Trustworthy AI Systems

Instructor: Guangjing Wang

guangjingwang@usf.edu

Instructor

- Guangjing Wang, Ph.D. --- Dr. Wang
  - https://guangjing.wang/
  - guangjingwang@usf.edu
  - Office @ BEH 311

- When you send me an email: **Include "[CIS6930 Trustworthy AI]" in the subject**.
- When you visit my office, if the door is closed, please **knock on the door** first.

CIS6930 Courses

- CIS6930 is 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
  - CAI6108 Fall 2025 (under plan)

Last Course Evaluation

Tampa - Engineering - Computer Science & Engineering		Instructor : Wang, Guangjing	Course Term : Fall 2024
Course Title : Trustworthy AI Systems		Course ID : CIS - 6930 - 004 / CRN : 97083	
Number Enrolled : 65	Number Responded : 62	Percent Responded : 95.00	

Comments Report

ITEM ID	ITEM	Excellent		Very Good		Good		Fair		Poor		No Response		Mean
		No.	%	No.	%	No.	%	No.	%	No.	%	No.	%	
E1	Description of Course Objectives & Assignments	36	58.06	21	33.87	5	8.06	0	0.00	0	0.00	0	0.00	4.50
E2	Communication of Ideas and Information	32	51.61	21	33.87	6	9.68	1	1.61	0	0.00	2	3.23	4.40
E3	Expression of Expectations for Performance	35	56.45	16	25.81	8	12.90	1	1.61	0	0.00	2	3.23	4.42
E4	Availability to Assist Students In or Out of Class	37	59.68	16	25.81	7	11.29	0	0.00	0	0.00	2	3.23	4.50
E5	Respect and Concern for the Students	39	62.90	17	27.42	4	6.45	0	0.00	0	0.00	2	3.23	4.58
E6	Stimulation of Interest in the Course	33	53.23	21	33.87	5	8.06	1	1.61	0	0.00	2	3.23	4.43
E7	Facilitation of Learning	32	51.61	21	33.87	7	11.29	0	0.00	0	0.00	2	3.23	4.42
E8	Overall Rating of the Instructor	32	51.61	23	37.10	4	6.45	1	1.61	0	0.00	2	3.23	4.43

https://fair.usf.edu/EvaluationMart/EvaluationsReport.aspx?reportid=39760&reporttype=D

There are Some Changes

Project-Midterm (Code)	12%
Project-Final (Code)	12%
Project-Checkpoints	6%
Essay	20%
Two quizzes	20%
Midterm Project Presentation	15%
Final Project Presentation	15%

Why Changes

Two quizzes are added…

```
The project based style of this course is good, however, I would have preferred if you had some in-class exams or quizzes in addition. It would have really cemented my understanding. But the course content and teaching style is fabulous.
```

Randomly recording attendance for **Extra Point**

```
Professor Guangjing Wang made the course interesting and overall engaging. The course was setup well, but the lectures were a bit lacking, and he was unable to keep full engagement of the class sometimes.
```

TA and Course Time

- TA
  - Aastha Sharma
  - Fahim Rahman
- Course Time
  - You are there, you know it.
  - We need to follow the schedule from the department.

```
My only problem of this course is the schedule timing. Why does it start from 6:30pm? It is dinner time and people feel sleepy. It is a really good course that deserves a decent schedule before 5pm.
```

Tips: Be Active to Seek for Feedback

- TA will evaluate your midterm projects and final projects
- Talk and discuss more to let them understand your efforts
- Ask TAs to give you more feedback…
- 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 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 ultimate goal in the course

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

Syllabus

- Check the syllabus for more details

- First-day attendance assignment
  - Deadline: Jan.14th 09:59 AM
  - Fail to finish will be automatically dropped

