transparency concerns training: understanding which algorithm, with which objective function, produced the final calibration. A model can be transparent at one level and opaque at another, and this has practical consequences when choosing a tool for the classroom.

A second element clarifies why the issue is so keenly felt:

"The demand for interpretability arises when there is a mismatch between the formal objectives of supervised learning (test set predictive performance) and the real world costs in a deployment setting" (Lipton, 2018).

In other words: the machine is optimised for a statistical objective (accuracy on a test set), whereas the real decision has human costs - a grade, a guidance recommendation, an admission - that do not enter the metric. It is precisely in that gap that the need for transparency arises and, with it, the responsibility of those who decide.

These levels help us understand why deep networks are opaque in a specific way. In terms of simulability they are inaccessible, because no one can contemplate billions of parameters at once; in terms of decomposability they are equally inaccessible, because the meaning of a single weight depends on the entire network. Algorithmic transparency remains: we know, in principle, which optimisation procedure was used, but we do not always know the exact data and the design choices that shaped it. This is why, for schools, the useful question is not "is this system transparent or not?", but "at which level, and with what consequences for the decision I have to take?".

3. Post-hoc explanations: the risk of the halo effect

Many tools offer "explanations": heat maps, word highlighting, natural-language sentences. They are useful, but Lipton warns that they must not be confused with transparency:

"post-hoc interpretations often do not elucidate precisely how a model works" (Lipton, 2018).

A post-hoc explanation can describe a prediction without clarifying the mechanism that produced it. The pedagogical risk is the halo effect: the presence of an explanation reassures and induces trust, even when the explanation is plausible but misleading. Lipton says it plainly:

"We caution against blindly embracing post-hoc notions of interpretability, especially when optimized to placate subjective demands. In such cases, one might - deliberately or not - optimize an algorithm to present misleading but plausible explanations" (Lipton, 2018).

For teachers this translates into a rule of cognitive hygiene: always ask at which level the explanation is being given, and do not mistake the elegance of the explanation for proof of reliability.

A borderline case of this misplaced trust is hallucination: fluent and plausible outputs that do not correspond to any verifiable source (Ji et al., 2023).

4. Designing Effective Pre-Solved Exemplars

DigComp 3.0 translates these themes into descriptive competences. Two are central:

"CS1.2.09: Recognise that some digital technologies, such as AI systems, may work as a 'black box', making it difficult to explain why or how a given output was produced" (Cosgrove & Cachia, 2025).

"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 European regulatory framework adds a dimension of rights. The ethics guidelines for trustworthy AI (HLEG, 2019) and the ALTAI set out its applications; in particular, Cynthia Rudin (2019) argues that for high-impact decisions it is preferable to use interpretable models "by design", rather than attempting to explain opaque models post hoc. This is not always possible, but it is a preference that schools can and must set as a preliminary criterion.

"CS1.2.08: Recognise that the data used to train AI systems and the ways in which they are trained affect the reliability of the information they provide" (Cosgrove & Cachia, 2025).

"CS1.2.11: Recognise that human prejudices (cognitive and emotional) and distortions (bias) of AI systems (in the data or in training) play a role in the generation and interpretation of information" (Cosgrove & Cachia, 2025).

Taken in sequence, these descriptors outline a competent path: understanding that data matter, that systems can be opaque, that output must be verified and that human and algorithmic bias are intertwined. Transparency, in this perspective, is not an isolated technical property but a weave of awarenesses that teachers are called to possess and to teach.

5. Few-Shot Prompting Through the STIAEM Pedagogical Framework

Translating awareness into practice means putting governance safeguards in place. We indicate four. First: demand technical documentation when adopting a tool, with training data, validity metrics and declared limitations. Second: claim the right to explanation, asking the provider not only what the system does, but how it enables the user to interpret the output. Third: carry out a preliminary impact assessment, verifying distortions with respect to one's own school population. Fourth: favour, where possible, transparent solutions for decisions that genuinely affect students' pathways.

