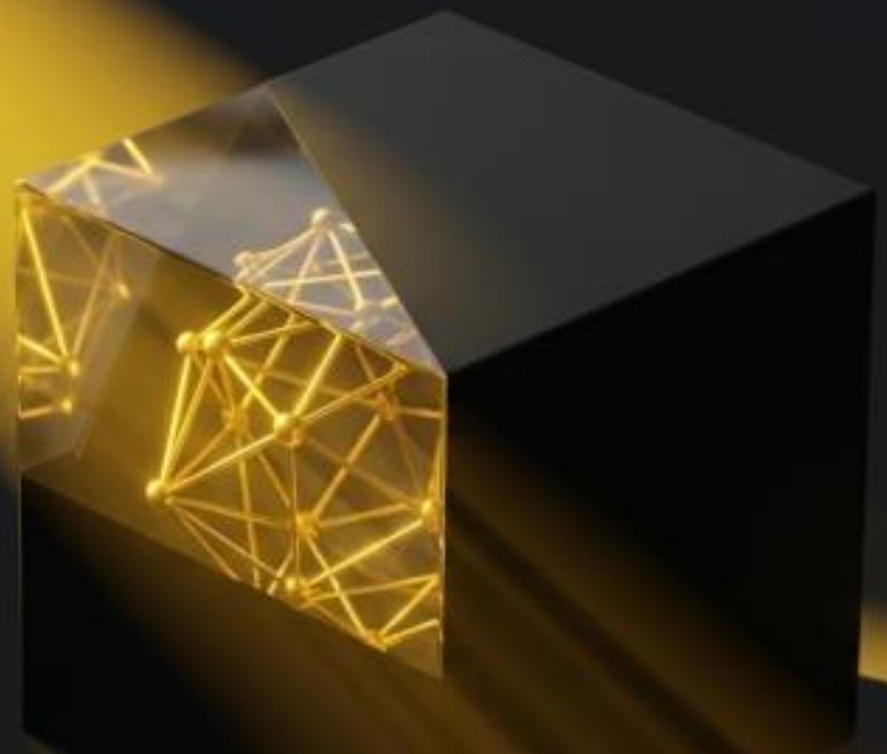
The Illumination of the Black Box

Medium opacity

Bridging the Gap Between Artificial Intelligence and Human Understanding

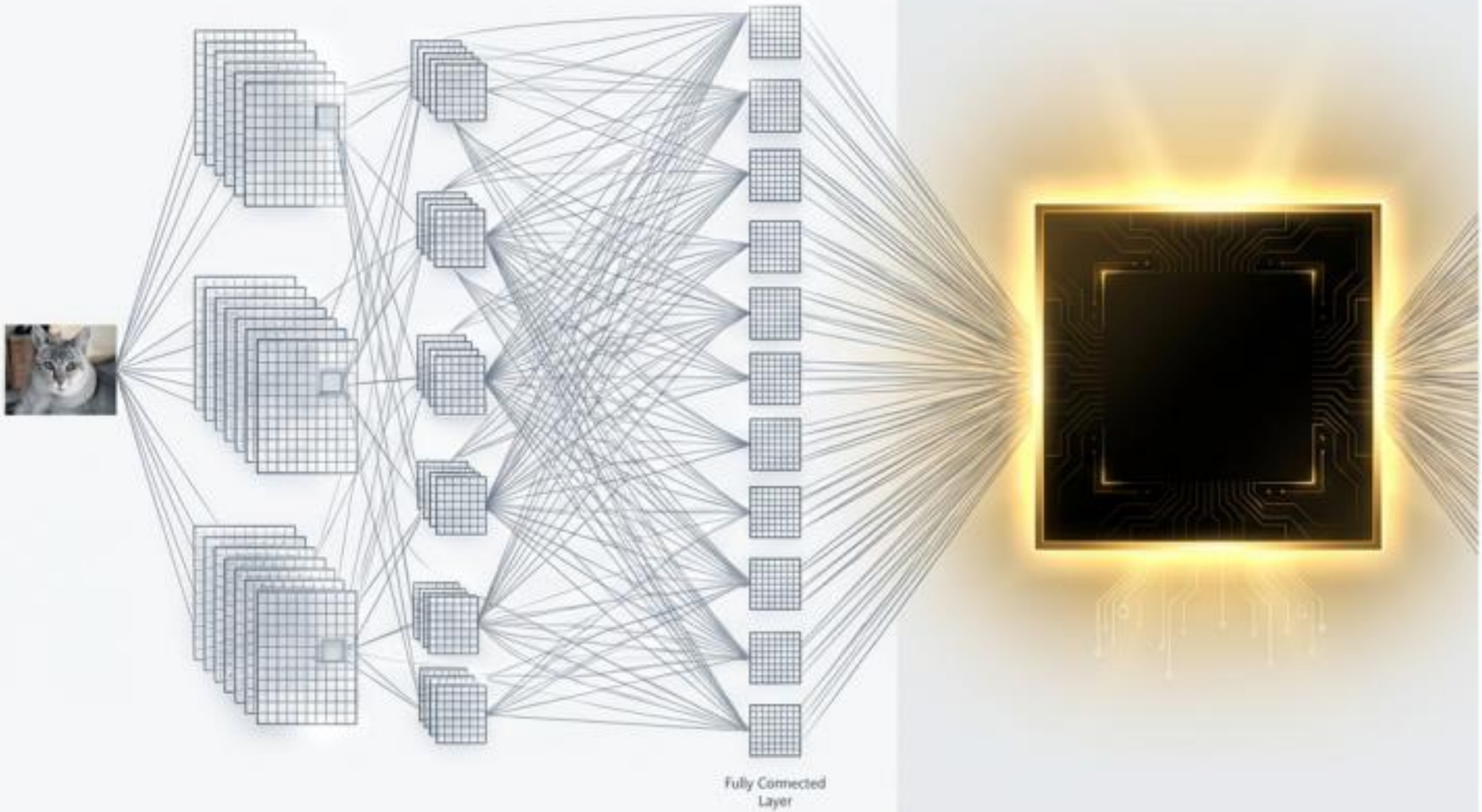


- Pr. Imen JDEY
- Associate Professor in Computer Science

The Outline

- Motivation & Context: The AI Black Box Problem
- Why Explainability Matters (Trust, Ethics, Regulation)
- Overview of Explainable AI (XAI) Approaches
- Key Methods & Techniques (Model-agnostic vs Model-specific)
- Applications & Case Studies
- Current Challenges & Limitations
- Future Directions & Conclusion

The Paradox of Power and Opacity



Deep Learning solves NP-hard problems with massive datasets (10M+ images).

