Teaching the Unknown: A Pedagogical Framework for Teaching With and About AI

Brian Ballsun-Stanton[1] and Jodie Torrington[2]

[1]Faculty of Arts, Macquarie University, Sydney, Australia,
brian.ballsun-stanton@mq.edu.au (corr. author).
[2]Macquarie School of Education, Macquarie University, Sydney, Australia.

Generative artificial intelligence (AI) has disrupted education systems worldwide. This disruption necessitates pedagogical approaches that embrace uncertainty while developing student agency. We examined how decoupling task success from assessment outcomes created environments where students developed critical AI literacy through structured risk-taking. Drawing on Transformative Learning Theory and Rumsfeld's epistemological matrix as interpretive frameworks, we analysed an experimental undergraduate AI unit across three disciplinary streams: Ancient History, Philosophy, and Politics and International Relations ($N = 23$). Data included student reflections, classroom observations, and AI interaction logs collected over a 13-week semester. Our pedagogical framework operationalised four interconnected pillars: risk-embracing assessment structures, intentional classroom culture development, systematic navigation of technological uncertainty, and facilitation of transformative learning experiences. This paper presents the implications of these pillars for 1) educational theory, where productive failure serves as an effective pedagogical strategy; 2) educator praxis, viewing AI as a textual technology that extends the capabilities of the humanities; and 3) implications for the university teaching context, where AI-enabled teaching should focus on reflection and process rather than demonstratable competencies.

Practitioner Notes

What is already known about this topic:

- Transformative Learning Theory demonstrates how adult learners restructure meaning perspectives through critical reflection on disorienting dilemmas, particularly when supported by intersubjective discourse.
- Students often approach AI systems with polarised misconceptions, viewing them as either infallible authorities or prohibited cheating tools. The technology's rapid evolution creates pedagogical challenges for educators working with frontier models.
- AI is a textual technology particularly suited to humanities methodologies, offering opportunities for disciplinary knowledge application.
- Productive failure enhances learning when properly contextualised and reflected upon.

What this paper adds:

- A four-pillar framework for teaching AI literacy through disciplinary applications in humanities contexts;
- Practical strategies including a collaborative "classroom grimoire" for prompt iteration and educator vulnerability as pedagogical method, demonstrating how epistemic humility enables effective AI literacy development; and
- Evidence that humanities educators possess foundational analytical competencies directly transferable to AI pedagogy, challenging deficit-based professional development models.

Implications for practice and/or policy:

- Effective AI pedagogy requires sustained investment in psychological safety through structured reflection cycles and formative assessment, challenging traditional mastery-based educational models.
- Institutional support must address the capability gap between advanced and basic AI models, as this differential fundamentally alters pedagogical possibilities and exacerbates educational inequities.
- Universities require new quality frameworks that value process-focused assessment and epistemic uncertainty navigation over demonstrable competency acquisition.

1 Introduction

Generative artificial intelligence (AI) has disrupted education systems worldwide, necessitating reconceptualisation of pedagogical approaches and assessment practices. While education leaders grapple with broad concepts of AI literacy and assessment reform (Baele et al., 2024; Liu & Bates, 2025; Lodge et al., 2023; UNESCO, 2023), institutional responses have often prioritised detection and restriction over integration and innovation. This reactionary approach frequently frames AI through the morally loaded lens of "plagiarism," attempting to maintain traditional educational paradigms. However, AI has exposed pre-existing problems with effort regulation and self-regulated learning (Gkintoni et al., 2025; Torrington et al., 2023). We observe that students increasingly defect, in the game theoretic sense, from educational activities they perceive as having diminished value (i.e. Hicks and Kitto, 2025; Ballsun-Stanton and Khalid, 2025 in this volume). As such, there is an urgent need to move beyond simplistic notions of AI as a "cheating engine" and develop pedagogical frameworks specifically designed for this technological paradigm shift[1].

Teaching with and about AI presents unique pedagogical challenges due to the technology's emergent properties, shifting capabilities, and inherent uncertainties. Our approach of teaching both the pragmatics and ethics of AI extends Beale's (2024) "critical AI literacy", a capacity to engage thoughtfully with AI's limitations, biases, and potential societal impacts (see also UNESCO, 2023). Students must not only engage thoughtfully, but must be able to judge when and how to use AI, and then also to use it.

In contrast, traditional university models that emphasise certainty, correctness, and predictable outcomes are in conflict with these characteristics (Hoidn & Kärkkäinen, 2014; Khalaf & Zin, 2018; Margetson, 1994). When student education centres on mimetic reproduction of factual content, AI-generated "safe outputs" present an attractive alternative to engaging in genuine learning processes, particularly for students who believe the technology produces superior results to their own efforts. This situation is exacerbated by systemic pressures on both educators and students. Overcrowded curricula, limited time, cost-cutting imperatives, and high-stakes assessment environments discourage experimentation. These pressures create conditions where experimentation and failure are perceived as unacceptably risky (Henderson et al., 2022; Huang et al., 2023; Rosenberg, 2023; Smith, 2020).

Students typically enter educational settings with what we term external AI-LOC (AI Locus of Control), viewing AI systems as possessing inherent agency and authority rather than functioning as tools directed by human operators (see Torrington et al., 2025). This conceptualisation, rooted in Rotter's Locus of Control theory (Rotter, 1966), represents a significant barrier to developing sophisticated AI literacy. Students initially defer to AI rather than direct it, accepting outputs without critical evaluation. As one of our EAL/D students reflected:

> Yeah, because I live in a [foreign] student accommodation, so I have lots of first year student friends. And I think it's quite interesting, like, some of them treat the AI as a God. So

[1] We use the term in the Kuhnian (1994) sense, insofar as our normal explanations and understandings of effective education no longer serve to describe the world we and our students live in. This threatens "revolutionary science" where two paradigms are competing for explanatory power: AI *qua* cheating versus AI *qua* deep-learning tool.