- Take a break

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

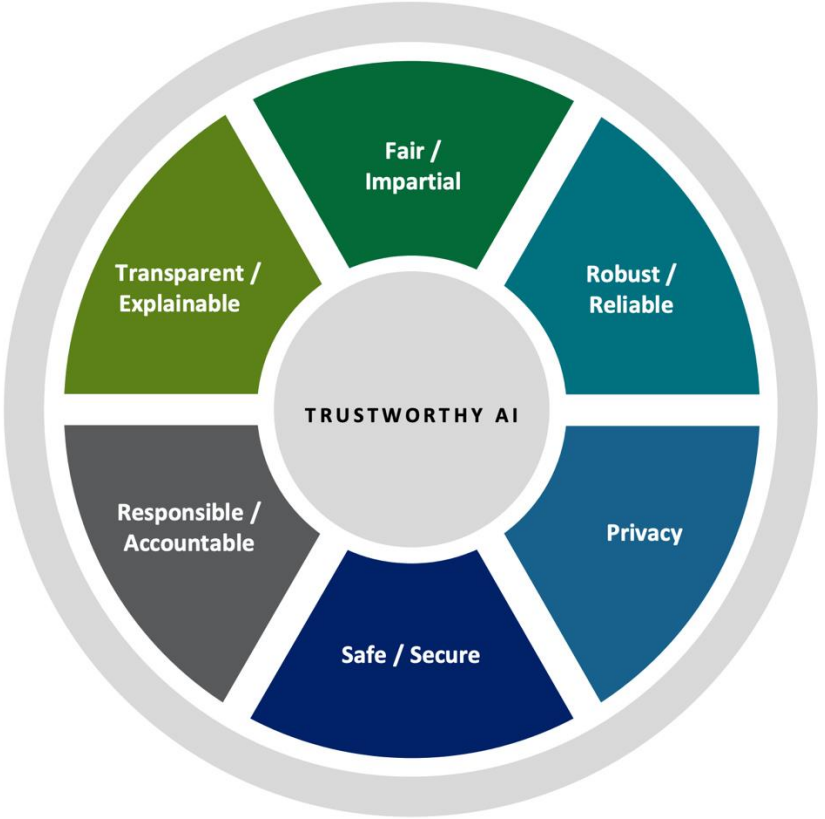
What is Trustworthy AI? (2)

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

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

Trustworthy AI principles (1)

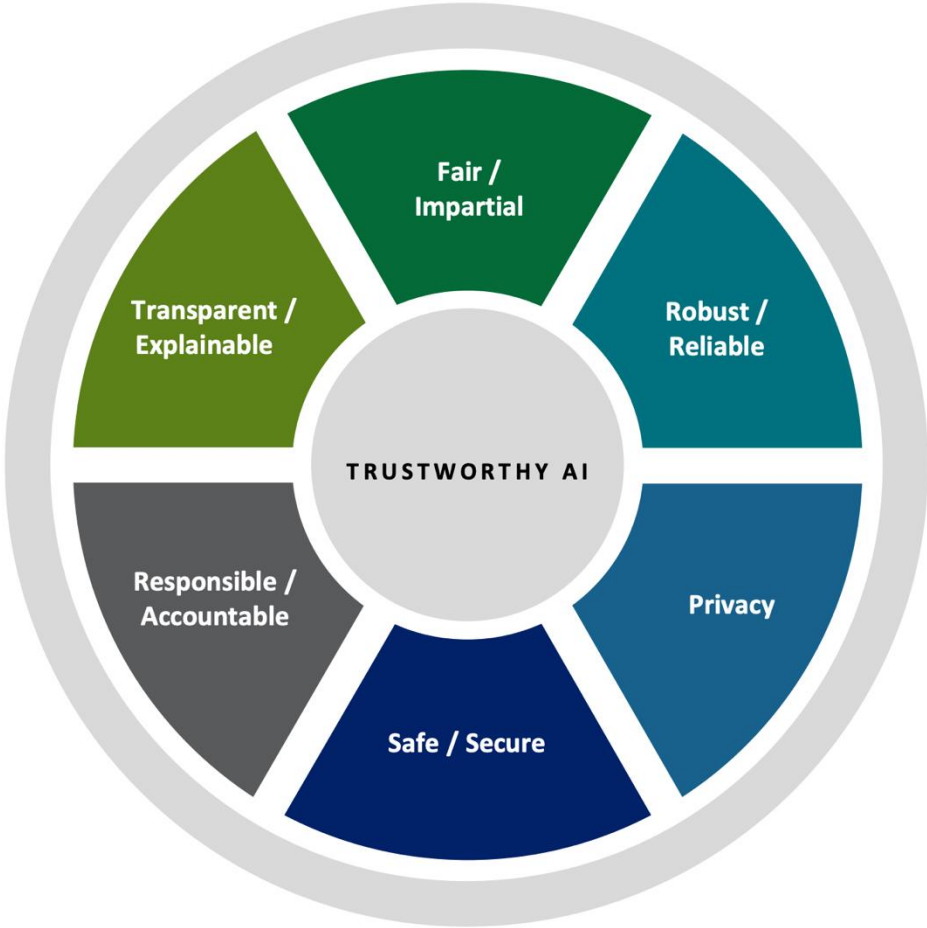
What is your understanding?



