Machine Learning and Deep Learning: Ethics and Challenges

Exploring the intersection of bias, data privacy, explainability, and the black box problem in modern AI systems

Group 6: Akhil, Riyaz, Deepika, Muhammad Gufran, Karthik Ram

The Questions We're Afraid to Answer

Trust & Control: The Black Box Dilemma

If our Professor just says if we passed or failed without any explanation, will that be ok?

Will you trust an AI doctor vs a human doctor? Now and in the future? When and how?

If you get a free Tesla Y, but you will have to use Autonomous Driving only, will you take it? Who will you blame if something wrong?

Value of Information: The Privacy Paradox

If granting AI access to everyone's **Apple Watch** health data could guarantee an Alzheimer's cure in 5 years, would withholding your data be immoral?

Why accept **Spotify** using our data for song recommendations, but resist insurers by not allowing to monitor your driving

Mirror to Society: The Bias Reality Check

Amazon scrapped its AI recruiting tool because it learned to discriminate against women. Did that fix the problem, or just hide the sexism back inside the human HR managers it learned from?

Apple Card gave women lower credit limits than their husbands despite shared assets. If the algorithm accurately reflected historical banking risk data, is the *math* sexist, or is reality?

Ethics and Challenges in Artificial Intelligence

Artificial Intelligence (AI) is transforming industries and daily life, but it also raises serious ethical challenges that must be addressed.

Key Ethical Issues:

- Bias and Fairness – when data or algorithms discriminate unfairly
- Data Privacy and Security – protecting sensitive personal information
- Explainability – ensuring AI decisions are understandable
- The 'Black Box' Problem – when AI becomes too complex to interpret

Goal: Balance innovation with ethical responsibility.

Bias and Fairness in Artificial Intelligence

AI learns from data — and biased data leads to biased outcomes.

Types of Bias:

- Historical Bias – reflects past inequalities in society
- Sampling Bias – underrepresentation of certain groups
- Label Bias – human subjectivity during data labeling

Bias and Fairness in Artificial Intelligence

Real-World Impact:

- Discrimination in hiring, credit scoring, and facial recognition

Solutions:

- Use diverse, representative datasets

- Conduct bias detection and fairness audits

- Maintain human oversight and ethical review

Data Privacy

- **Data privacy refers to protecting personal and sensitive information from unauthorized access or misuse.**

- **In AI, data privacy ensures that systems using large datasets do not expose user or confidential information.**

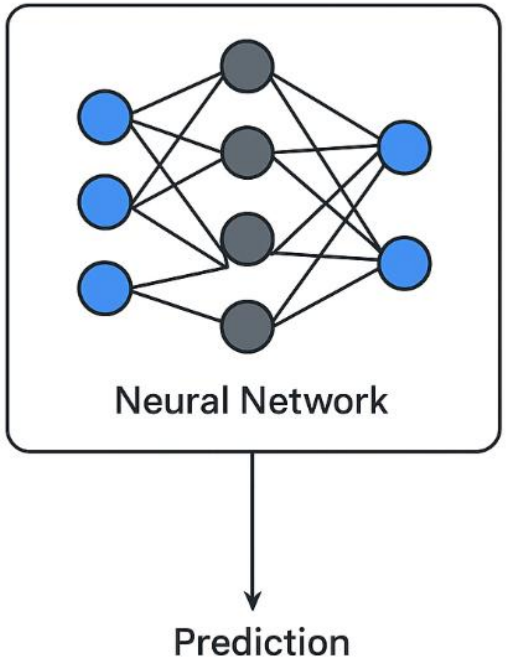
- **Protecting data builds trust and ensures ethical and legal compliance.**

