

Trustworthy AI Systems

Instructor: Guangjing Wang

guangjingwang@usf.edu

Instructor

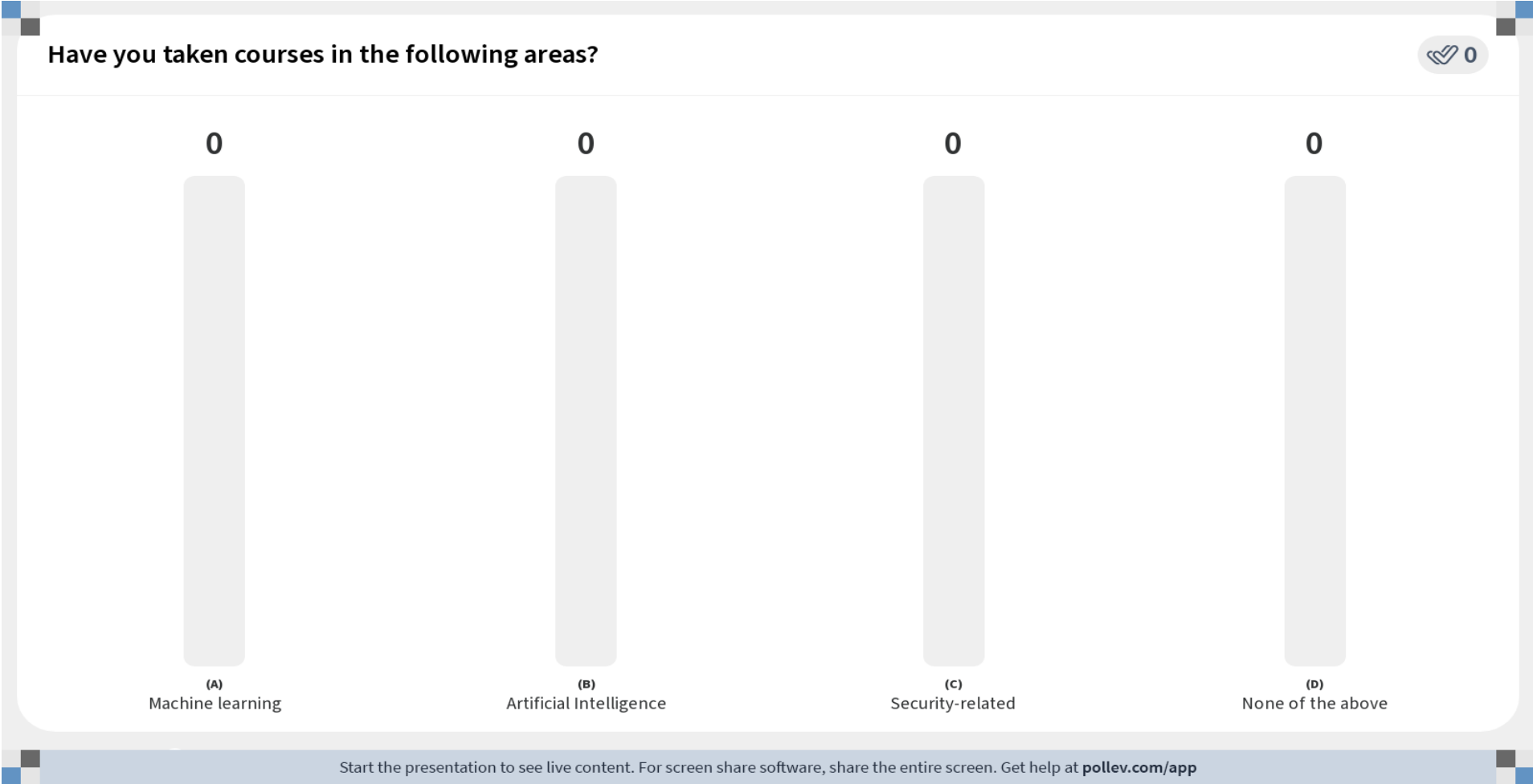
- Guangjing Wang
- Ph.D. degree from Michigan State University
- Ph.D. thesis topic: Applied AI for security and privacy in IoT
- Research interest:
 - LLM Agents: Exploring large language models for various applications.
 - Security and Privacy: Addressing challenges in safeguarding data and systems.
 - Sensing and Its Data Management: Leveraging various devices to collect, analyze and manage multimodal data from the physical world.

