A Framework for Thinking About Trust in Education and AI

Ruben R. Puentedura, Ph.D.

1. Why Look at Issues of Trust?

Many conversations about AI and education these days focus on issues of *trust*, i.e. questions about specific worries regarding the trustworthiness of specific uses of AI, in specific education contexts. Errol St. Clair Smith has done an excellent job of collecting the most salient areas of concern and summarizing them in an **Areas of Trust Concern × AI Typologies** matrix. The matrix identifies nine areas of primary trust concern and connects them with three different AI types: Gen AI (creating content), Agentic AI (taking actions autonomously), and Ambient AI (pervasively present in the background). The result is a rich and valuable resource, providing a snapshot of where related conversations are happening today. At the same time, this poses a challenge: the matrix lists 27 items with examples for each one, and it's easy to either lose focus or over/underestimate the relevance of many of these items.

It would therefore be desirable to have a simpler alternative approach that groups the matrix items into a smaller set of categories, allowing for reasonable productive discussion, but also highlighting patterns that might otherwise remain opaque. I decided to look for a suitable set of categories, taking a set of cues from approaches associated with the work of Pierre Bourdieu (e.g. Multiple Correspondence Analysis (MCA) and related techniques – see the recent article by Will Atkinson for an excellent review/critique of these approaches). Unfortunately, at this point in time the dataset for AI and trust in education is not large enough to support a full formal statistical analysis. But the exploration did suggest a categorization that is, I believe, strong enough to help with conversations about trust in educational uses of AI.

2. The Four Trust Categories

Four trust-related categories account reasonably well for 24 out of the 27 items in the original matrix. Each can be seen as a different **object of trust**: a reframing from *"what we worry about"* to *"what is actually being trusted."*

Category I — Trust in the *Output*

- **Construct:** accuracy, sourcing, verifiability, and authenticity of what the AI produces.
- **Core question:** *"Can I believe and legitimately claim as my own what comes out?"*
- **Related concepts:** information quality, epistemic trust, authorship/authenticity.

Category II — Trust in the *Process*

- **Construct:** whether consequential decisions are explainable, contestable, and attributable, and whether harm can be remedied.
- **Core question:** *"Can I see how it decided, challenge it, and find who is responsible?"*
- **Related concepts:** procedural justice, algorithmic accountability, explainability, due process.

Category III — Trust in the *Data Relationship*

- **Construct:** capture and use of personal and behavioral data, monitoring, consent, covert inference, and data-driven sorting and discipline.
- **Core question:** *"Who is watching, capturing, and sorting me, and did I meaningfully agree?"*
- **Related concepts:** surveillance/dataveillance, informational power asymmetry, contextual integrity, consent.

Category IV — Trust in the *Human Role*

- **Construct:** erosion of human judgment, skill, relationships, and labor market position as AI assumes cognitive and decision work, in the classroom now, and in the economy later.
- **Core question:** *"What human capability and role is being displaced, and is that displacement chosen or drifting?"*
- **Related concepts:** deskilling, automation and autonomy, human in the loop, future of work.

3. Trust Categories and the Original Matrix

Each of the 24 intersections accounted for in the original matrix is mapped in the table below to a primary category (●) and, where relevant, a secondary one (○).

Category	Matrix Intersection	C-I	C-II	C-III	C-IV
I Trust in the *Output*	Safety × Gen AI	●	·	·	·
	Fairness × Gen AI	●	○	·	·
	Academic integrity × Gen AI	●	·	·	○
	Transparency × Gen AI	●	○	·	·
II Trust in the *Process*	Fairness × Agentic AI	·	●	○	·
	Transparency × Agentic AI	·	●	·	·
	Accountability × Gen AI	○	●	·	·
	Accountability × Agentic AI	·	●	·	·
	Accountability × Ambient AI	·	●	○	·
III Trust in the *Data Relationship*	Safety × Ambient AI	·	○	●	·
	Fairness × Ambient AI	·	○	●	·
	Privacy × Gen AI	·	·	●	·
	Privacy × Agentic AI	·	·	●	·
	Privacy × Ambient AI	·	·	●	·
	Academic integrity × Ambient AI	·	·	●	○
	Transparency × Ambient AI	·	○	●	·
IV Trust in the *Human Role*	Safety × Agentic AI	·	○	·	●
	Academic integrity × Agentic AI	○	·	·	●
	Human role × Gen AI	·	·	·	●
	Human role × Agentic AI	·	·	·	●
	Human role × Ambient AI	·	·	○	●
	Future jobs × Gen AI	·	·	·	●
	Future jobs × Agentic AI	·	·	·	●
	Future jobs × Ambient AI	·	·	·	●

4. A Separate Category Type: *Environmental Impact*

The three **Environmental Impact** cells in the original matrix are not connected directly to any of the four trust categories. They concern physical resource consumption (energy, water, carbon, e-waste), not the trustworthiness of an interaction or decision. That does not mean that AI-related environmental questions are not important – quite the contrary – but it does mean that they may be most clearly discussed as a category unto itself, independently of trust concerns. Thus, for instance, the environmental impact of a data center needs to be discussed and weighed in terms of resources consumed, pollution generated, jobs created, and value created, independently of trust issues – not that different from the conversation around any other factory, really.

5. Changes Over Time

Looking at the table above, a pattern emerges. As we progress from the first to the fourth category, the primary AI types for each trust category look like:

- Category I (*Output*): 100% Gen AI
- Category II (*Process*): 20% Gen AI, 60% Agentic AI, 20% Ambient AI
- Category III (*Data Relationship*): 14% Gen AI, 14% Agentic AI, 72% Ambient AI
- Category IV (*Human Role*): 25% Gen AI, 50% Agentic AI, 25% Ambient AI

But Gen AI, Agentic AI, and Ambient AI have not all come onto the scene simultaneously and fully developed. They have followed each other in sequential order, with each going through stages that we could describe as:

Long-Term Prospective ➔ Short-Term Prospective ➔ Active

Right now, Gen AI is fully Active, Agentic AI is transitioning from Short-Term Prospective to Active, and Ambient AI is transitioning from Long-Term Prospective to Short-Term Prospective. And as each form of AI more fully approaches its active form, we should correspondingly expect the trust categories to become more salient over time, in proportion to the percentage of the AI form involved:

- All four categories are present from the get-go, since Gen AI is present in all four, and no category "goes away" by itself – all four remain in the mix;
- Trust in the *output* is most salient at the beginning, since it is solely dependent upon Gen AI; trust in the *process*) and trust in the *human role* increase in salience next; and trust in the *data relationship* is the last to increase in salience;
- The fact that later categories connect to larger numbers of matrix items may be perceived as an "acceleration" in trust issues;
- *Environmental Impact* is present throughout, but how it is perceived/discussed may shift as a result of the evolving salience of the four categories.

Thinking about these patterns may both help understand the shifts in public conversations about AI and education, and suggest possible directions for information communication and policy design.

6. Some Final Thoughts

- The four trust categories were developed in the context of the current conversation about AI and education. However, a second look reveals that all four have been present in some fashion in the past – for instance, trust in *output* applies just as well to a paper textbook, trust in *process* to traditional learning assessment procedures and measures, trust in the *data relationship* to the institutional framing and certification of student learning, and trust in the *human role* to the roles and actions of the people staffing those institutions.
- While the current categories focus on perceptions and uses of AI in education, it is likely that a similar approach would also be of use in analyzing and developing policies for wider workplace scenarios – but that's a task for another day.