> they believe what AI said, they don't do the reading anymore because they think the AI can do the summary for them. (Class Transcript, 10 October 2024)

As educators, we need to support students in shifting to an internal AI-LOC, where students perceive themselves as having primary agency in the human–AI interaction. This mental shift requires substantial time, psychological safety, and structured learning experiences currently absent from most educational and workplace contexts (A. C. Edmondson, 2018, p. 8). This pedagogical reorientation necessitates institutional support for capability development. Beyond this, as AI is rapidly evolving and the contexts and affordances of use are still being discovered, there is no settled set of skills that we can expect educators to possess. Therefore, the development of pedagogical frameworks specifically designed to navigate technological uncertainty thus becomes an essential first step, along with intentional classroom culture change, educator training, and a renegotiation of social norms about what it means to participate in education.

This paper argues that effective AI education requires pedagogical approaches that embrace uncertainty, structure risk-taking, and develop student agency amid technological indeterminacy. Discussing our insights from teaching an experimental undergraduate AI unit, we contextualise them with Transformative Learning Theory (TLT, Mezirow, 1978, 1997) and the epistemological framework of Rumsfeld's matrix (Daase & Kessler, 2007; Pawson et al., 2011; Rumsfeld, 2002). We discovered that, in teaching this class, four pillars provided the foundation for our pedagogical approach:

- Risk-Embracing Assessment Structure: decoupling task success from assessment outcomes creates educational environments where students can learn how to play with (and therefore use) AI;
- Navigating Technological Uncertainty: embracing the unknown unknowns of the jagged AI technological frontier;
- Intentional Classroom Culture Development: emphasising metacognitive knowledge and process as well as building trust and rapport through educator vulnerability; and
- Facilitating Transformative Learning: applying disorienting dilemmas, critical reflection, and communal discourse to change how students thought about AI use.

This offering of the class was a launching-off point for broader normalisation of AI teaching as part of a humanities curriculum. Our intention was to discover, through trial and error, what aspects of teaching with and teaching about AI had salient resonance with students: we had to figure out how to teach the unknown. While this classroom experimentation is hypothesis-generating rather than hypothesis-testing (Nosek et al., 2018), we believe that what we discovered when teaching lays the groundwork for further experimentation and testing of how to teach with and about AI.

This chapter lays the groundwork for a pedagogical framework on teaching AI using evidence collected from our teaching. We contextualise and justify our experiences by referring to longitudinal documentation of student and educator experiences, including assessments, class transcripts, and class observations. The context of our reporting is an experimental undergraduate AI unit offered by the Macquarie University Faculty of Arts in 2024. This unit applied discipline-specific knowledge to investigate the implications of the use of AI for three humanities disciplines' problems: the Ancient History stream built a 160 source annotated bibliography on Caligula (Green et al., 2024), philosophers investigating AI and propaganda, and simulations on Politics and International Relations topics (Torrington et al., 2025). This class was offered as a university-sanctioned experiment: students consented to having deidentified transcripts of lecture recordings, their assessments, and classroom observation of teaching recorded as part of our research on how to teach this topic[2].

Our argument develops through several stages. First, we establish a dual theoretical framework: TLT illuminates how students transform their understanding of AI through navigating disorienting dilemmas and engaging in critical reflection, while Rumsfeld's matrix provides an epistemological structure for categorising and converting technological uncertainties into actionable knowledge. We then detail our pedagogical design, beginning with the course context and three disciplinary streams before presenting our four-pillar pedagogical framework as described above. The discussion section examines implications across three levels: for educational theory (challenging traditional knowledge progression models), for educator praxis (recognising humanities methodologies as foundational for AI literacy), and for the university teaching context (addressing institutional requirements for systematic implementation). We conclude by synthesising how these theoretical and practical insights provide colleagues with foundations for engaging with AI-mediated educational transformation while acknowledging the ongoing adaptation required.

2 Theoretical Frameworks

We employ two theoretical frameworks: Transformative Learning Theory addresses the psychological processes of transformation, while Rumsfeld's matrix provides tools for navigating epistemological uncertainty in AI contexts.

2.1 Transformative Learning Theory (TLT)

TLT describes how adult learners restructure their meaning perspectives when confronted with experiences that existing understandings cannot accommodate (Mezirow, 1978). At its core, TLT posits that transformative learning occurs through critical reflection on assumptions that unconsciously influence thought and action. Mezirow (1997) articulates this as "the process of effecting change in a frame of reference" (p. 5). These frames of reference comprise two dimensions: habits of mind (broad, abstract, orienting predispositions) and points of view (specific belief systems). Transformation occurs when learners encounter experiences that expose the inadequacy of existing meaning schemes, prompting fundamental reassessment of underlying assumptions.

The transformative process unfolds through ten phases that Mezirow (1991) identifies. While these phases suggest linear progression, subsequent research demonstrates that transformation often occurs through recursive cycles rather than sequential stages

[2] This data is available at the Australian Data Archive under mediated access controls to any researcher with appropriate Ethics clearance

(Moström Åberg, 2023). Table 1 illustrates how these theoretical phases manifested in our AI education context, drawing primarily from one student's documented journey (POIR Student 23) as a representative case of the transformative progression.

Table 1 Mezirow's Ten Phases of Transformative Learning Applied to AI Education

	Transformative Process	**Classroom Example**
1	Experiencing a disorienting dilemma	Being instructed to use AI for Greek Civil War preparation; receiving unexpectedly acceptable responses
2	Undergoing self-examination	"Many of my issues stemmed from anthropomorphising Claude" (Student 23, Week 6)
3	Conducting critical assessment of assumptions	Annotating 41 pages of transcripts identifying effective/ineffective prompts (see Figure 1)
4	Recognising shared experiences	Class-wide debriefing sessions
5	Exploring options for new roles	Experimenting with different prompting strategies
6	Planning action courses	Pre-simulation practice runs to test character mechanics
7	Acquiring necessary knowledge and skills	Iterating prompts from classroom grimoire; choosing when/how to engage AI
8	Provisionally trying new roles	Designing subsequent simulations
9	Building competence and confidence	Delivering AI-driven simulations (Student 23, Week 12)
10	Reintegrating based on new perspectives	Creating policy briefs for university AI integration (Student 23, Week 13)

This social dimension proves particularly relevant for AI education contexts. Fleming (2018) introduces Honneth's recognition theory to address critiques that TLT inadequately theorises social dimensions of transformation. Fleming argues that intersubjective recognition constitutes a necessary precondition for critical reflection and discourse, positioning social validation as foundational rather than supplementary to individual transformation (see phases 4 and 7 in Table 1). Student sensemaking was evident in group and classroom discussion, building on weekly practice and individual reflection. When they could see how others were reacting to their experiences, it helped them reflect and contextualise their achievements and failures.