TAs



- Saandeeep Aathreya, Ph.D. Candidate
 - His research focuses on **affective computing**, which is a branch in Computer Science that deals with developing tools to perform human behavior analysis. The affective components in humans includes emotions, expressions, action units using various modalities.
- Md Imran Hossain, Ph.D. Candidate
 - His research focuses on developing and implementing **Explainable AI (XAI)** algorithms to enhance the interpretability of deep learning models, particularly for computer vision tasks such as image and video classification.

Your Background?



What is AI? (1)

- AI: behaving like an Intelligent being, planning, reasoning, human computer interaction
- ML: a subset of AI to find patterns from a large scale of data

What is AI? (2)

From a technical perspective:

- Machine Learning (deep learning, statistical learning, etc.)
- Natural Language Processing, Computer Vision
- Data Mining, Multiagent Systems, Knowledge Representation
- Information Retrieval, Human-in-the-loop AI, Search, Planning, Reasoning, Robotics and Perception

AI Algorithm and AI System

AI Algorithm

- Data representation
- Algorithm accuracy

AI system

- Data: data drift, concept drift
- Algorithm: generalization
- Computer System: efficiency, scalability, etc.
- User, Society: trustworthiness

The AI system is not the algorithm itself, it is about how the algorithm is implemented, situated within the human context.

What is Trustworthy AI? (1)

What is trust?

- Trust in AI is earned from a person or community
- Continuing demonstration of robustness and reliability
- Trustworthiness is for particular audiences, must have the target

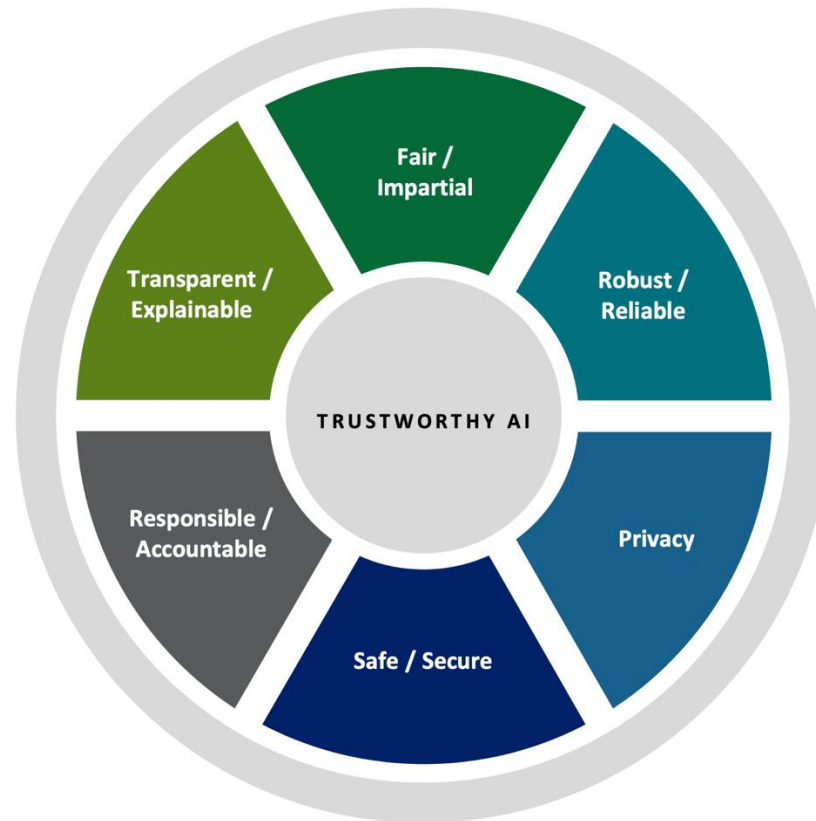
What is Trustworthy AI? (2)



Note: there is no single answer or standard, as trustworthiness depends.