Trustworthy AI principles (2)

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

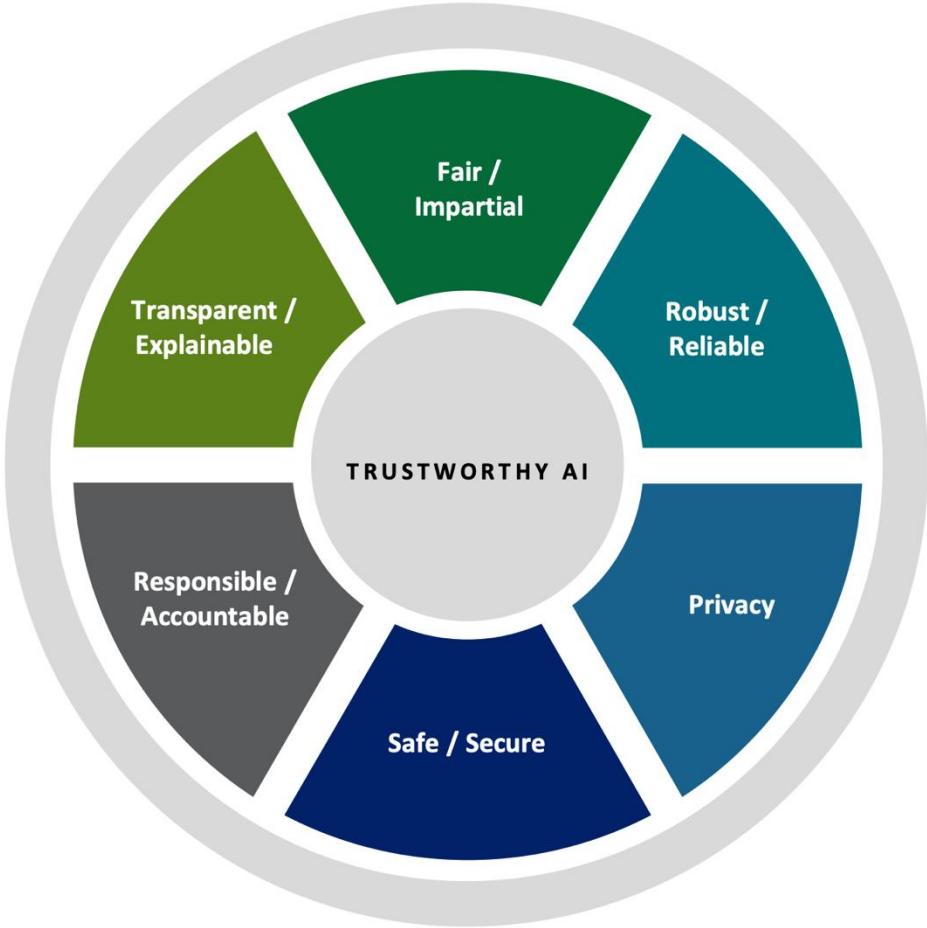
- Privacy: data should not be used beyond its intended usage



Trustworthy AI principles (3)