The concept of disorienting dilemmas has received particular theoretical attention. Laros (2017) distinguishes between externally imposed dilemmas (life crises, job loss) and intentionally designed educational experiences that surface assumptions for examination. This distinction suggests that educators can deliberately structure learning environments to catalyse transformation through carefully calibrated challenges to existing meaning schemes. Recent empirical work by Feng et al. (2025) demonstrates how interdisciplinary learning contexts generate particularly potent disorienting experiences by exposing disciplinary assumptions typically left unexamined. In a semester-long class, this afforded us multiple opportunities to trigger a disorienting dilemmas within their professional context and within AI tool use and then resolve the dilemma through reflection and discussion.

For AI education contexts, TLT offers several theoretical advantages. First, the framework explicitly addresses how adults reconstruct meaning when confronted with paradigm-shifting phenomena. Second, its emphasis on critical reflection aligns with the need to interrogate assumptions about knowledge, authority, and human-technology relationships. Third, the theory's evolution to incorporate social and cultural dimensions provides tools for understanding how collective meaning-making shapes individual transformation. These theoretical resources prove essential given AI's capacity to challenge fundamental assumptions about learning, creativity, and human agency.

The processual nature of transformative learning particularly suits technological contexts characterised by rapid change. Rather than viewing transformation as a destination, TLT now conceptualises ongoing cycles of assumption examination and perspective revision (Moström Åberg, 2023). This theoretical orientation accommodates the reality that AI capabilities evolve continuously, requiring learners to maintain what Mezirow (1997) terms "reflective judgment" as a permanent disposition rather than achieving fixed understanding:

> Education that fosters critically reflective thought, imaginative problem posing, and discourse is learner-centered, participatory, and interactive, and it involves group deliberation and group problem solving. Instructional materials reflect the real-life experiences of the learners and are designed to foster participation in small-group discussion to assess reasons, examine evidence, and arrive at a reflective judgment. Learning takes place through discovery and the imaginative use of metaphors to solve and redefine problems. (p. 10)

Such theoretical flexibility becomes essential when AI in an educational context undergoes constant transformation, rendering static competencies obsolete while elevating critical reflection and adaptability as core educational outcomes.

When teaching AI, the technology itself functions as a primary disorienting dilemma. Some of our students entered with preconceived notions about AI, often viewing it dichotomously as either a "cheating engine" or an infallible authority (Torrington et al., 2025). These simplistic conceptions create significant barriers to developing sophisticated and critical AI literacy. One student reflected this initial disorientation:

> Initially approaching AI as a mere tool for information retrieval, I gradually discovered the complexity and nuance required for meaningful engagement with these systems. ... It became increasingly apparent that I struggled to develop truly reflective questions and move beyond surface-level technical analysis. (Week 13, Student 8)

TLT allows us to understand how to teach AI through the provocations of navigating uncertainty, fostering critical reflection, and cultivating adaptability. These are core competencies required to engage with AI's emergent capabilities and ethical complexities. TLT thus provides a scaffold for designing pedagogies that prepare learners to interrogate AI's black box nature (Chaudhary, 2024), challenge assumptions about knowledge production, and develop agency in an AI-augmented world (Mollick, 2024a).

Classroom technology applications typically remain stable throughout a semester, allowing educators to maintain consistent learning designs. Expecting to teach AI disrupts this stability. Educators must navigate continuous cycles of disorientation across multiple dimensions: evolving usage norms, rapidly changing models and services, and the iterative development of effective prompting strategies. Traditional institutional constraints compound these challenges, as university IT departments' procurement cycles inevitably lag behind frontier model releases, limiting access to current capabilities.

However, by using a service offering access to the latest frontier models, showcasing this rate of change to students presents opportunities for disorienting dilemmas. TLT offers a framework for educators to guide students through sense-making of change: "[I]t is so important that adult learning emphasize contextual understanding, critical reflection on assumptions, and validating meaning by assessing reasons" (2000, p. 3). AI use and development continually produces disorienting dilemmas. These experiences create cognitive dissonance that cannot be resolved within one's existing belief system, necessitating a fundamental reconsideration of assumptions and perspectives.

2.2 Rumsfeld's Matrix as Epistemological Framework

Rumsfeld's epistemological matrix, developed from his 2002 formulation and refined by scholars (Daase & Kessler, 2007; Pawson et al., 2011; Rumsfeld, 2002), categorises knowledge into four states:

- known knowns (facts we are aware of and understand);
- known unknowns (recognised gaps in knowledge);
- unknown knowns (tacit knowledge we possess but do not recognise); and
- unknown unknowns (gaps in knowledge we are not yet aware of).

The matrix offers epistemological descriptions for AI education contexts characterised by emergent and probabilistic properties. Students entered the class with significant unknown unknowns: they lacked awareness of both what they knew and did not know about AI systems. Teaching AI requires transforming these unknown unknowns into known unknowns through structured experimentation and critical reflection.

This epistemic challenge parallels what Pawson et al. (2011) identify in policy contexts as "the predicament of evidence-based policy," where acknowledging uncertainty becomes prerequisite to developing sophisticated understanding. In AI education, this predicament intensifies: the knowledge corpus remains unsettled, pedagogical conventions are emerging, and usage norms continue evolving through active negotiation among educators, students, and institutions.

Together, TLT and Rumsfeld's matrix provided complementary frameworks that informed our pedagogical framework. While TLT gave us a framework for the processes through which students navigate perspective transformation when confronted with AI's capabilities and limitations (Section 3.2.4), Rumsfeld's matrix offered an epistemological structure for systematically converting "unknown unknowns" into actionable knowledge (Section 3.2.2). The progression from theoretical conceptualisation to practical implementation required careful attention to both assessment structures and classroom cultural development.

