Trustworthy AI Systems

Instructor: Guangjing Wang

guangjingwang@usf.edu

Instructor

Guangjing Wang

- https://guangjing.wang/
- guangjingwang@usf.edu
- Office @ BEH 311

My Requests

- When you send me an email: **Include "[CAI 6605]" in the subject**.
- When you visit my office, if the door is closed, please **knock on the door** first.
- Just call me Dr. Wang

CAI 6605 Course

- CAI 6605 is developed based on CIS6930, a selected topics course: a very high-level graduate course, research-oriented.

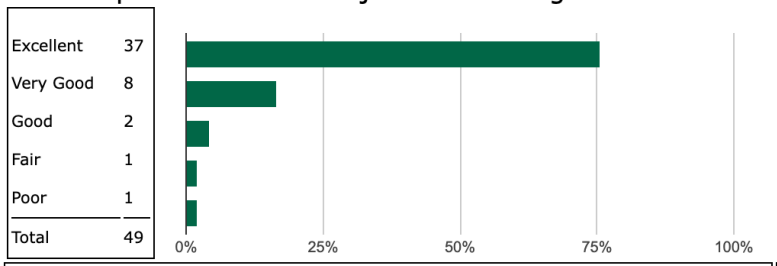
- If you want to learn some basics, choose courses such as machine learning, deep learning, Introduction to AI…

- This course
  - CIS6930 Fall 2024 (65 enrolled)
  - CIS6930 Spring 2025 (72 enrolled)
  - CAI6605 Fall 2025

Last Course Evaluation

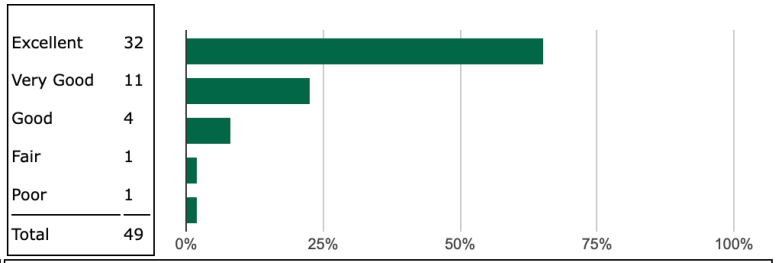
- https://fair.usf.edu/EvaluationMart/EvaluationComment.aspx?recid=306026&reportid=60805&reporttype=D

1. Description of Course Objectives & Assignments



Statistics	Value
Mean	4.6
Median	5.0
Mode	5
Standard Deviation	+/-0.8

2. Communication of Ideas and Information



Statistics	Value
Mean	4.5
Median	5.0
Mode	5
Standard Deviation	+/-0.9

Rubrics

Project-Midterm (Code)	12%
Project-Final (Code)	12%
Project-Checkpoints	6%
Essay	20%
Two quizzes	20%
Midterm Project Presentation	12%
Final Project Presentation	12%
Peer Feedback in Written	6%

TA and Course Time

TA:

- Mahit Venkat Gautam Kumpatla
- Email: mahitvenkatgautam@usf.edu
- TA's office hour: appointment by email

Course Time:

- M, W: 6:30pm
- Instructor's office hour: Tuesday, 2-5 pm
- Previous slides: https://guangjing.wang/CIS6930/lectures/

Tips: Be Active to seek for feedback

- TAs will evaluate your midterm projects and final projects
- Talk and discuss more to let them understand your efforts
- Please ask TAs to give you feedback in advance
- I will give detailed feedback and suggestions during your midterm and final presentations, group by group in person.

```
I think the course needs more TA support
```

```
The material is engaging, and the discussions encourage critical thinking about important
topics in the field. While some areas, such as assignment feedback, could be slightly
improved, the course overall offers a solid learning experience.
```

More tips: Try to sit in the front

```
He can speak a little more loud and clear
```

I will try my best to speak louder and clearer…

Another tip: **Be a graduate instead of an undergraduate**

- Do not think you can understand everything by just listening to lectures.
- You are expected to read papers in more detail by yourself before and after class.
- Lectures are for guidance in this course.

Difference between Graduates and Undergraduates?

- Undergraduate:
  - I give you the problem, I give you the solution, you implement it
- Master:
  - I give you the problem, you find the solution, you implement it
- Ph.D.:
  - You find the problem, you give the solution, you implement it

The goal in the course

- You are confident to write your course project on your resume and introduce it to your interviewer/recruiter.
    - Be confident in your contribution
    - Be familiar with every detail of your solution
    - Good presentation to introduce the problem
    - Solid understanding of the related work and challenges

Syllabus

- Check the syllabus for more details

- First-day attendance assignment
  - Deadline: Aug.26th 08:59 AM
  - Fail to finish will be automatically dropped

- Questions AND take a break

What is Artificial Intelligence? (1)

- Artificial: made or produced by human beings rather than occurring naturally, especially as a copy of something natural.

