



AI Guide

ONLINE COURSE CONTENT

MODULE 2: AWARENESS OF LIMITATIONS, BIAS AND TRANSPARENCY

2025-1-LT01-KA220-HED-000354784

Artificial Intelligence for Green Understanding, Insights & Decision Empowerment



SQUAREDEV



Co-funded by
the European Union

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 National Agency of Erasmus+ Programme in Lithuania – Education Exchanges Support Foundation (ŠMPF). Neither the European Union nor the ŠMPF can be held responsible for them.

Table of Contents

Awareness of Limitations, Bias and Transparency.....	3
A. General Information	3
Introduction.....	3
<i>Understanding the Limitations of AI Systems.....</i>	<i>3</i>
<i>Bias in AI-Generated Answers.....</i>	<i>4</i>
<i>Transparency in AI answers</i>	<i>5</i>
References.....	5
Green Decision Context	6



Co-funded by
the European Union

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 National Agency of Erasmus+ Programme in Lithuania – Education Exchanges Support Foundation (SMPF). Neither the European Union nor the SMPF can be held responsible for them.



Awareness of Limitations, Bias and Transparency

A. General Information

Module No	Module2
Subtopic	Awareness of limitations, bias and transparency
AI Literacy Component	AI bias, AI limitations, black-box problem, transparency, and explainability.
Decision Making Skills	Critical judgement, bias detection, fairness and accountability evaluation.
Learning Objectives for STEM HE Teachers	Explain AI limitations, bias mechanisms and transparency issues relevant to STEM and green decisions.
Learning Objectives for STEM Students	Identify bias, limitations and lack of transparency in AI-supported outputs.
Suggested AI Tools	Model cards; dataset documentation; chatbots;
Estimated Duration	60 minutes

Introduction

An essential aspect of AI literacy is understanding that AI systems are probabilistic models rather than repositories of verified knowledge. They predict the most likely sequence of words based on patterns learned from vast amounts of training data. Consequently, AI-generated answers may contain factual inaccuracies, incomplete information, logical inconsistencies, outdated knowledge, or unintended biases.

Teachers, researchers, and students must therefore develop awareness of three fundamental concepts:

- limitations – what AI systems can and cannot do;
- bias – how AI may systematically favor certain viewpoints;
- transparency – understanding how AI generates answers, what information it uses, and how confident users should be in its output.

Developing this awareness enables users to critically evaluate AI-generated information rather than accepting it uncritically.

Understanding the Limitations of AI Systems

One of the most common misconceptions is that AI "knows" information in the same way humans do. In reality, modern language models do not possess human understanding or factual knowledge. Instead, they generate responses by predicting statistically probable word sequences learned during training. For example, when asked:

Who discovered penicillin?

the AI generates the answer "Alexander Fleming" because this association frequently appears in its training data, and not because it independently verifies historical facts.

This distinction is crucial because AI cannot correctly distinguish between truth and falsehood unless supported by reliable information during training or connected to external knowledge sources.

Although AI often appears capable of sophisticated reasoning, much of its performance results from recognizing patterns rather than genuine logical understanding. AI may successfully solve many reasoning problems but can also fail on relatively simple tasks when [1]:

- multiple logical steps are required;
- assumptions are ambiguous;
- contextual information is missing;
- mathematical precision is essential.

For example, an AI may correctly explain Newton's laws but incorrectly solve a multi-step physics problem if intermediate reasoning is flawed.

Perhaps the best-known limitation of generative AI is hallucination [2]. An AI hallucination occurs when the system confidently produces information that is false, fabricated, or unsupported by evidence. Examples include [2]:

- inventing scientific references;
- fabricating quotations;
- creating non-existent legal cases;
- generating incorrect statistics;
- describing fictional events as real.

A particularly dangerous aspect is that hallucinated answers are often presented with high linguistic confidence, making them difficult for inexperienced users to recognize. For example:

"According to Smith et al. (2022)..."

may appear entirely credible even though no such publication exists.

Hallucinations become more likely when users ask highly specialized, obscure, or recent questions for which reliable information is limited.

AI models are trained on data collected during specific time periods. Unless connected to external search systems or live databases, they cannot automatically know:

- recent scientific discoveries;
- current legislation;
- breaking news;
- newly released products;
- recent policy changes.

Consequently, responses may accurately reflect historical knowledge while omitting recent developments.

Although modern AI systems can process long conversations, they still have practical context limits [3]. Long interactions may lead to [4]:

- forgotten earlier instructions;
- inconsistent answers;
- contradictions;
- loss of important details.

For lengthy projects, users should periodically restate objectives and verify that AI continues following the intended requirements.