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

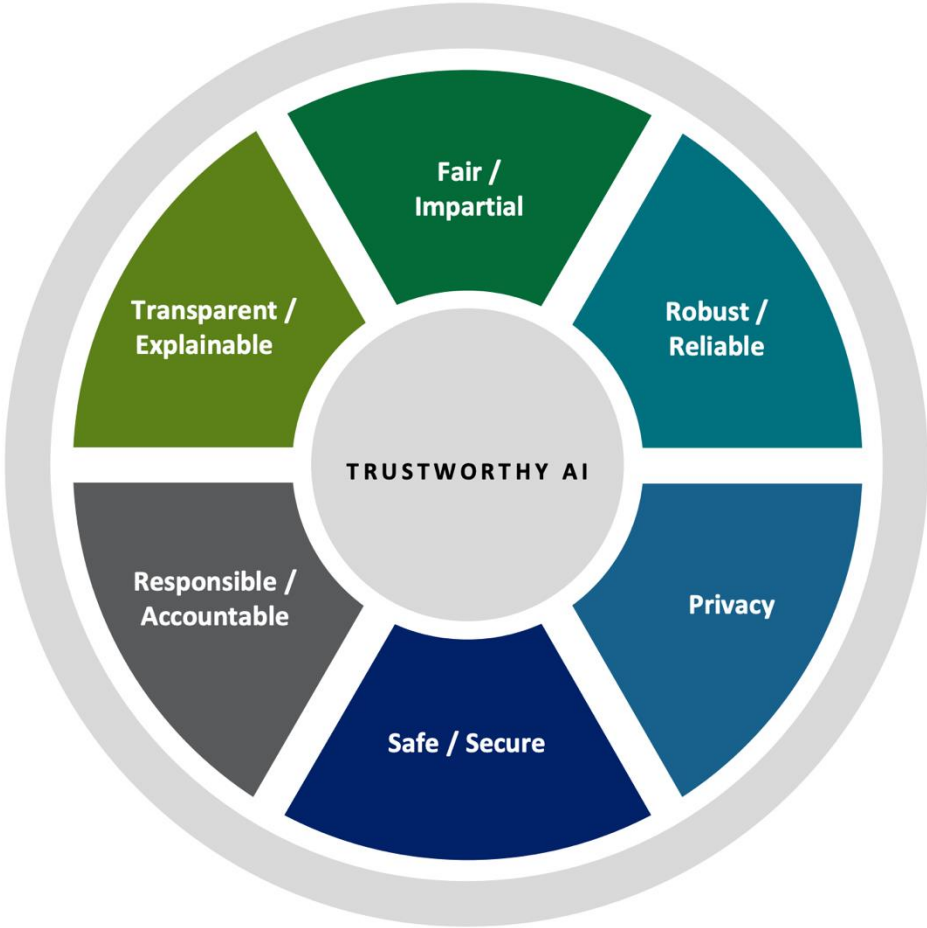
- Fairness: equal application to all applicants



Trustworthy AI principles (4)

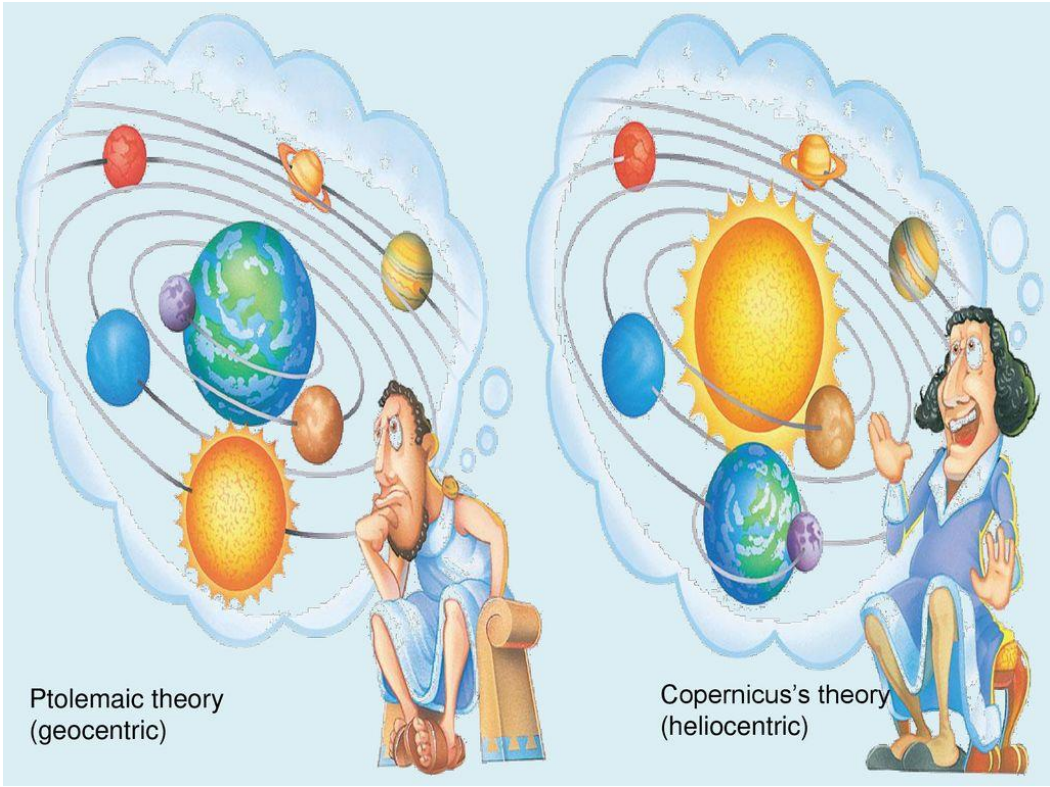
- Explainability: algorithm, policy of data, data sharing, and usage