But as complexity rises, transparency falls.

We have built systems that work, but we cannot see HOW they work.

The Trust Deficit: When the Black Box Fails

Healthcare.
IBM Watson (2011) rejected by hospitals. Doctors could not trust decisions they couldn't understand.

Military & Defense.
Autonomous systems must distinguish combatants from civilians. 'Why' is a matter of life and death.

Algorithmic Bias.
From sentencing disparities to recognition errors. Opaque systems amplify systemic data biases.

The Global Mandate for Transparency

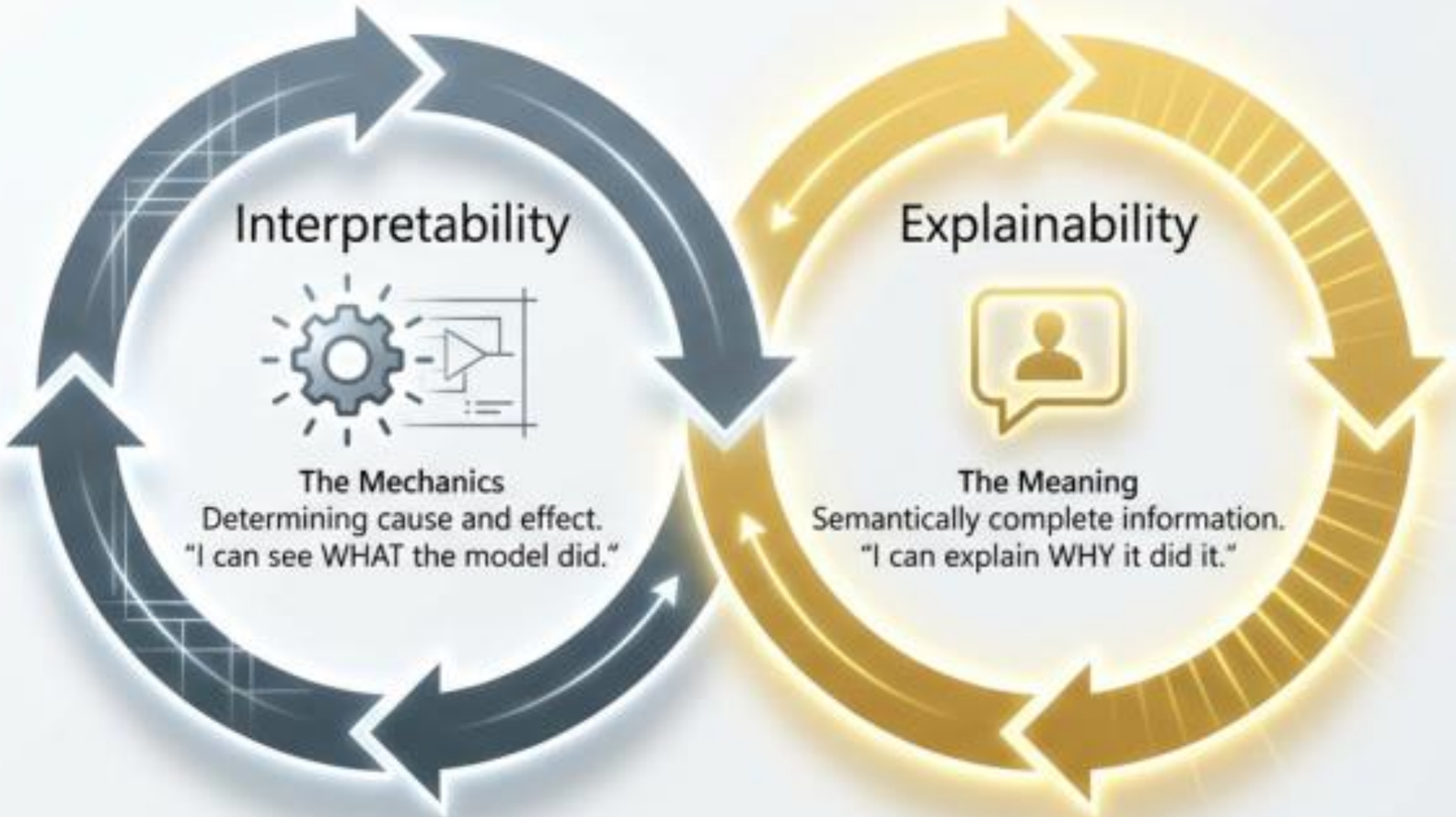
Explainability is no longer optional; it is the law