3 Pedagogical Design

This section examines the practical operationalisation of our theoretical frameworks, detailing both the structured design elements and classroom implementation strategies that facilitated student development of AI literacy and agency. Our pedagogical design integrated disciplinary knowledge with exploratory AI interaction through a scaffolded progression of learning experiences. Rather than positioning AI literacy as a generic skill set divorced from contextual application, the course design embedded technological exploration within discipline-specific inquiries, allowing students to develop simultaneously their subject knowledge and technological agency.

3.1 Course Design

The experimental unit was implemented within ARTS3500, the Faculty of Arts capstone unit at Macquarie University. This institutional context provided a suitable framework for innovative pedagogical approaches, as the capstone structure emphasised integrative learning experiences that synthesise disciplinary knowledge with transferable skills applicable to post-graduation contexts. Students self-selected into the experimental offering through an expression of interest process that outlined the streams, objectives, and research participation components.

The unit was structured around three distinct disciplinary streams, each applying domain-specific methodologies and epistemologies to the common challenge of effective AI interaction:

Ancient History Stream: This stream ($N = 3$) engaged students in a comprehensive investigation of the historical accounts and modern interpretations of the Roman Emperor Caligula. Students employed AI tools to translate and analyse multilingual sources, developing a collaborative annotated bibliography comprising approximately 160 sources. This approach integrated traditional historiographical methods with technological affordances, enabling students to interrogate how AI-assisted analysis might transform historical research practices. The primary deliverable was a substantive annotated bibliography examining Caligula's "madness" across historical periods. This deliverable served as both an authentic scholarly artefact and a vehicle for meta-cognitive reflection on AI-human collaborative research.

Philosophy Stream: Focusing on the intersection of AI capabilities and philosophical inquiry, this stream ($N = 12$) examined questions of "Existential risk, Truth, and Democracy" through the lens of propaganda analysis. Students systematically employed AI tools to both generate and deconstruct propaganda, creating philosophical dialogues that explored the ontological and epistemological implications of AI-generated content. This stream foregrounded ethical considerations, investigating how philosophical frameworks could inform responsible AI utilisation while simultaneously examining how AI might transform philosophical discourse itself.

Politics and International Relations Stream: This stream ($N = 8$) investigated power dynamics in domestic, regional, and international systems through AI-augmented simulations. Students engaged in the Council on Foreign Relations (CFR) "Opposing Communism in 1947" mini simulation (2021), using AI to help roleplay as Truman's cabinet ministers. The class then decided, rather than running a second scenario from CFR, to split into groups and develop and then run their own scenarios. The central theme of this stream was to examine how the use of AI may introduce bias, both as a function of prompting, but also inherent to the model itself.

While we were identifying the boundaries of AI in these streams, however, each stream relied on discipline-specific knowledge as we targeted metacognitive knowledge within evaluation and creation tasks (Krathwohl, 2002, p. 214). Applying students' discipline-specific knowledge was essential to ground their judgement such that they could understand, contextualise, and evaluate the utility of AI in their soon-to-be professional contexts.

The course design also incorporated specific structural elements to address the challenge of ongoing technological change during the semester. We implemented a recurring "Show Me Your Prompts" discussion at the beginning of each class session, creating a "ritual" allocating approximately thirty minutes of each two-hour seminar for everyone, students and educators, to discuss and reflect on AI interactions, both successful and unsuccessful, from the previous week. This practice served multiple functions: activating prior knowledge (Dinsmore et al., 2008; Zimmerman, 2002), normalising failure and experimentation (Darabi et al., 2018), and creating a communal repository of evolving technical insights, a shared "grimoire" of effective prompts (see Mollick, 2023). This communal grimoire provided a collaborative mechanism for documenting, iterating upon, and disseminating emerging technical knowledge. The grimoire served as a student-sourced source-of-prompts complimenting and iterating on the educator's prompts. The presence of a communal document allowed an environment of psychological safety by explicitly permitting the growth of a community of practice (Lave, 1991) around prompting, allowing iteration from established prompts, rather than *ab novo* prompt creation each time.

Each disciplinary stream maintained its distinct methodological approaches and epistemological foundations while sharing common pedagogical elements designed to support AI literacy development. The streams converged through cross-stream mentoring activities, shared reflective practices, and collaborative development of effective prompting strategies. The integration of disciplinary objectives with technological exploration created authentic learning contexts where students could simultaneously develop subject matter expertise and technological agency.

3.2 Pedagogical framework

Our pedagogical framework was structured around four interconnected dimensions:

1. Risk-embracing assessment design;
2. Navigating technological uncertainty;
3. Intentional class-culture development; and
4. Facilitating transformative learning experiences.

This approach addressed the challenge of teaching with a rapidly evolving technology characterised by emergent properties and inherent indeterminacies. Through these complementary strategies, we established conditions where students could progress from external to internal AI-LOC, developing increasingly sophisticated mental models of AI capabilities and limitations while cultivating their own agency in human–AI collaboration.

3.2.1 Risk-Embracing Assessment Structure

The foundational principle guiding our assessment design was the deliberate decoupling of task success from assessment outcomes. This approach directly addressed a critical challenge in AI education: students' resistance to experimentation stemming from grade-related anxieties, as reported in classroom debriefs. By evaluating students' thoughts-about-AI rather than the quality of AI outputs, we established conditions where productive failure became a valuable learning resource rather than a punitive experience.

Our assessment framework employed backward design principles (Wiggins & McTighe, 2005). The institutional constraints of the existing capstone unit structure served as our starting point. Within these predetermined requirements, we identified essential capabilities for effective AI interaction in professional contexts. Critical evaluation of the utility of students' prompts and consequent AI outputs emerged as a primary capability, alongside recognition of model limitations and development of domain-specific prompting strategies. These core competencies informed our construction of progressive assessment activities that scaffolded students toward increasing technological agency. We prioritised the development of critical AI literacy (Baele et al., 2024; UNESCO, 2023) over tool-specific competencies throughout the assessment design. This approach acknowledged a fundamental reality of AI education: the ability to critically evaluate and adapt to new AI systems holds greater long-term value than mastery of any particular platform or model. As Picasso et al. (2024) argue, academics designing authentic assessments must embed "layers of critical data and AI literacy" to help students "develop the skills needed to thrive in the datafied society... explaining why these are relevant to their future professional and everyday lives" (p. 296). In rapidly evolving technological landscapes, these transferable evaluative competencies provide students with sustainable professional capabilities rather than ephemeral technical skills.