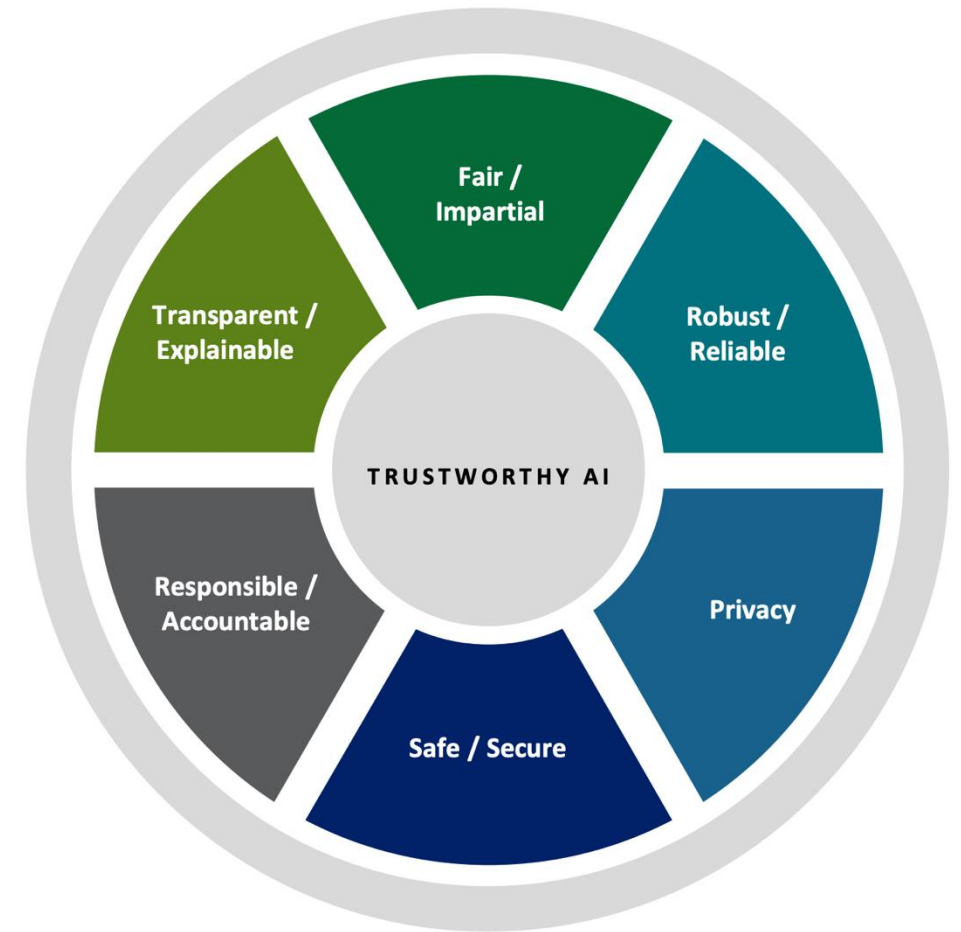
Trustworthy AI principles (1)

What is your understanding?