2017: Montreal Declaration.
A framework for Responsible AI.

2018: GDPR
Established the 'Right to Explanation' for automated decisions.

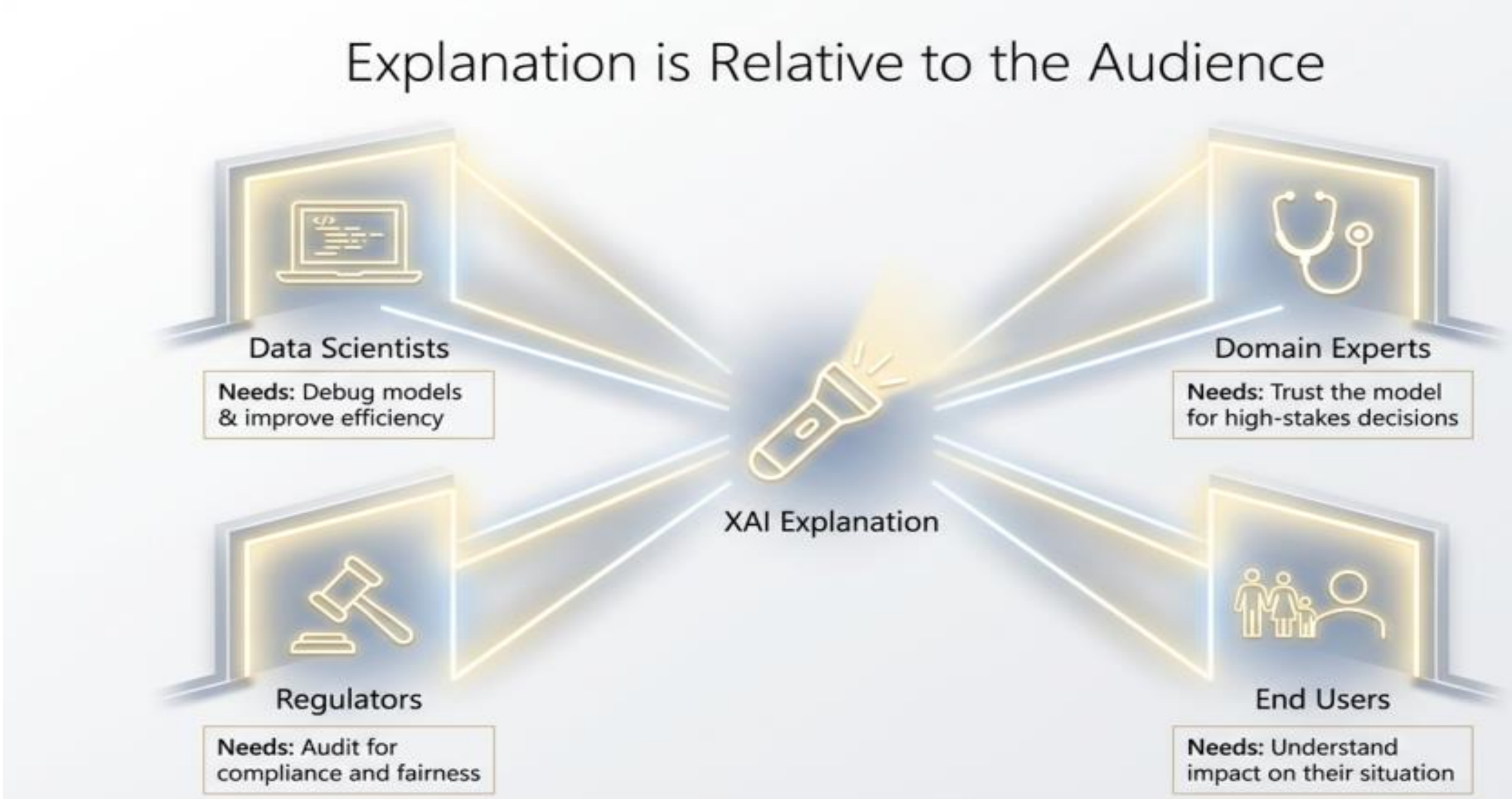
2020: DARPA XAI Program.
Investment in trusted systems.