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



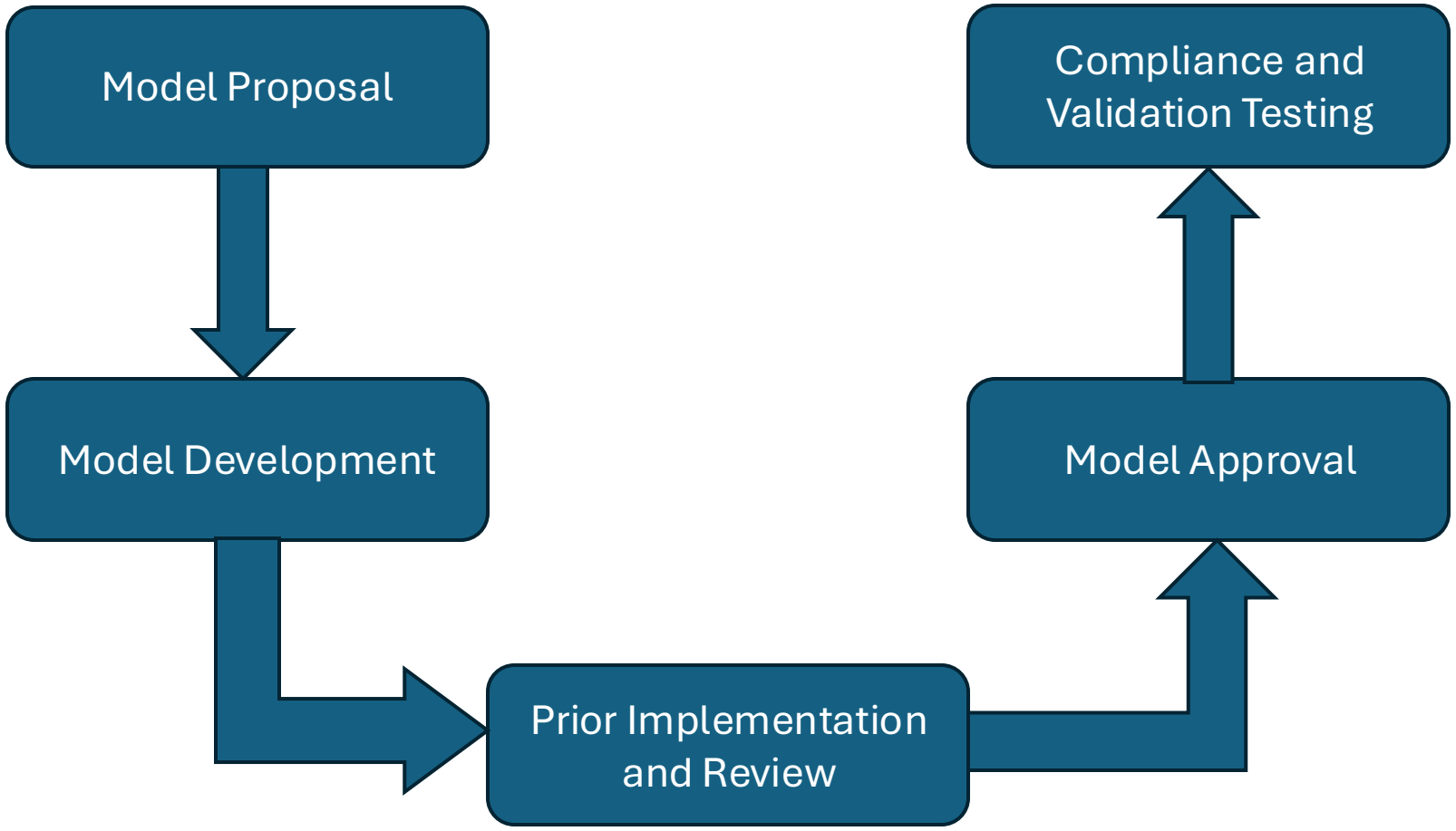
Be Critical!

- 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