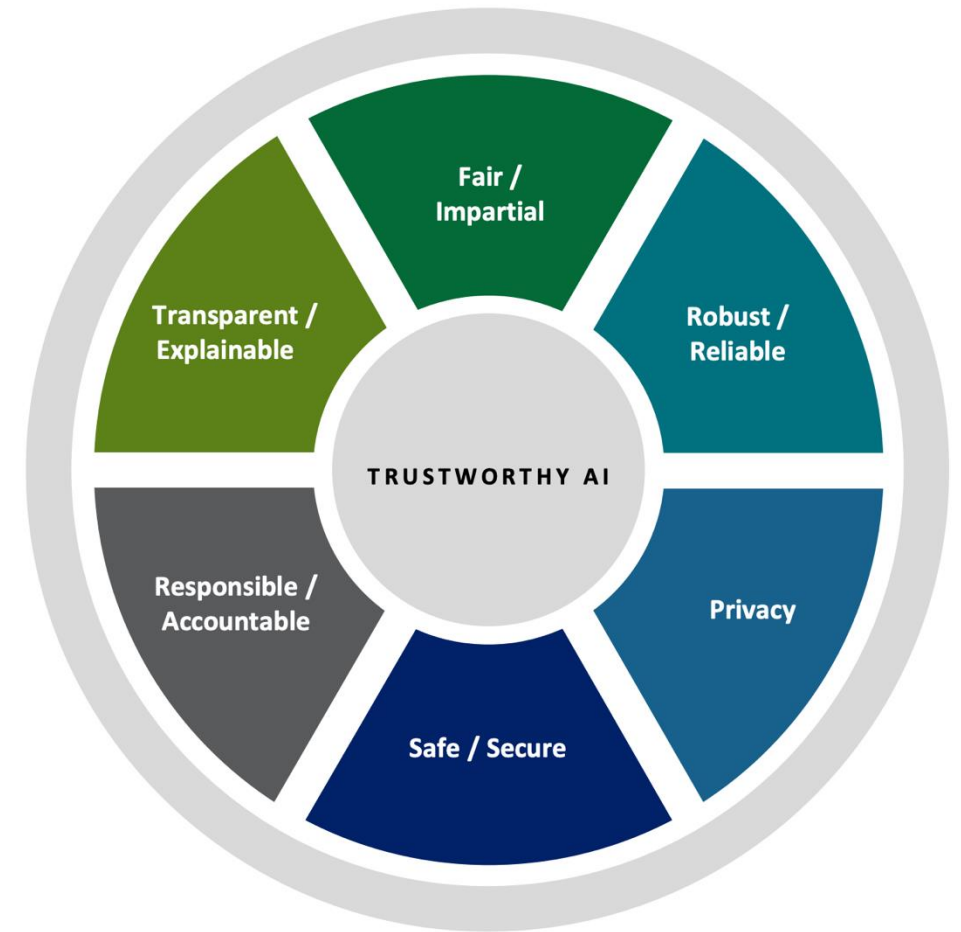
Trustworthy AI principles (2)

- Safety: avoid risks that cause physical/digital harm to any individual, group and entity
- Privacy: data should not be used beyond its intended usage



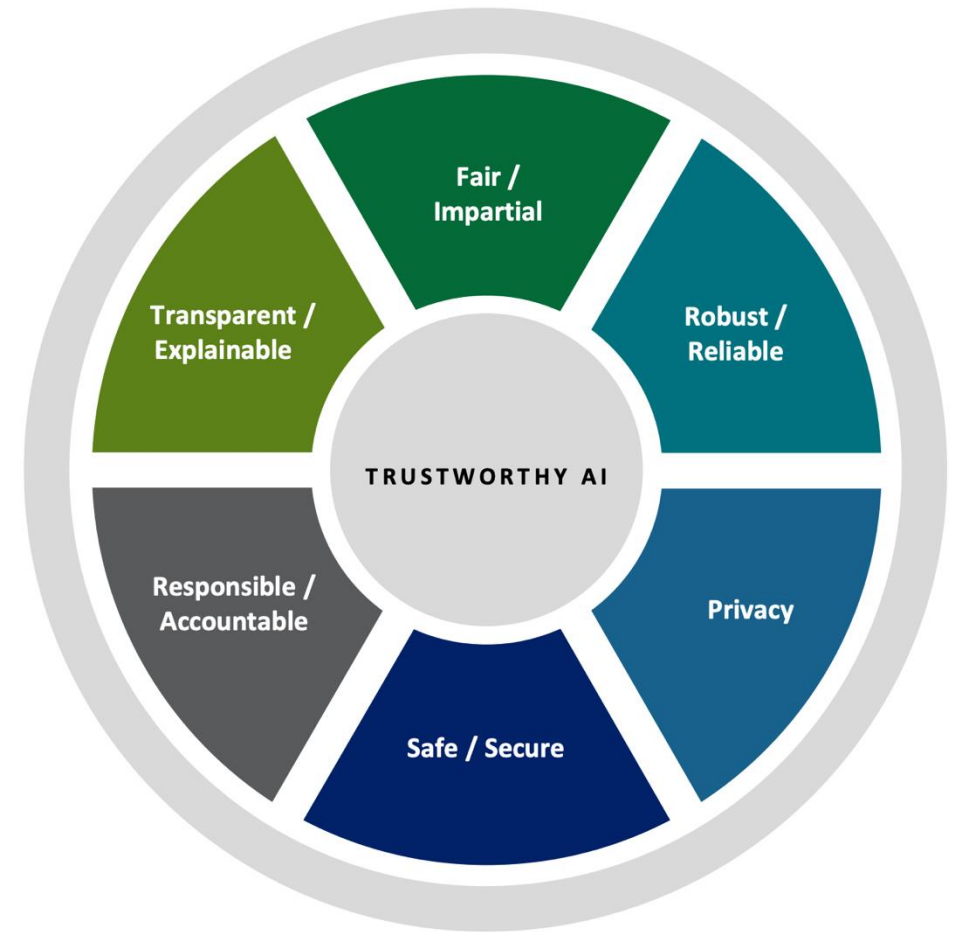
Trustworthy AI principles (3)

