Trustworthy AI Final Project: Report Option

In the Trustworthy AI domain, we often say "trust is construct", which means "Let's start from the idea that trust is something we humans make up and agree on—it's not something that just appears on its own." Over the course of this semester, we've talked about different ways that people are trying to understand AI systems, and understand what it would mean to trust AI systems, and perhaps understand how to build systems that can be trusted. These discussions remain incomplete, and what "trust" is as a construct related to AI is both unclear and changing as the technology (and our relationship to it) changes.

So, this assignment focuses on understanding what is a construct and why there is so much confusion about what things mean. For Part I, we ask that you identify and digest 10 readings that center on trust (or trust by another name) and for each reading chart the move from fuzzy concept to operationalization.

We ask that you take note of the various ways trust is named and list those out in Part II. This will help us understand what we each mean when we say "trust or "trustworthiness" in specific contexts. Not only are words often used imprecisely, the meanings attached to them can vary widely from discipline to discipline.

For Part III, we ask you to construct a schema that represents the construct that is most relevant to your problem domain or area of interest. In other words, "locate" trust as a relationship among things and/or actors.

Part I – From Fuzzy concept to operationalization

Constructs or fuzzy concepts are key to theorization – they are the "what" of a theory – but they are not directly measurable without first having conceptualization. A conceptualization pins down a fuzzy concept into something precise that can then be measured or counted (operationalization). One example is listed and the related reading is in the DTAIS Summer Incubator 2025 folder.

Paper/Measure (author/s, date)	Fuzzy Concept	Conceptualization	Operationalization
example (topic is complexity)	Complexity is the thing that makes software hard to test and maintain	Graph-theoretic measure of paths through a program	$C_{MCC} = E - N + 2 * P$ (1)
1.			
2.			
3.			
4.			
5.			
6.			
7.			
8.			
9.			
10.			

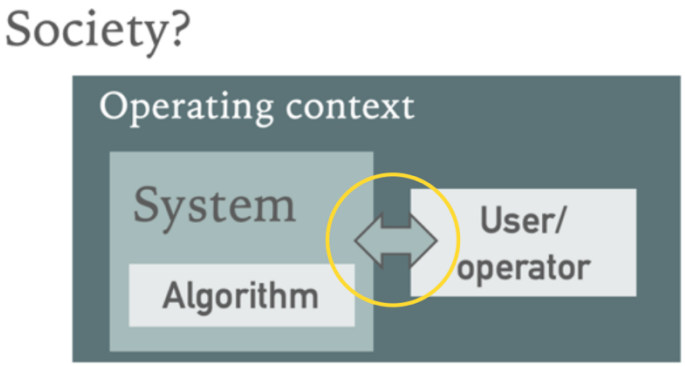
Once you have listed each for your readings, take a moment to reflect:

- Looking across the reading you have selected, do they agree with each other about what they are trying to measure?
- Is there agreement (across the measures) about *should* be measured?
- Are the operationalizations doing a good job at measuring that?

Part II – Constructs of trust
List constructs (trust by other names) you have come across in your reading or interactions.

Part III – Create your own version of a trust schema and identify where is trust located?
A trust schema is a visualization of the contexts and relationships where trust (or trust by another name) matters to your area of interest.

In the example below, the arrow indicates that trust is located or is critical between the user/operator and the system containing the algorithm. This is where your attention will be focused. Use shapes and arrows of your choosing, consider adding labels or a brief explanation to make your schema legible. These



As I've been talking to some students about this potential assignment, I've found that people with a mostly CS background struggle to understand what is expected in these schema. To help, I'm sharing below a version of a "trust schema" in the world of Frozen, which is basically like AI—built with good intentions but prone to accidentally freezing everyone out[1].

> In this extended analogy, we focus on a schema for *interpersonal trust in Disney's Frozen*.
>
> **Step 1: Identify Key Actors and Relationships**
>
> Here's a simplified schema that shows how trust might operate in **Frozen**:
>
> - **Anna → Elsa**: Trust is based on sibling bonds but is strained by fear and secrecy.
> - **Elsa → Anna**: Trust wavers due to Elsa's desire to protect Anna by distancing herself.
> - **Anna → Hans**: Initially high trust (romantic excitement), then betrayal (false identity).
> - **Anna → Kristoff**: Starts low, grows as actions prove reliability and care.
> - **Olaf → Everyone**: Unconditional trust, symbolizes innocence and optimism.
>
> **Step 2: Represent the Construct as a Schema**

[1] Just suggested by ChatGPT

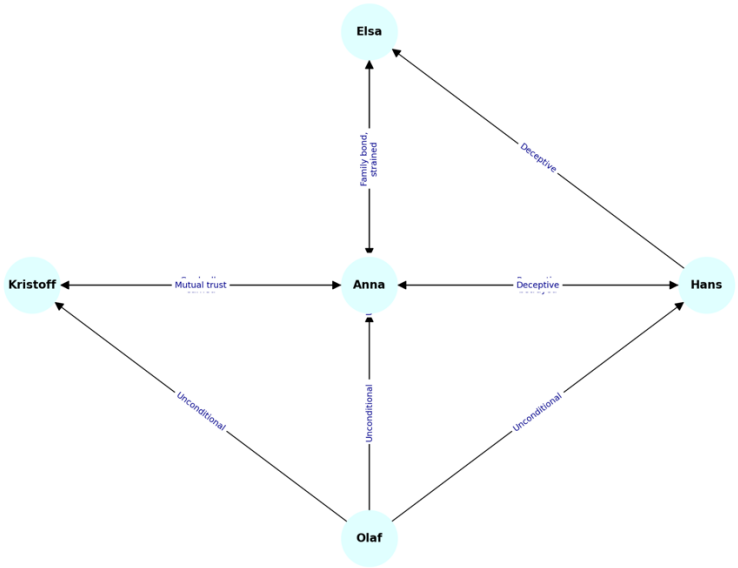
You could draw a diagram showing:

- **Characters as nodes**
- **Arrows labeled with trust level and basis** (e.g., “earned through shared adventure,” “broken by betrayal,” “based on family bond”)
- Optional: Color-code arrows (e.g., green for trust, red for broken trust, gray for uncertain trust)

Step 4: Reflect on the Construct

In this world:

- Trust isn’t just a feeling; it’s **based on roles** (sibling, stranger, hero), **history of interactions**, and **expectations of behavior**.
- Trust can change quickly with new information.
- What counts as trustworthy varies by character and context (e.g., Elsa isolates herself to be trustworthy, but it feels untrustworthy to Anna).



Please note that this is a somewhat strained analogy. I don’t expect snowmen to be key stakeholders in your trustworthy AI schema, instead you might have end-users, data-creators, data-owners, AI companies, governments, domain practitioners, etc…