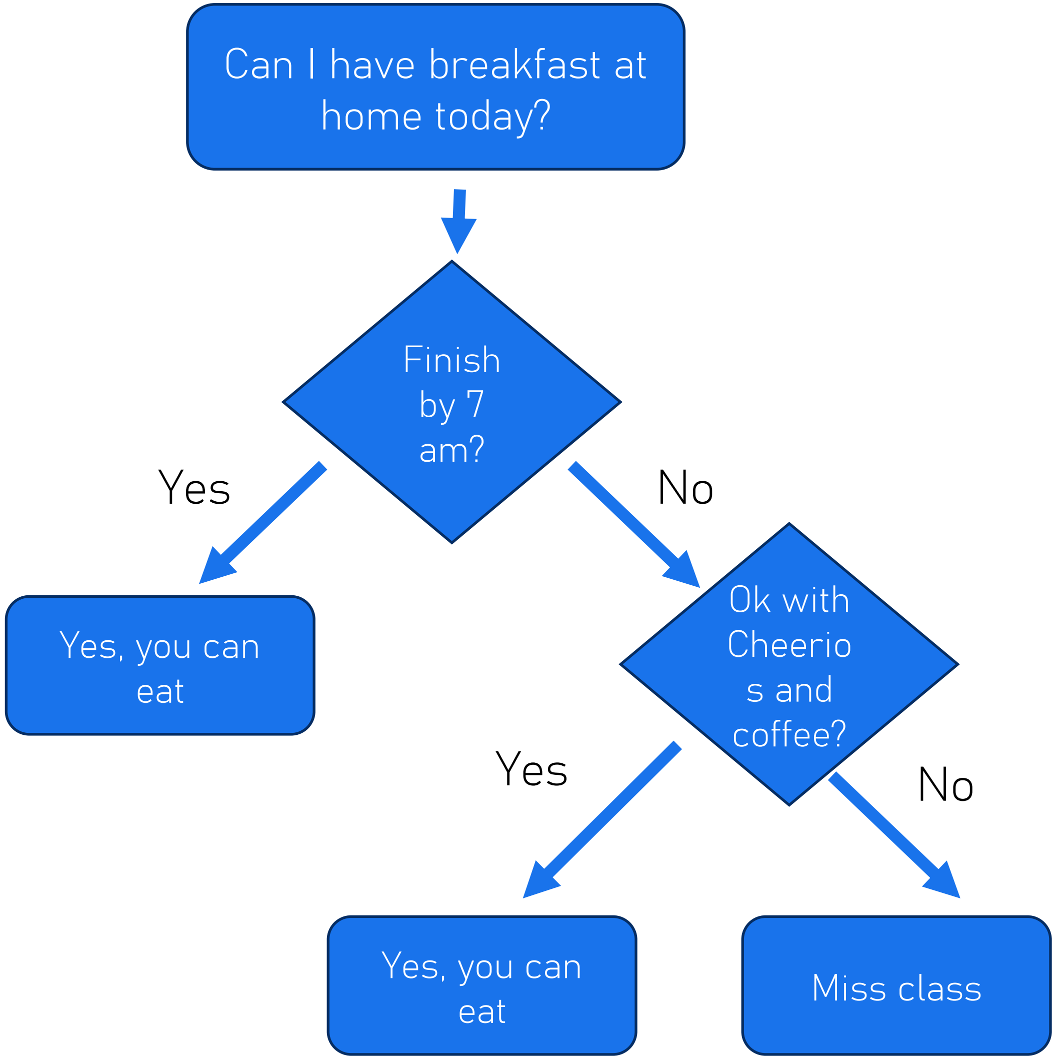
Data Privacy

- **Massive Data Collection – Ensuring user consent is difficult at scale.**
- **Data Leakage Risks – AI models may unintentionally expose private data.**
- **Bias and Misuse – Training data can contain sensitive or biased information.**
- **Third-Party Sharing – Integrations increase exposure and risk.**

Why is it "Black Box"?

What time should I wake up for my class today?
Inputs: All my historic behaviors, traffic patterns, driving history, time to get ready etc..
Output: One answer – 6:48 am





Understanding the Opacity

Why Neural Networks Are Opaque

Deep Learning models, like those powering ChatGPT or self-driving cars, rely on artificial neural networks. These can have billions of parameters—tiny conceptual "knobs" that the model tweaks during training.

A decision isn't made by one logical "if-then" statement. It's statement. It's made by the aggregate, non-linear interaction of interaction of these billions of parameters. It's like trying to trying to understand human consciousness by looking at at individual firing neurons; the complexity is too vast to easily to easily interpret.

Real-World Consequences

This isn't just an academic puzzle; it has real-world consequences. If an AI denies your mortgage application, or application, or worse, misdiagnoses a patient in a hospital, we hospital, we need to know why.

Without interpretability, we cannot fully trust the system. We system. We can't easily detect if the model is making decisions decisions based on sound logic or if it has learned dangerous dangerous biases hidden in the training data—for example, example, rejecting a resume based on gender rather than than qualifications.

The Path to Explainable AI

The industry is now heavily focused on "Explainable AI" (XAI). We are developing techniques that attempt to reverse-engineer these models, giving us a window into which features drove a specific decision. The goal is to move from "Black Boxes" to "Glass Boxes," ensuring that as AI becomes more autonomous, it remains accountable to human understanding.