- Robustness: accurate and reliable outputs that are consistent with the original design
- Fairness: equal application to all applicants



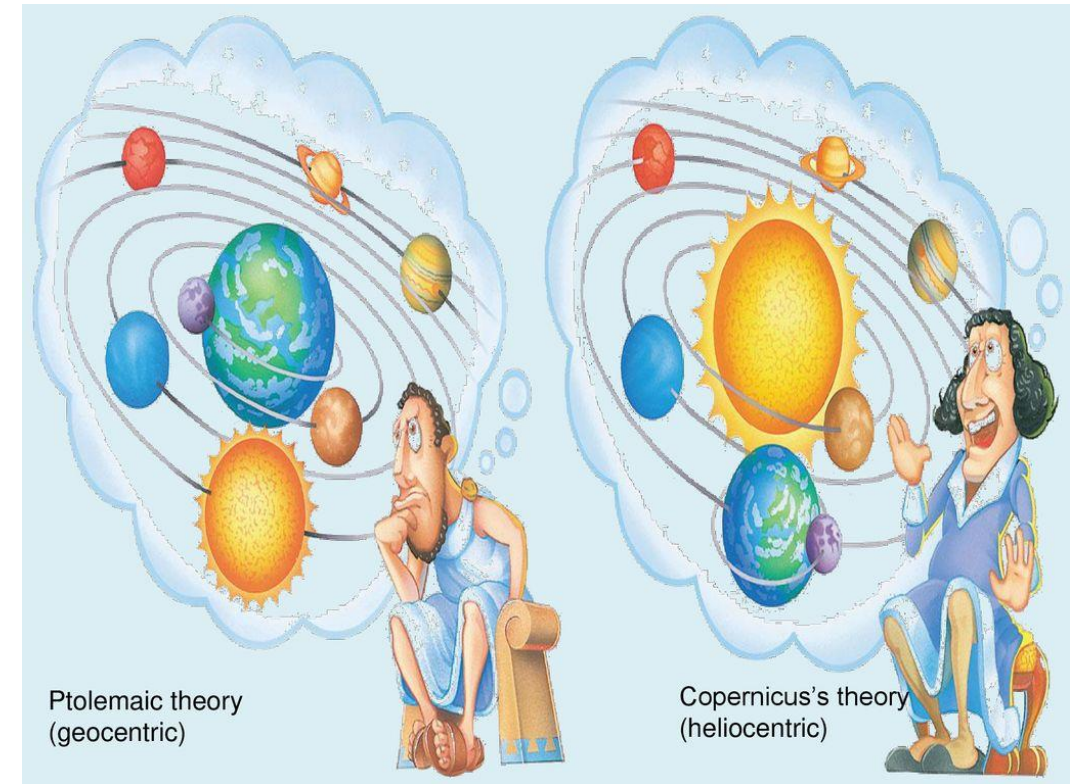
Trustworthy AI principles (4)

- Transparency: algorithm, policy of data, data sharing, and usage
- Accountability: outline governance and who is responsible for all aspects of AI solutions