Core to our assessment design was the implementation of a formative scaffolding, progressively building AI prompting sophistication across the semester (Weeks 6, 9, 12, and 13), allowing for incremental capability development through low-stakes formative experiences (within the constraints allowed by the grading-structure of the unit). This structure created multiple opportunities for students to recalibrate their understanding of AI capabilities and limitations while receiving consistent reinforcement that exploration and critique were valued over perfection. All assessments deliberately incorporated significant reflective components to establish psychological safety, as defined by A. C. Edmondson (2018) as:

> [T]he belief that the work environment is safe for interpersonal risk taking. The concept refers to the experience of feeling able to speak up with relevant ideas, questions, or concerns. Psychological safety is present when colleagues trust and respect each other and feel able – even obligated – to be candid. (p. 24)

The assessment structure particularly emphasised documentation and analysis of AI failures, directly countering students' initial resistance to experimentation. Across all streams, assessment rubrics explicitly rewarded "innovative, well-documented use of multiple AI tools" and "comprehensive, insightful annotations of AI logs." In the Ancient History stream, for instance, the "Effective Use of AI Tools" criterion specifically evaluated students' ability to document processes and critically analyse failures rather than produce successful outputs. This focus on process, reflection, and iteration as "Effective Use" rather than output allowed students freedom to experiment, even as the scope grew.

By establishing assessment structures that valued documenting learning processes over products, we created educational and environmental conditions where students could progressively develop an internal AI-LOC through structured experimentation. As one student said in their final reflection:

> My perceptions of AI and risk have changed throughout the semester. By interacting with these tools with scrutiny, their limitations become very obvious. I think prior to the unit I had overestimated how well these tools would perform on basic tasks. ... So it can take quite a lot of trial and error to even get an LLM to complete basic tasks if specificity is needed. Overall I think this soothed some of the anxiety I had around these tools. That is not to say all anxiety around AI has been soothed. (Week 13 Student 1)

The systematic reinforcement of reflection over results effectively mitigated students' initial tendency to defer to AI outputs, facilitating their transition from passive consumers to active directors of technological tools. Students even recommended this "fail-forward" approach in their prescriptions for AI in future university classes:

> These specialised courses should create safe, experimental spaces where students can actively engage with AI technologies, share experiences, and develop practical skills through hands-on learning. ... Crucially, these learning environments should embrace a 'fail-forward' approach, where students are encouraged to experiment with AI prompting and applications without fear of negative consequences. This acceptance of failure as a valuable learning tool is essential for developing advanced AI literacy, as it allows students to identify, analyse, and learn from their mistakes, ultimately leading to more sophisticated and ethical AI usage. (Week 13, Student 3)

The classroom context, educator expertise, repeated formative assessments, and repeated exposure to AI and critique of AI models provided the foundational psychological safety required for productive engagement with the inherent uncertainties of emergent AI systems.

3.2.2 Navigating Technological Uncertainty

A central challenge in teaching with and about AI lies in its inherently probabilistic nature and emergent capabilities; qualities that resist traditional educational approaches predicated on stability and predictability. Our pedagogical design explicitly incorporated technological uncertainty as a fundamental learning dimension rather than an incidental challenge. Our approach acknowledged and embraced what Feng et al. (2025) identifies as essential for transformative learning: the creation of structured environments where students could safely encounter disorienting dilemmas that prompt critical reflection. The basis of which was that everyone in class (students and teachers) was provided paid subscriptions to Perplexity, so that they could experiment with and experience all of the latest frontier models as they were released throughout the semester.

Additionally, beyond the fact that model availability changed month-by-month, we also deliberately explored and tested the boundaries of the consequences of AI output: creating situations where bias, refusals, and other problems were more likely to come to light. In contrast to educational approaches that position technology as a reliable tool with predictable behaviours, we deliberately engineered instances where students could induce and reflect on AI failure: hallucinations, inconsistencies, and biased outputs. Beyond these obvious failure modes, we regularly pushed students to think about model prompting and output in terms of utility. We asked them to annotate their prompts and think about if their prompts produced useful output as a consequence (see blue/pink highlighting in Figure 1). This reflection on the quality of the output caused students to think about the consequences of their prompts and choice of model when considering AI responses.

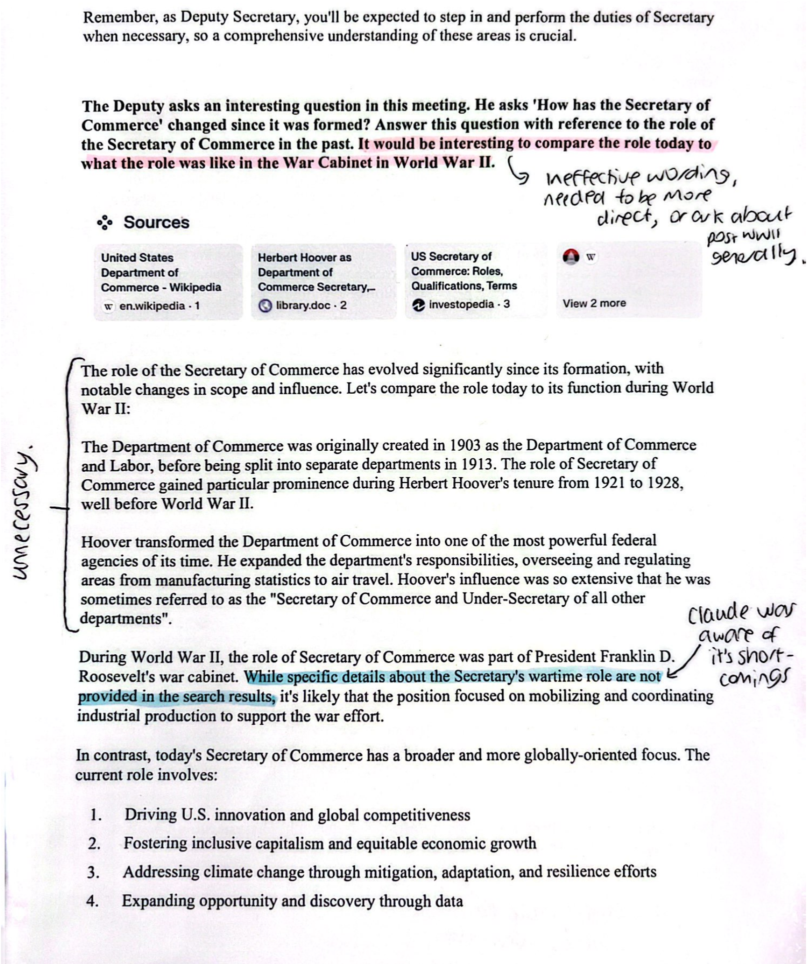
Our approach to navigating technological uncertainty aligned with Manz and Suárez (2018)'s suggestions for teaching under conditions of scientific uncertainty, which emphasises:

> [A] context where students encounter ambiguity and explore decisions about how to investigate and make sense of natural phenomena establishes a need for practices, helps students frame their activity as about finding something out, and provides a context for authentic discussion in which practices can be developed and refined (p. 773).

By positioning technological uncertainty as a central learning dimension rather than a peripheral concern, our course design embraced risk and created educational conditions where students could develop critical AI literacy (Baele et al., 2024).

Rather than prescribing specific modes of AI interaction, we created bounded experimental spaces where students could explore capabilities and limitations within a psychologically safe environment. This experiential approach served as the primary mechanism for epistemological transformation. Students progressively developed granular understanding of when, how, and why to employ specific AI tools through cycles of experimentation, failure, and reflection. The process directly countered the "black-boxing" (Chaudhary, 2024) of AI by revealing the technology as a complex system with differentiated capabilities and constraints rather than a monolithic entity. As Pawson et al. (2011) conclude regarding evidence-based policy, "the whole point is the steady conversion of 'unknowns' to 'knowns'" (p. 543). Our classroom methodology provided the temporal and psychological space necessary for this conversion through productive failure.

Clarifying Known Knowns: The epistemological progression manifested across all four categories of Rumsfeld's matrix. Initial instruction on prompt structuring and model capabilities established a foundation of known knowns, though we intentionally limited these direct teaching moments to avoid prescriptive interaction patterns. This restraint proved essential given what Daase and Kessler (2007) identify as the dynamic nature of knowledge categories in complex systems. The rapid evolution of AI capabilities meant that today's established techniques could have become obsolete



Remember, as Deputy Secretary, you'll be expected to step in and perform the duties of Secretary when necessary, so a comprehensive understanding of these areas is crucial.

The Deputy asks an interesting question in this meeting. He asks 'How has the Secretary of Commerce' changed since it was formed? Answer this question with reference to the role of the Secretary of Commerce in the past. It would be interesting to compare the role today to what the role was like in the War Cabinet in World War II.

ineffective wording, needed to be more direct, or ask about post WWII generally.

Sources

United States Department of Commerce - Wikipedia
en.wikipedia · 1

Herbert Hoover as Department of Commerce Secretary,...
library.doc · 2

US Secretary of Commerce: Roles, Qualifications, Terms
investopedia · 3

View 2 more

The role of the Secretary of Commerce has evolved significantly since its formation, with notable changes in scope and influence. Let's compare the role today to its function during World War II:

The Department of Commerce was originally created in 1903 as the Department of Commerce and Labor, before being split into separate departments in 1913. The role of Secretary of Commerce gained particular prominence during Herbert Hoover's tenure from 1921 to 1928, well before World War II.

Hoover transformed the Department of Commerce into one of the most powerful federal agencies of its time. He expanded the department's responsibilities, overseeing and regulating areas from manufacturing statistics to air travel. Hoover's influence was so extensive that he was sometimes referred to as the "Secretary of Commerce and Under-Secretary of all other departments".

unnecessary.

During World War II, the role of Secretary of Commerce was part of President Franklin D. Roosevelt's war cabinet. While specific details about the Secretary's wartime role are not provided in the search results, it's likely that the position focused on mobilizing and coordinating industrial production to support the war effort.

Claude was aware of it's short-comings

In contrast, today's Secretary of Commerce has a broader and more globally-oriented focus. The current role involves:

1. Driving U.S. innovation and global competitiveness
2. Fostering inclusive capitalism and equitable economic growth
3. Addressing climate change through mitigation, adaptation, and resilience efforts
4. Expanding opportunity and discovery through data

Fig. 1 Assessment from Student 23, Week 6. Used under extended consent provisions in Macquarie University HREC Project 16084. This simulation was individuals in class engaging in CFR's "Mini Simulation: Opposing Communism in 1947." This is an annotated printout of the student's prompts and Claude 3.5 Sonnet's response from Perplexity. Effective prompt and outputs are highlighted in blue. Ineffective prompting is highlighted pink. Scan and annotation copyright remains with Student 23.

within months. As they observe, "the foundations of the known knowns crack open when basic concepts of political discourses are contested, when analogies do not work and mechanical if-then sentences are invalidated" (p. 420). The parallel to our technological context was striking: each major model update potentially invalidated previous interaction strategies.

Boundary-making with Known Unknowns: Students' progression from known unknowns to operational knowledge proved particularly significant for developing agency. By week 8, classroom discussions revealed sophisticated differentiation between model capabilities. Students articulated specific limitations and contextual applications, moving beyond binary evaluations to nuanced understanding. This evolution was evident when students began selecting models strategically: “I used Opus for this one, and I used Sonnet for that one,” demonstrating conscious tool selection based on task requirements (Class Transcript, 24 September 2024).

Experiencing Unknown Knowns: The transformation of unknown knowns represented perhaps the most profound shift. Many students initially failed to recognise how their disciplinary expertise could inform AI interaction. Through structured reflection and collaborative experimentation, they discovered that their existing skills provided essential capacities for evaluating and directing AI outputs. This activation of tacit knowledge through reflexive practice demonstrated the value of creating space for discovery rather than instruction.

