expert insights

Responsible AI Lives in Gray Boxes

Vlad Goodkovsky, PhD

LEONARDO • Institute

Q: What are the Responsible AI considerations for using generative AI in learning and instructional design?

A: There's potential and challenges in incorporating generative AI, like ChatGPT, into learning and instructional design.

Understandable AI Models as Part of Responsible AI

AI models should be interpretable (understandable for users) for effective use in learning and teaching. The idea of the unknowable nature of the human mind was referred to as a black box in the field of psychology. When using an AI tool in education, the black box approach, where the internal workings are not transparent to users, is suboptimal. The need for AI systems that users can understand and trust is a critical aspect of integrating AI into learning environments.

Augmenting AI with Systemic Frameworks

Augmenting AI with Systemic Frameworks involves enhancing AI's transparency (clarity in how AI makes decisions), interpretability (ease of understanding AI processes by humans), and responsibility (accountability for AI's actions and decisions), along with its effectiveness and efficiency. This is achieved by integrating AI with systemic frameworks (holistic approaches considering all aspects of a system), structure algebras (mathematical structures for organizing data), inferences (logical deductions made by AI), and generative engines (systems that create new content or solutions). This augmentation strives to humanize AI, creating a synergy between human and artificial intelligence, thus improving its applicability in educational contexts.

Developing Responsible AI-Infrastructure for ChatGPT Utilization

The development of an infrastructure that enables users to create effective prompts for ChatGPT and validate its responses. This includes a deep study of human learning theories and exploring interdisciplinary models like systems theory. The goal is to create a more precise and structured application of AI in learning.

Q: How can an organization transition from Blackbox to Glassbox Models?

A: A significant part of the discussion is dedicated to transitioning from the black box approach (where the AI's decision-making process is opaque) to a glass box approach (which is more transparent and explainable). I suggest a gray box approach, combining the strengths of both, to improve the clarity and responsibility of AI systems. This approach is seen as essential for integrating AI into learning design, making it more systematic and interpretable. The glass box approach, in contrast to the black box model, refers to systems (especially in artificial intelligence and machine learning) where the internal workings are transparent and understandable to users.

Transparency in Glass Box Models

In a glass box model, the processes and decision-making criteria within the system are clear and open to inspection. This transparency allows users, developers, and regulators to see and understand how the system arrives at its conclusions or actions. It’s like looking through a glass box where the internal mechanisms are visible.

Relation to Black Box Models

The black box approach, often used in complex AI models like deep neural networks, is characterized by its opacity; the internal logic or the way it processes inputs to produce outputs is not easily discernible. In contrast, the glass box approach aims to make these processes intelligible. While black box models can be highly effective and powerful, their lack of transparency can be a drawback, especially in fields where understanding the decision-making process is crucial (like healthcare, finance, or law).

Importance in AI and Machine Learning

The shift from black box to glass box models in AI is driven by the need for accountability, trust, and ethical considerations. As AI systems become more integrated into critical decision-making processes, understanding how these systems make decisions becomes vital. This is particularly important in situations where AI decisions need to be explained or justified, such as in credit scoring, medical diagnoses, or legal decisions.

Implementing Responsible AI Glass Box Approaches

Achieving a glass box model can involve various techniques like using simpler, more interpretable models (e.g., decision trees instead of deep neural networks), or developing methods to interpret and explain the decisions of more complex models. It also involves designing AI systems with explainability as a core feature from the ground up, rather than treating it as an afterthought. The glass box approach aims for clarity and understandability in AI systems, standing in contrast to the opaque nature of black box models. It’s a move towards making AI more accountable, ethical, and user-friendly, enabling a broader range of individuals to understand and trust AI-driven decisions.

Q: Can Responsible AI live in a Glass Box?

A: The feasibility and reasonableness of implementing a glass box approach in AI depend on various factors, including the complexity of the AI system, the specific application, and the current state of technology.

Considerations for Responsible AI

Complexity of AI Models Simpler AI models, such as decision trees or linear regression, are inherently more interpretable and can naturally fit the glass box approach. However, more complex models like deep neural networks, which are often used for tasks involving large datasets and require a higher level of abstraction, are more challenging to make transparent.

Advancements in Explainable AI (XAI)

There is a growing field of research known as Explainable AI, which focuses on making AI decisions more understandable to humans. This includes developing new algorithms and methods to explain and interpret complex models. The progress in this field is promising, but it's still an active area of research.

Trade-off Between Performance and Interpretability

Often, there's a trade-off between the performance of an AI model and its interpretability. Complex models that provide high accuracy (like deep learning models) are typically less interpretable. Achieving a balance between these two aspects is a key challenge in the development of glass box models.

Regulatory and Ethical Considerations

In fields where understanding the decision-making process of AI is crucial (e.g., healthcare, finance, legal), a glass box approach is not just desirable but may be required by regulations. In these cases, developing interpretable models is essential, even if it means compromising somewhat on performance.

User Trust and Adoption

Transparency and interpretability in AI systems can increase user trust and facilitate wider adoption, especially in critical applications. Users and stakeholders are more likely to trust and rely on AI systems if they can understand how decisions are made.

Technical Feasibility of Responsible AI

While it's challenging, making AI systems more interpretable is technically feasible. The implementation might involve combining complex models with interpretability layers or using inherently interpretable models for certain tasks. While a glass box approach is challenging, especially for complex AI systems, it is increasingly seen as necessary and feasible, particularly for applications where transparency and accountability are critical. The balance between interpretability and performance, and the continuous advancements in XAI, will shape the future of how this approach is implemented in various AI applications.

Practical Challenges and Solutions

There are practical challenges like the AI's tendency to 'hallucinate' and the necessity of human oversight for prompt engineering and result verification. The 'gray box' approach is proposed as a solution to these challenges, facilitating a smoother integration of AI into learning systems. Overall, I advocate for a nuanced, responsible, and interpretable integration of AI into learning design, underscoring the need to bridge the gap between AI technology and its practical application in education.