In the classroom this takes the form of education in the critical reading of tools: pointing out to students when an assistant cannot explain its choices, inviting them to ask for sources, showing that opacity is not a magical mystery but a circumstantiated design feature. Protecting students' intellectual freedom therefore means teaching them that they have the right not to accept an output simply because it came out of a machine, and giving them the words to say so.

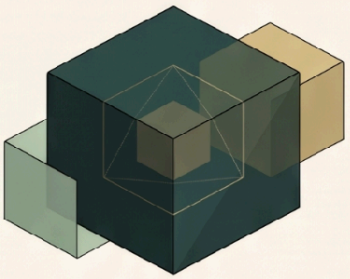
The last point concerns professional autonomy. Knowing that a system is opaque does not mean having to renounce it: it means adopting it with awareness, deciding which decisions to delegate to it and which to reserve for human judgement. Slow Tech offers a compass here: technology is assessed by the human consequences it produces, not by its novelty. If an automated proofreader, for example, pushes students towards a standardised register, impoverishing their voice, the human cost outweighs the technical benefit, and the choice to limit its use is a legitimate professional choice, not a step backwards.

Level of transparency	What I ask for	Guiding question	Risk if absent
Simulability	Understanding the entire model at once	Can I hold it entirely in my head?	Blind trust in very complex models
Decomposability	Understanding each component/parameter	Do I know which part produces which effect?	Impossible to attribute an outcome to a cause
Algorithmic transparency	Understanding the training and the objective function	Do I know how it was calibrated?	The mismatch between metric and decision remains invisible
Post-hoc explanation	Obtaining a description of the prediction	Does it explain the mechanism or only the outcome?	Halo effect: plausible but misleading

Visual Scaffolding

GOVERNING THE BLACK BOX:

Safeguarding Intellectual Freedom in the AI Classroom

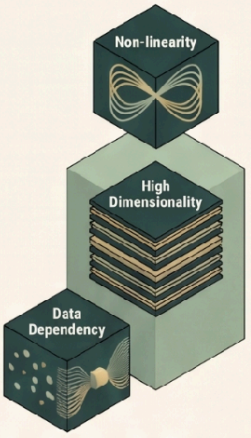


THE 'BLACK BOX' IS A RELATIONSHIP

Opacity is a structural relationship, not a temporary glitch.

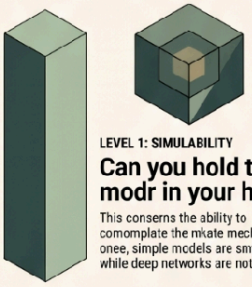
We call a system a "black box" when we cannot reconstruct its internal workings from the outside; governing it requires asking "opaque for whom and for what decision?"

THE THREE PILLARS OF OPACITY



Opacity stems from **Non-linearity**, **High Dimensionality**, and **Data Dependency.**

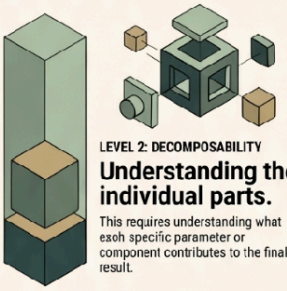
LIPTON'S THREE LEVELS OF TRANSPARENCY



LEVEL 1: SIMULABILITY

Can you hold the modr in your head?

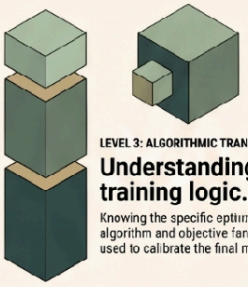
This conserns the ability to comomplate the mkate mechanism at onee, simple models are smuletable, while deep networks are not.



LEVEL 2: DECOMPOSABILITY

Understanding the individual parts.

This requires understanding what exoh specific parameter or component contributes to the final result.



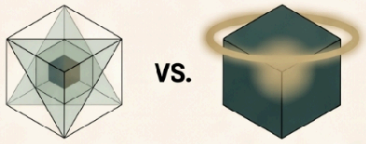
LEVEL 3: ALGORITHMIC TRANSPARENCY

Understanding the training logic.

Knowing the specific optimization algorithm and objective function used to calibrate the final model.