Unknown Unknowns: The presence of unknown unknowns created productive pedagogical tension throughout the semester. Neither educators nor students could predict how rapidly evolving AI capabilities would reshape interaction possibilities. This fundamental uncertainty removed the possibility of definitive answers about “correct” usage, instead requiring continuous adaptation and epistemic humility. As Rumsfeld (2002) observed, these unknown unknowns “tend to be the difficult ones” (line 335). Yet embracing this difficulty as a pedagogical resource rather than an obstacle enabled authentic engagement with technological emergence. The almost monthly appearance of new capabilities and limitations transformed the classroom into a living laboratory for navigating uncertainty.

This systematic engagement with all four epistemic categories prepared students for the deeper epistemic work of developing technological agency. The intellectual capacities for understanding uncertainty needed to be complemented by social structures that supported risk-taking and collaborative learning. The following section examines how intentional culture development created the interpersonal conditions necessary for this transformative engagement with AI.

3.2.3 Intentional Classroom Culture Development

The implementation of our pedagogical approach centred on establishing classroom rituals and modelling behaviours that systematically built psychological safety. The recurring “show me your prompts” practice served as the cornerstone of our classroom culture and the basis for psychological safety when experimenting and failing with AI.

Each class session began a variation on with this deceptively simple invitation: “How I really want to start all of my classes is show me your prompts. What have you folks been doing with AI this week in any regard?” (Class Transcript, 8 August 2024). This ritual served multiple pedagogical functions beyond information sharing. Students presented both successful and failed AI interactions, normalising the full spectrum of technological engagement. The practice activated prior knowledge while simultaneously creating a collaborative learning environment where reflective experimentation was more important than successful performance.

The classroom dialogue revealed how this practice facilitated deeper engagement with AI limitations. When a student reported reliability issues with data visualisation, the ensuing discussion transformed individual frustration into collective learning:

> Student: “First of all, you can’t read the diagram very accurately because I have to input the data by myself, yes, and I find some answer is not correct as well.”
> Ballsun-Stanton: “This is one of those fundamental rules. Thou shalt not suffer an error to live. If there is an error or a refusal in the conversation, stop. Because once it refuses, it will keep doing that. Once it makes an error, it will keep doing that.” (Class Transcript, 8 August 2024)

This exchange demonstrates how the ritual created opportunities for immediate, contextualised instruction grounded in student experience rather than abstract principles. Ballsun-Stanton’s response reframed errors as systemic patterns requiring strategic intervention rather than isolated failures.

Educator vulnerability functioned as a deliberate pedagogical strategy rather than an incidental aspect of teaching. Live demonstrations frequently included real-time error recognition and correction, as evidenced during a prompting demonstration:

> “Never ever, getting an AI to do more than one thing. Because it’ll lose some of one of the things, right? Here I’m getting it to think through and then explain. And that’s a mistake. But only by walking [back to the lecture computer] did I realise that I made a mistake in this prompt.” (Class Transcript, 22 August 2024)

This public acknowledgement of error served multiple functions: modelling metacognitive awareness, demonstrating that expertise includes recognising mistakes, and establishing that the classroom valued learning processes over preformative competence.

The classroom culture explicitly celebrated discovery of AI limitations. Ballsun-Stanton’s response to student-reported errors, “So honestly, I’m delighted that you’re finding errors. I’m delighted that you’re running into these problems because that’s what I want you to play with and to get those spooky moments,” inverted traditional educational values that position errors as failures. This reframing is consistent with Darabi et al. (2018)’s findings that productive failure enhances learning when properly contextualised and reflected upon.

The communal grimoire emerged from these classroom interactions. As students shared prompting strategies and failure patterns, collective knowledge accumulated in both verbal exchanges and documented repositories. This collaborative knowledge construction exemplified what Lave (1991) identify as legitimate peripheral participation: students progressively moved from observing others’ AI interactions to sharing their own experiments to mentoring peers in effective strategies.

Physical and temporal considerations proved essential for cultural development. The two-hour seminar format provided sufficient time for extended discussion and experimentation. The consistent allocation of 30 minutes for prompt sharing created predictable space for vulnerability and exploration. The development of shared vocabulary emerged through regular interactions. Students progressively adopted technical terminology (“context window,” “system prompt,” “model refusal”) through experience and experimentation. Students knew that each session would begin with opportunities to share discoveries and difficulties, establishing psychological safety through routine rather than exhortation.

By the semester’s midpoint, students had internalised these cultural norms sufficiently to initiate peer mentoring. Data collected from classroom transcripts captures a student sharing their experiments with AI for witness examination preparation, including sophisticated strategies like preventing external searches and managing context windows. This unprompted knowledge sharing on topics outside of our specific class demonstrated that the classroom culture had successfully shifted from educator-centred information transmission to community-based collaborative learning. Students were taking processes from this unit, experimenting with them in the context of their other studies, and then bringing back their insights into the success and failure of these approaches to their other students—all because of "show me your prompts."

These intentional cultural practices created the interpersonal foundation necessary for the deeper transformative work of reconceptualising human-AI relationships. Without psychological safety and normalised experimentation, students could not have engaged productively with the disorienting dilemmas that AI interaction necessarily produces.

3.2.4 Facilitating Transformative Learning

Our classroom implementation deliberately structured learning experiences that facilitated transformative learning through systematic engagement with disorienting dilemmas, critical reflection, and communal discourse (Mezirow, 2000). Rather than treating these theoretical principles as abstract concepts, we embedded them within concrete instructional practices that progressively developed students' agency with AI technologies.

Disorienting dilemmas, fundamentally, were because students were entirely out of their comfort zone: they had no prior experience with AI, the normal academic achievements of demonstrating knowledge through writing essays were absent, and they were expected to create and display their discoveries to their peers. Critical reflection was an essential sense-making activity, grounding the dilemmas in retrospective and prospective student deliberation. By dwelling on prompting strategy and judgement, students demonstrating increasingly sophisticated conceptual understanding of how and when to use AI to support their learning and activities. This evolution in critical awareness directly evidenced what Teng and Yue (2023) identify as development of metacognitive regulatory strategies: the ability to monitor and evaluate one's own cognitive processes when engaging with complex tasks.