Bias in AI-Generated Answers

Bias refers to systematic tendencies that cause AI-generated responses to favor particular perspectives, assumptions, or groups over others [5]. Bias is not always intentional. Instead, it often



Co-funded by
the European Union

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 National Agency of Erasmus+ Programme in Lithuania – Education Exchanges Support Foundation (SMPF). Neither the European Union nor the SMPF can be held responsible for them.



reflects patterns present within the data used to train the model [5]. Because AI learns from billions of documents written by humans, it inevitably inherits some characteristics of human language, culture, and society.

Several factors contribute to AI bias [6]:

- Training data bias – if certain viewpoints appear more frequently in training data, AI is more likely to reproduce them;
- Historical bias – training datasets include historical documents;
- Selection bias – not all information available on the Internet is included in AI training;
- Prompt bias – the phrasing of a user's question can itself introduce bias.

Types of Bias in AI Response [7]:

- Cultural bias – responses may reflect assumptions common in particular regions or societies while overlooking alternative cultural perspectives;
- Language bias – high-resource languages generally receive better support than less commonly spoken languages;
- Confirmation bias – if users ask leading questions, AI may reinforce the assumptions embedded in those questions;
- Availability bias – frequently discussed topics receive richer and more detailed responses than subjects with limited documentation;
- Popularity bias – widely accepted opinions often dominate AI outputs even when legitimate minority viewpoints exist.

Transparency in AI answers

In the context of AI-generated answers, transparency refers to the extent to which users understand how an AI system produces its responses, what information the responses are based on, and what limitations or uncertainties may affect their reliability. Unlike traditional software, where the logic behind outputs can often be explicitly traced, modern generative AI systems operate through complex neural networks that learn statistical relationships from vast amounts of data. As a result, the internal decision-making process is not fully interpretable, making transparency particularly important for responsible AI use.

From a user's perspective, transparency means being informed about the nature of AI-generated content. Users should understand that AI systems do not "know" facts in the human sense, nor do they independently verify every statement they generate. A transparent AI system should communicate the limitations of its answers rather than presenting every response as unquestionably correct.

An important aspect of transparency is the communication of confidence and uncertainty. Well-designed AI systems increasingly acknowledge when information may be uncertain, incomplete, or dependent on interpretation. For example, an AI may indicate that evidence on a particular topic is limited, that multiple scientific viewpoints exist, or that recent developments could have changed the available knowledge. Such statements help users distinguish between well-established facts and areas where additional verification is recommended. Transparency therefore encourages informed decision-making rather than blind trust in AI-generated outputs.

References:

[1] Hadi, M. U., Qureshi, R., Shah, A., Irfan, M., Zafar, A., Shaikh, M. B., ... & Mirjalili, S. (2023). Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea preprints*, 1(3), 1-26.

- [2] Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 1-55.
- [3] Kuratov, Y., Bulatov, A., Anokhin, P., Rodkin, I., Sorokin, D., Sorokin, A., & Burtsev, M. (2024). Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *Advances in Neural Information Processing Systems*, 37, 106519-106554.
- [4] Wang, X., Salmani, M., Omid, P., Ren, X., Rezagholizadeh, M., & Eshaghi, A. (2024). Beyond the limits: A survey of techniques to extend the context length in large language models. *arXiv preprint arXiv:2402.02244*.
- [5] Wei, X., Kumar, N., & Zhang, H. (2025). Addressing bias in generative AI: Challenges and research opportunities in information management. *Information & Management*, 62(2), 104103.
- [6] Ranjan, R., Gupta, S., & Singh, S. N. (2024). A comprehensive survey of bias in LLMs: Current landscape and future directions. *arXiv preprint arXiv:2409.16430*.
- [7] Tiwari, S. (2025). Biases and Fairness in LLMs. In *Generative AI: Techniques, Models and Applications* (pp. 229-242). Cham: Springer Nature Switzerland.

Green Decision Context

AI is increasingly used to support green and sustainability-related decisions, such as:

- analyzing environmental and climate data
- recommending sustainable farming and resource management practices
- supporting renewable energy planning and energy-efficiency measures
- assessing environmental impacts of projects and technologies
- assisting in biodiversity conservation and ecosystem management

When using AI for green decision-making, users should understand that:

- AI may generate incorrect or oversimplified explanations of environmental processes.
- AI may reflect biases present in its training data, favoring certain regions, technologies, or viewpoints.
- AI may not include the latest environmental regulations, climate research, or sustainability standards.
- AI recommendations should be interpreted in the context of local environmental conditions and expert knowledge.
- Transparency about the assumptions, uncertainties, and limitations of AI outputs is essential for responsible environmental decision-making.