Black Box Models

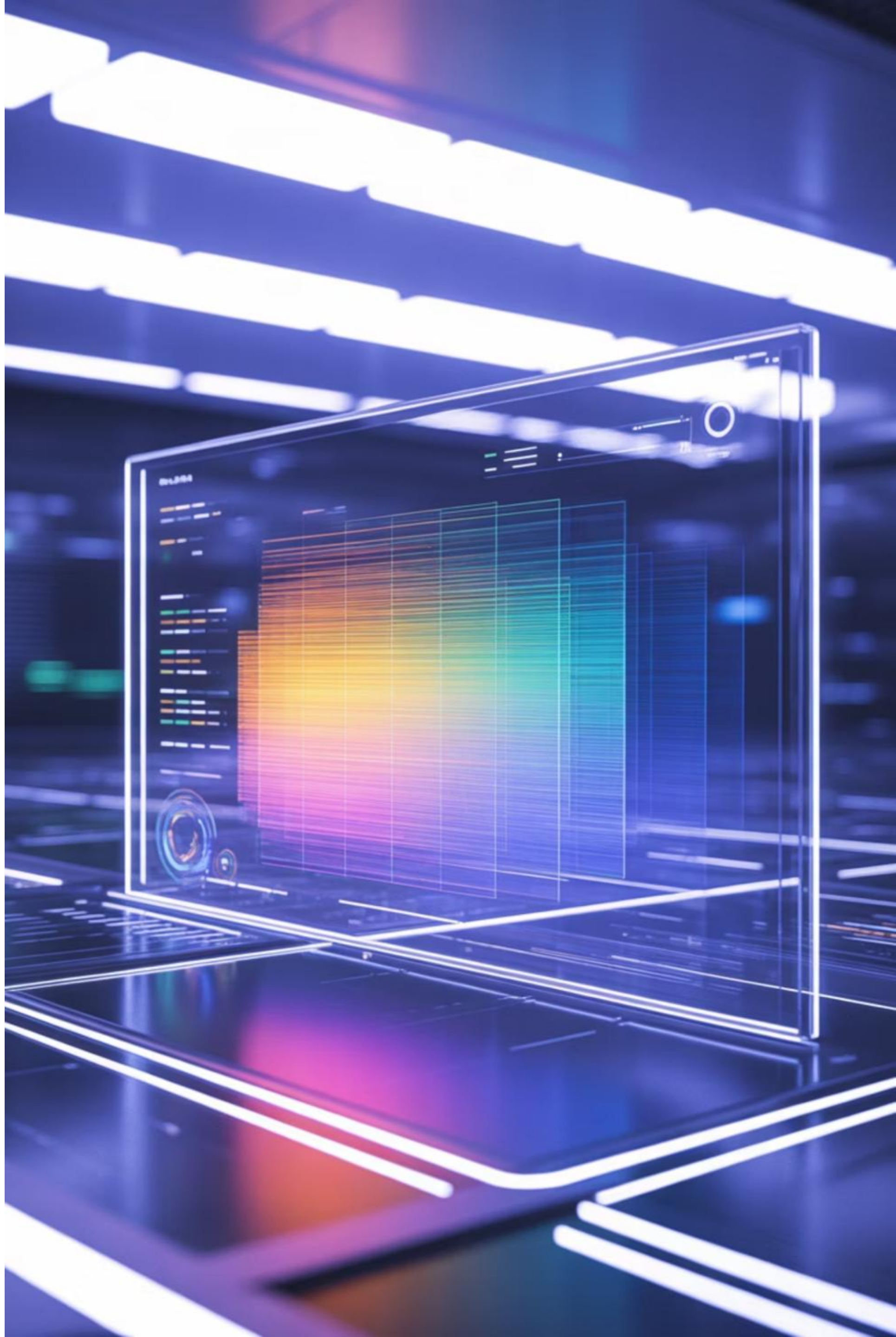
Opaque decision-making with billions of parameters

XAI Techniques

SHAP values and LIME reverse-engineer decisions

Glass Box Future

Transparent, accountable AI systems



The Imperfect Progress: Embracing Practical Utility

We've often accepted inherent risks when benefits are clear

1

The "Privacy" Trade-off We Already

We voluntarily share our exact location with apps like Google Maps for tangible benefits like avoiding traffic. This convenience greatly outweighs theoretical privacy concerns; we accepted this exchange years ago.

2

The "Black Box" We Trust With Our Money

Credit card fraud detection systems make instant, unexplainable decisions. We accept the slight inconvenience of false positives for the massive security benefit, not requiring detailed explanations from the AI.

3

"Bias" That is Actually Medical Accuracy

An AI identifying skin cancer might flag certain lesions more in lighter-skinned patients due to a higher melanoma base rate. In medicine, reflecting accurate biological risk factors isn't bias, but precision.

4

The "Unexplainable" Move That Was Superior

In 2016, Google DeepMind's AlphaGo's "Move 37" against Lee Sedol was initially baffling to human Go masters, yet it was a brilliant, game-winning move that redefined Go theory. We accepted its superiority without understanding its logic.

References & Resources

- **For the Amazon AI Recruiting Tool:** Dastin, J. (2018, October 10). *Amazon scraps secret AI recruiting tool that showed bias against women*. Reuters. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G
- **For the Apple Card Gender Bias:** Kelly, M. (2019, November 10). *Apple Card partner Goldman Sachs investigated over gender bias claims*. The Verge. https://www.google.com/search?q=https://www.theverge.com/2019/11/10/20958223/apple-card-goldman-sachs-gender-bias-investigation-new-york
- **For LIME (Local Interpretable Model-agnostic Explanations):** Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM. https://doi.org/10.1145/2939672.2939778
- **For SHAP (SHapley Additive exPlanations):** Lundberg, S. M., & Lee, S. (2True). A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30* (pp. 4765–4774). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf
- **For AlphaGo's "Move 37":** Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489. https://doi.org/10.1038/nature16961
- **For Bias in Medical AI (Skin Cancer Example):** Heaven, W. D. (2021, July 21). AI skin-cancer classifiers are less accurate on dark skin. *Nature*. https://www.google.com/search?q=https://doi.org/10.1038/d41586-021-01991-y *(Note: This is a "Nature News" article, providing context on the ethical issue of dataset diversity.)*
- LLMs like ChatGPT and Google's Gemini models were used for research