- Artificial Intelligence: behaving like an intelligent being, planning, reasoning, human-computer interaction

- ML, pattern recognition, data mining: 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 a target

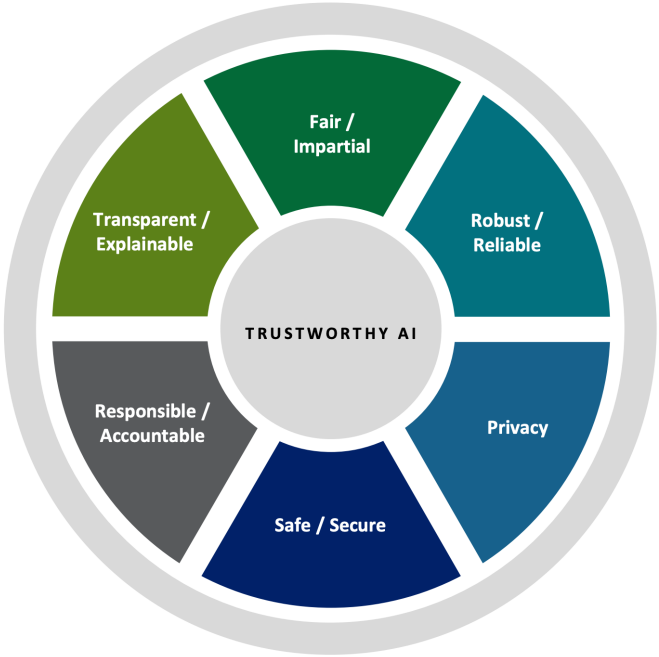
What is Trustworthy AI? (2)

https://www.youtube.com/watch?v=V7kWAZ-dV0w

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

Trustworthy AI principles (1)

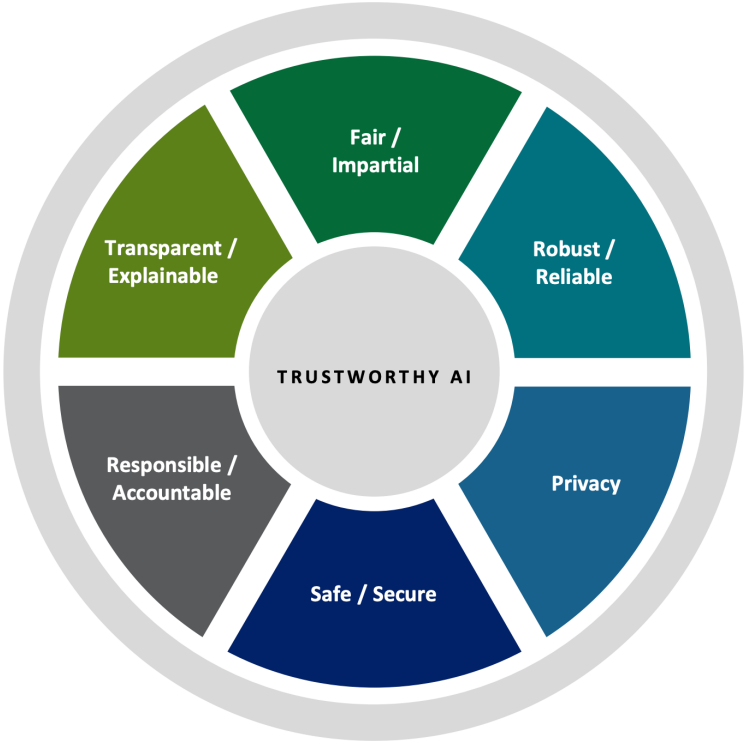
What is your understanding?