Interpretability vs. Explainability



"Explainable models are interpretable by default, but the reverse is not always true."

Explanation is Relative to the Audience



The Accuracy vs. Interpretability Trade-off

The Taxonomy of Illumination

Ante-hoc

Glass Box

Transparent by Design
Interpretable models built from the ground up.
Examples: Linear Regression, Decision Trees.

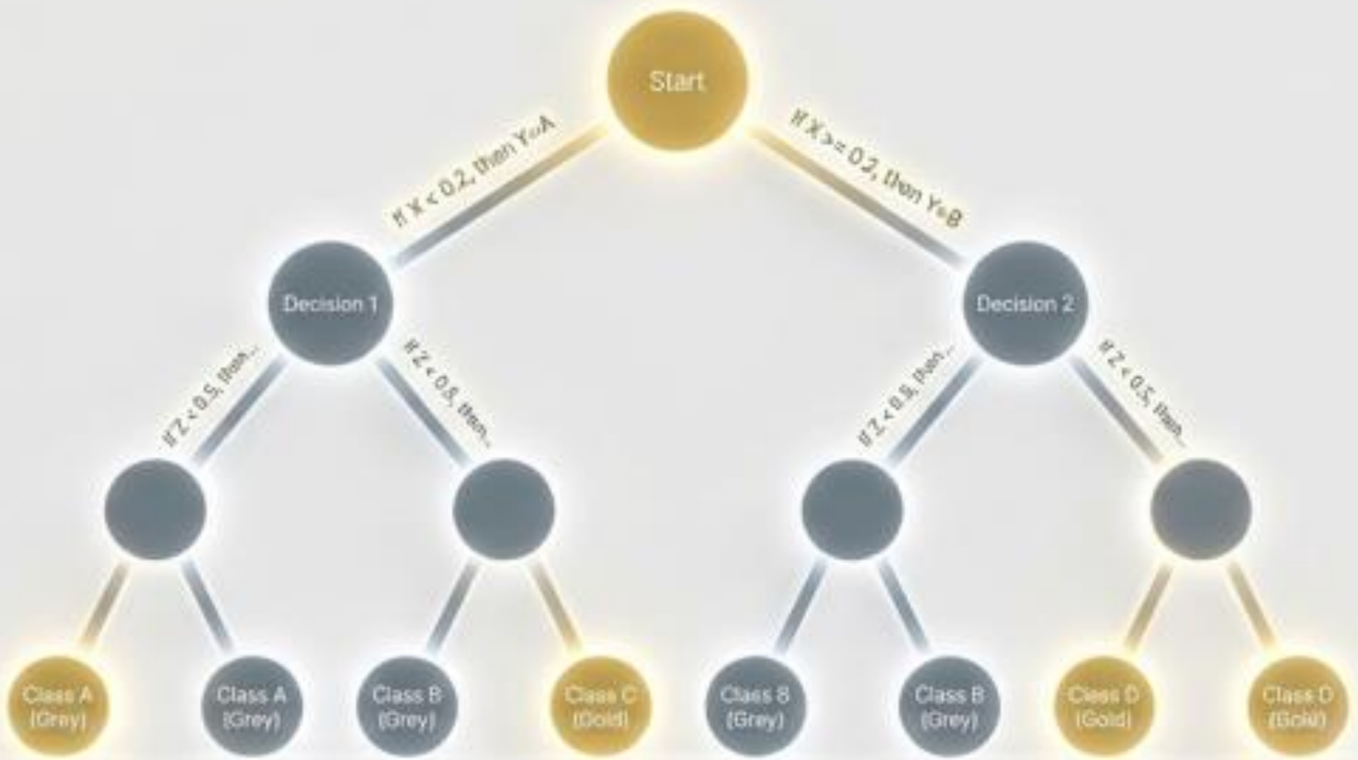
Post-hoc

Black Box

The X-Ray Approach
Techniques to explain opaque models AFTER training.
Examples: LIME, SHAP, Saliency Maps.

Ante-hoc: The Glass Box Approach

Frameworks for direct interaction between human knowledge and data. Transparent, but limited in handling complex, unstructured problems.



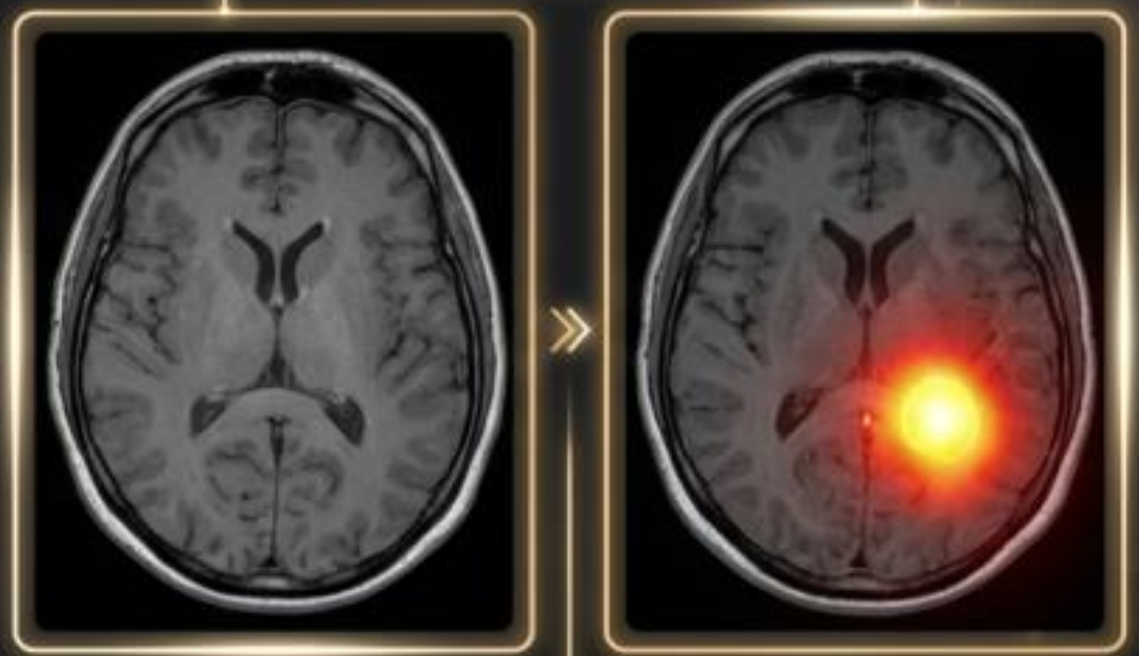
Visualizing the ‘Why’

Saliency Maps and Heatmaps provide intuitive visual evidence.

Correct Identification

Beagle (Correct)

Focus Attention



Tumor (Focus)

Challenges on the Horizon

1

Interdisciplinary Complexity.

XAI sits at the crossroads of informatics, psychology, and cognitive science.

2

The Consensus Problem.

No standard evaluation metrics yet. Explanations are difficult to quantify.

3

Psychological Hurdles.

Explaining a complex system to a human requires understanding human cognition, not just code.

Bridging the Gap Between Deep Learning Explainability and Privacy

PRIVACY

EXPLAINABILITY

The Regulatory Vise: HIPAA vs. GDPR

Balancing mandated secrecy with the right to explanation in clinical AI.

Conclusion: The Human in the Loop

Building trusted, inclusive, and ethical AI for tomorrow.

Today

Why did you do that?
When can I trust you?

Future with XAI

I understand why.
I know when to trust you.

Human Intuition + Machine Intelligence.

Pr. Imen JDEY
imen.jdey@ieee.org
FSEGS, Sfax
University, Tunisia.