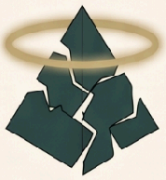
THE 'HALO EFFECT' (POST-HOC VS. REAL TRANSPARENCY)

COMPARISON: REAL TRANSPARENCY VS. POST-HOC EXPLANATIONS



Don't mistake elegance for reliability.

Post-hoc interpretations describe a prediction but often do not clarify the actual mechanism that produced it.



THE RISK OF MISLEADING TRUST

The Halo Effect can mask algorithmic "hallucinations."

Plausible-sounding explanations can induce false trust, leading users to accept fluent but inaccurate outputs that lack verifiable sources.

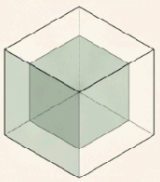
RIGHTS, FRAMEWORKS, AND HUMAN RESPONSIBILITY



DIGCOMP 3.0 COMPETENCIES

Recognizing black box mechanisms is a core digital skill.

Frameworks CS1.2.09 and CS1.2.10 emphasize recognizing opacity and maintaining final human responsibility for verifying AI-generated content.



INTERPRETABLE-BY-DESIGN

Favor interpretable models for high-impact decisions.

Following Rudin (2019), schools should prioritize models that are inherently understandable over attempting to "explain" opeque models after the fact.

FOUR GOVERNANCE SAFEGUARDS FOR SCHOOLS

1. DEMAND DOCUMENTATION	2. CLAIM THE RIGHT TO EXPLANATION	3. IMPACT ASSESSMENT	4. PRIORITIZE INTELLECTUAL FREEDOM
Require technical proof from providers.	**Ask "How" not just "What."**	**Verify distortions for your specific student population.**	**Reserve high-stakes decisions for transparent tools.**
Ask for training data cuts, validity mottles, and declared limitations before arlopting any classroom tool.	Providers must explain how their tool enables users to interpret outputs, rather than just delivering a result.	Conduct preliminary checks to ensure the test does not introduce blas or herm student intellectual freedom.	Protect student pathways by ensuring that any ñt used for signifist significant is fully transparent and gevemable.

Gemini Notebook

Conclusions and Metacognitive Reflection Prompts

The opacity of AI systems is not a single, elusive property, but a set of distinct levels that can be named. Naming them is the first act of governance: it allows precise demands to be made on providers, prevents an elegant explanation from being mistaken for proof of reliability, and enables the choice, where it matters, of more transparent tools. This is how schools protect students' intellectual freedom: not by rejecting technology, but by insisting on understanding it well enough not to be subjected to it.

To transfer these core ideas into your own practice:

- Which AI tool do you use, or would you like to use, in the classroom, and at which of the three levels of transparency do you feel able to defend it?
- When an assistant shows you an explanation, can you distinguish the mechanism from the halo effect? How could you train yourself to do this with your students?
- What concrete request for documentation could you make, as early as tomorrow, to your institution or to a provider?

Technical Metadata Sheet

Metadata Field	Specification / Parameter Value
Identifier	OER 3.1.1
Title	Managing 'Black Box' Systems: Governing Opacity to Safeguard Intellectual Freedom
Creator	Konak Science and Art Center (BILSEM) - Türkiye
Description	The resource does not aim to teach how to explain AI systems but to enable teachers to govern their opacity. It distinguishes three levels of transparency - simulability, decomposability and algorithmic transparency - and contrasts post-hoc explanations with real transparency, warning against the halo effect. The practical focus is on translating the right to explanation into concrete demands on providers and into classroom choices that protect students' intellectual freedom.
Keywords	black box; opacity; transparency levels; explainability; post-hoc explanations; halo effect; intellectual freedom; AI governance
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 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
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	distinguish the three levels of transparency - simulability, decomposability and algorithmic transparency - and not reduce opacity to a single metaphor; recognise the difference between a post-hoc explanation and real transparency, using it as a criterion for evaluating tools; translate the right to explanation into concrete demands on providers and into teaching choices that protect students' intellectual freedom.
Rights	Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)
Date Available	14/09/2026
Date Modified	

References

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