Trustworthy AI principles (2)

- Security: avoid risks that cause physical/digital harm to any individual, group and entity

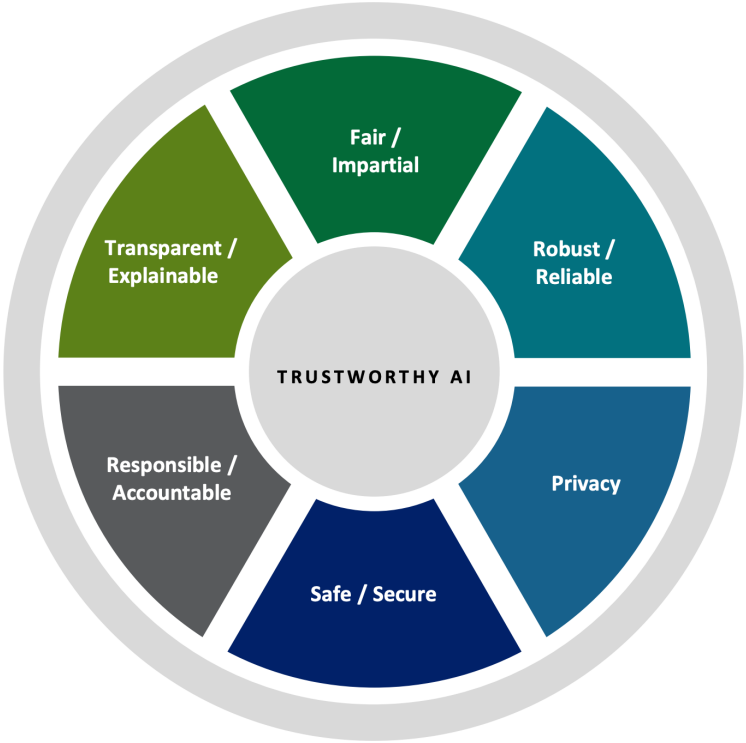
- Privacy: data should not be used beyond its intended usage



Trustworthy AI principles (3)

- Robustness: accurate and reliable outputs that are consistent with the original design

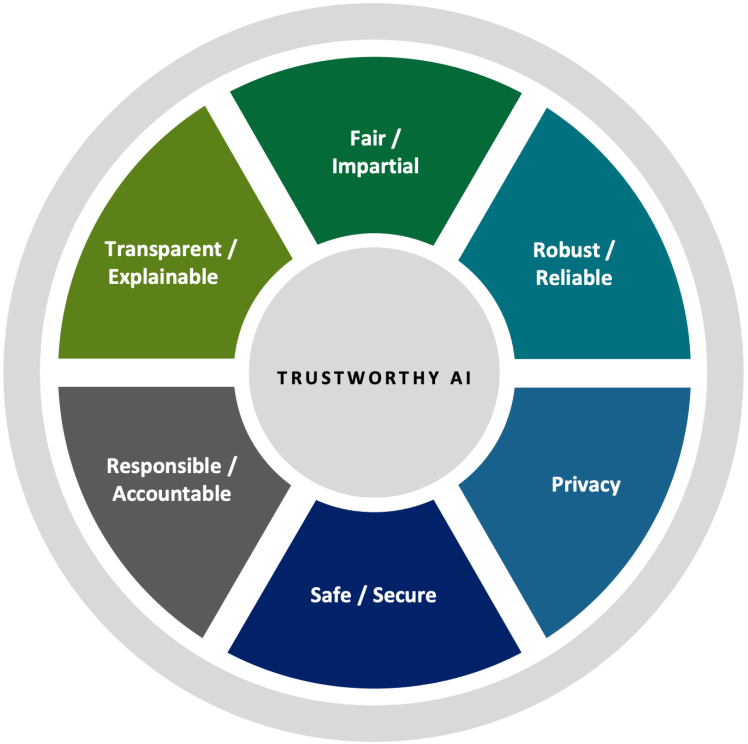
- Fairness: equal application to all applicants



Trustworthy AI principles (4)

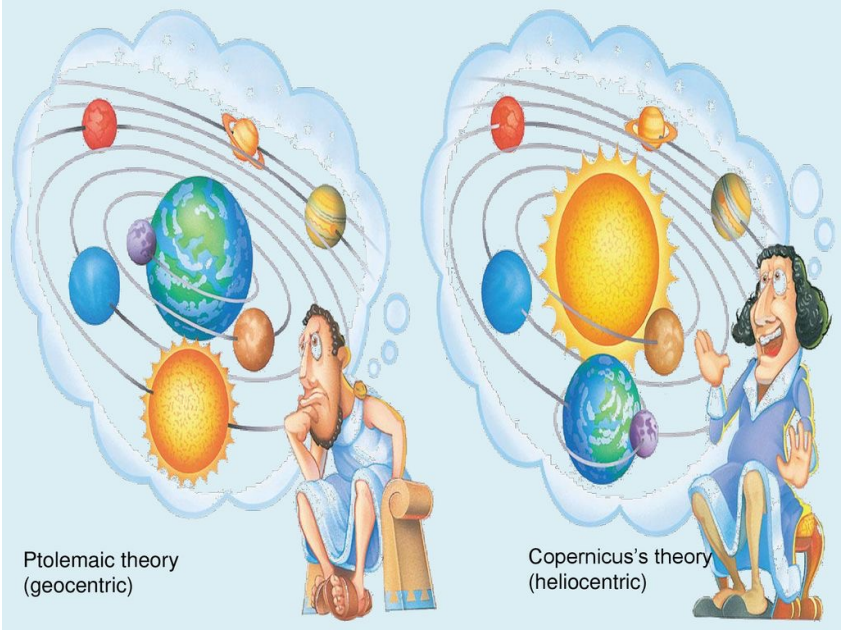
- Explainability: algorithm, policy of data, data sharing, and usage