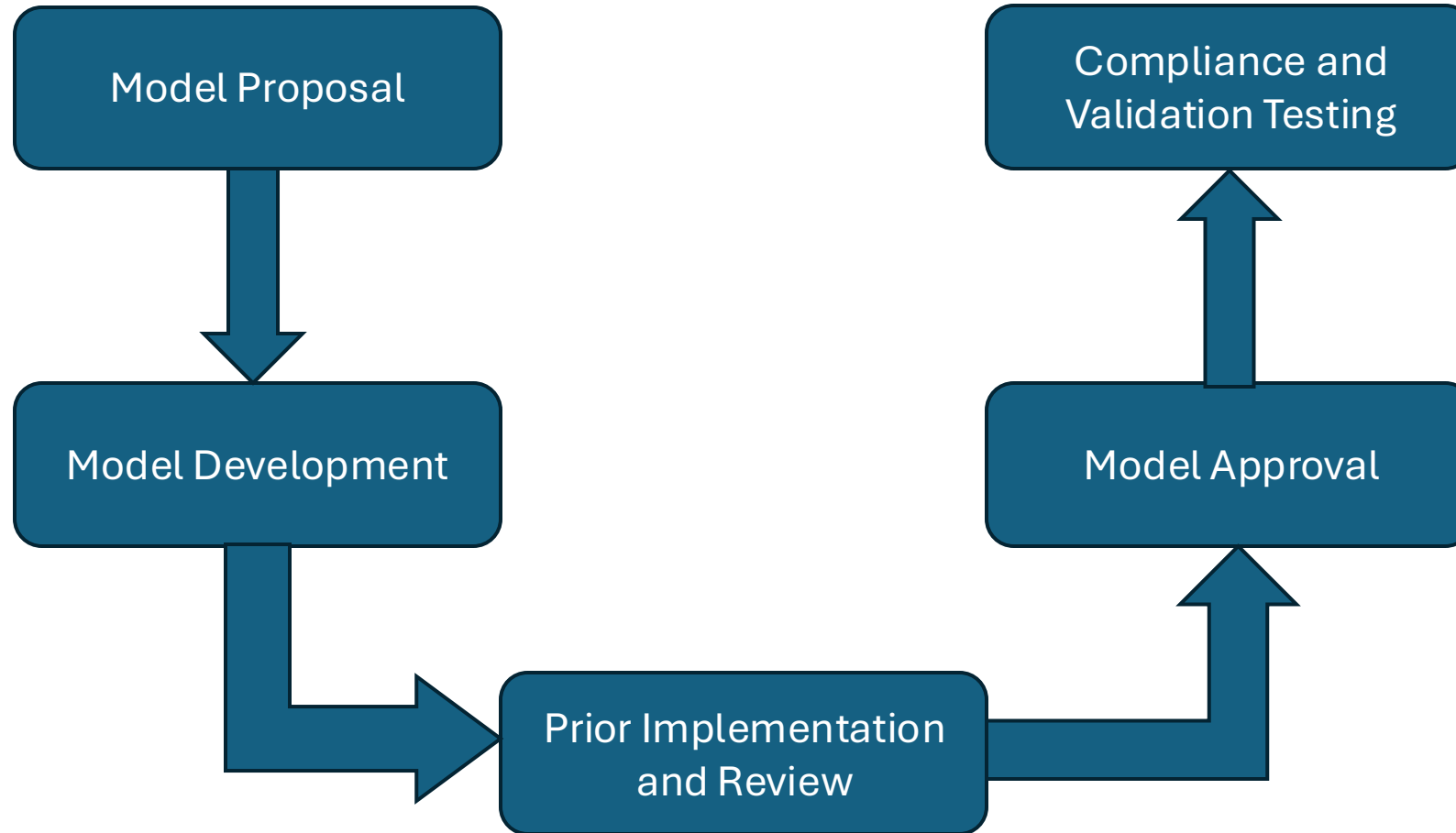
Be Critical!

- The existing theory of AI could be incomplete
- Algorithm explainability: can be misleading
- Something is explainable does not mean that the explanation is correct



<https://slideplayer.com/slide/16121923/>

Achieving Trustworthy AI System



What will we learn this semester?

- Take a break
- First day attendance
- Syllabus review
- Questions?

References

- <https://www.youtube.com/watch?v=0EW3uUCCoUc>
- <https://www.youtube.com/watch?v=V7kWAZ-dV0w>
- <https://www.hhs.gov/sites/default/files/hhs-trustworthy-ai-playbook.pdf>