- Accountability: outline governance and who is responsible for all aspects of AI solutions



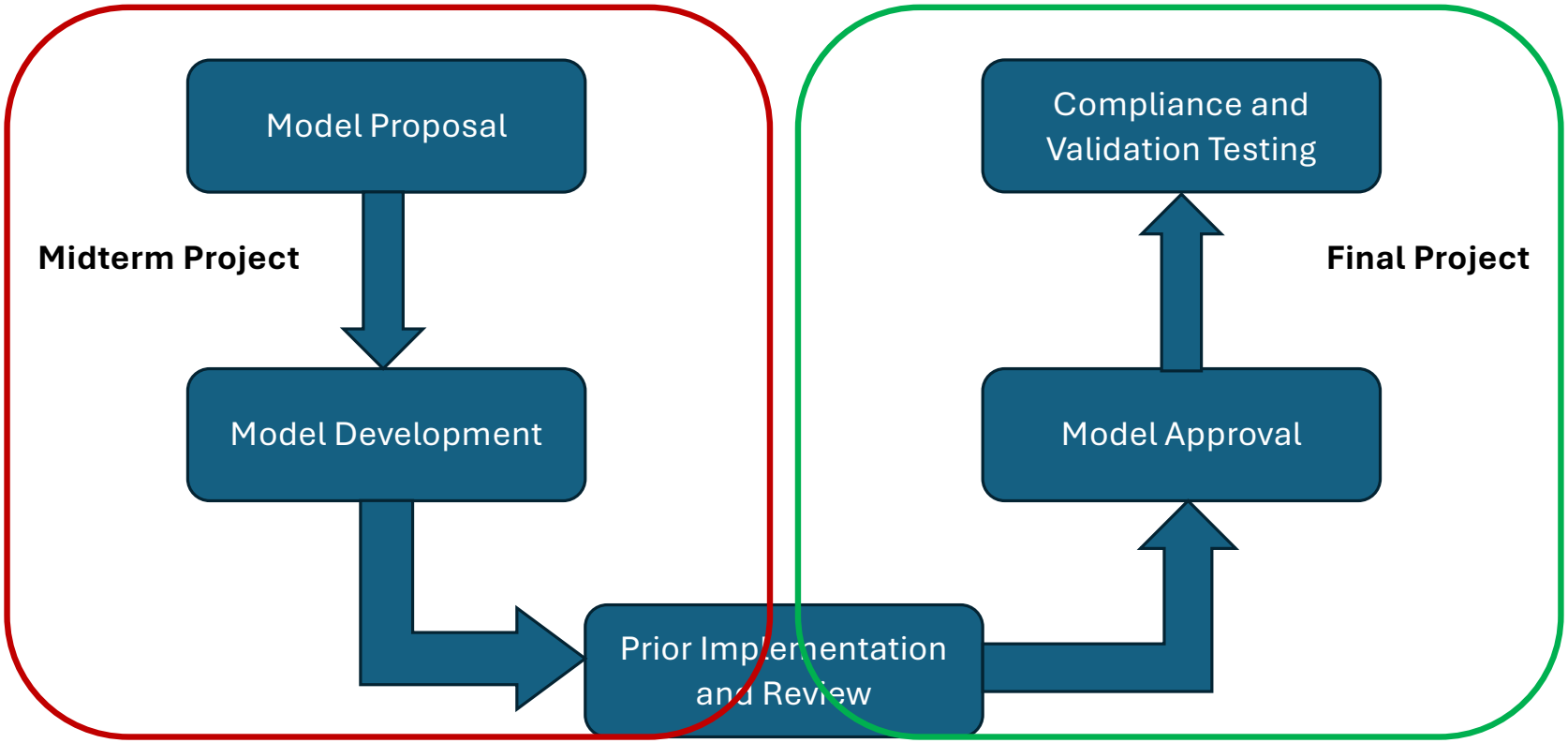
Be Critical!

- The existing theory of AI could be incomplete
- E.g., 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



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