Communal discourse evolved from simple information sharing into sophisticated exchanges where students developed understanding of prompt strategies, output evaluation metrics, and differential model capabilities, complimented by the written grimoire. By week 8, students regularly experimented with different models and shared their insights about model use and prompting strategies with each other, both live and in the classroom grimoire:

> When I was using Claude the way that generative AI had been used stereotypically (as a search engine), it was clear early on that I was unable to achieve my required goals. But it was through learning how to prompt, and as mentioned, fail and collaborate, that this was overcome. ... [C]ollaboration has enabled us to become inquisitive and experiment with AI from a place of excitement and collective learning. ... [M]entoring and helping our peers has enabled us ourselves to learn, rather than the learning experience being an isolated process. Through engaging with peers via Teams, as well as through group iMessage chats, we've

> been able to mentor each other and share both failures and successes. (Week 9 reflection, Student 17)

This progression demonstrated the development of increased student technological agency, the capacity to make informed choices about technological tools based on critical evaluation of their affordances and limitations.

4 Discussion and Implications

Our pedagogical experiment surfaced tensions between traditional educational approaches and the realities of teaching with emergent AI technologies. Here we examine how our findings challenge existing theory, inform educator practice, and expose institutional barriers to meaningful AI integration. We also acknowledge both the limitations of our hypothesis-generating experimental context and the broader applicability of our insights.

4.1 Implications for Education Theory

Teaching effective AI use requires systematic navigation of all four epistemic states identified in Rumsfeld's matrix rather than the traditional progression from "known unknowns to known knowns" that characterises conventional higher education. Pawson et al. (2011) observe that evidence-based (geopolitical) practice involves "the steady conversion of 'unknowns' to 'knowns.'" However, in a teaching context, we observed that AI-enabled education fundamentally disrupted this linear progression by introducing technological contexts where unknown unknowns emerged continuously throughout the learning process. We observed that students necessarily moved through cycles of unknown unknowns: encountering AI capabilities they were unaware existed; to known unknowns: recognising specific limitations and biases. This cycle was a necessary precondition to developing stable known knowns about effective interaction patterns. Our experience challenges the traditional educational social contract that assumes stable knowledge foundations and predictable learning outcomes.

TLT provides an essential interpretive framework for understanding why AI education necessarily involves sustained discomfort and cognitive disruption rather than smooth skill acquisition. Mezirow (1997) positions disorienting dilemmas as catalysts for perspective transformation, and our classroom implementation demonstrated that technological uncertainty functioned precisely as such catalysts. Students' progression from external to internal AI-LOC occurred through structured confrontation and reflection upon AI's unpredictable outputs, incorrect responses, and contextual limitations. These scaffolded experiences systematically challenged students' initial assumptions about technological authority and AI output reliability. Rather than representing pedagogical failures, these discomforting encounters constituted essential learning mechanisms that enabled students to develop sophisticated technological agency. We were able to achieve these shifts through deliberate educator vulnerability and risk-taking as legitimate pedagogical strategies, countering traditional approaches that prioritise certainty and educator authority.

AI education requires educators to acknowledge and navigate genuine unknown unknowns in real-time classroom contexts, fundamentally challenging assumptions about pedagogical preparation and educator expertise. Traditional pedagogy assumes educators possess mastery of stable knowledge domains, enabling systematic curriculum planning and predictable learning progressions (Khalaf & Zin, 2018). However, when AI tools undergo frequent significant capability updates, interface modifications, and policy changes, educators cannot have definitive answers about optimal tools or interaction strategies at the start of a unit and expect those recommendations to remain valid throughout the teaching period. This technological volatility creates pedagogical adaptation requirements that demand epistemic humility: educators must model learning processes rather than knowledge transmission, demonstrating how to navigate uncertainty rather than providing predetermined solutions. Our classroom approach incorporated this uncertainty as a learning dimension, teaching students that effective AI interaction requires continuous adaptation and critical evaluation rather than mastery of fixed techniques.

These theoretical frameworks offer educators practical conceptual tools for navigating AI's pedagogical challenges. TLT reframes the discomfort of not having answers as pedagogically productive rather than professionally threatening. When educators understand technological confusion as a disorienting dilemma that catalyses learning, they can embrace vulnerability as a teaching strategy rather than concealing uncertainty behind false confidence. Similarly, Rumsfeld's matrix provides accessible vocabulary for discussing different types of uncertainty with students and colleagues, moving beyond simplistic "we do not know" admissions to nuanced recognition of what we know, what we know we do not know, and what we have not yet recognised as unknowable. This granular approach to uncertainty helps educators design learning experiences that explicitly address each epistemic category rather than treating all uncertainties as equivalent. Together, these frameworks help educators recognise that developing critical AI literacy requires fundamentally different pedagogical approaches than teaching established disciplinary content.

Understanding navigation between unknown unknowns through to known knowns enables intentional curriculum design for technological uncertainty, challenging traditional outcome-based educational models that assume predictable learning progressions. Our assessment structures deliberately rewarded students' exploration of unknown unknowns. They succeeded when they documented failed AI interactions, analysed unexpected outputs, and reflected on technological limitations. Our approach aligns with Darabi et al. (2018)'s findings that productive failure serves as an effective pedagogical strategy when properly structured and reflected upon. TLT and Rumsfeld's geopolitical epistemology provide a useful lens in thinking through pedagogical theory in an AI-enabled world. These frameworks provide guideposts on how we might transfer our competencies for AI use by ourselves and our students.

4.2 Implications for Educator Praxis

Students demonstrated competency transfer when analysing AI outputs using established techniques for textual criticism. For instance, Ancient History stream students evaluated AI-generated translations of Latin sources by applying standard historiographical methods: examining translation choices for bias, assessing accuracy against established scholarly sources, and contextualising outputs within broader historical interpretive contexts. Similarly, Philosophy stream students applied argumentative analysis techniques to AI-generated propaganda, identifying (and creating) logical fallacies, examining assumption structures, and evaluating persuasive mechanisms. Students realised they could use their discipline-specific competencies to engage with AI, figuring out how to prompt effectively and critique output using their well-established skills.

The transferrable competencies demonstrated by students also suggest that the cognitive load of adoption of these textual technologies is not necessarily as high as educators may imagine from the outside. Humanities educators already possess foundational skills applicable to AI pedagogical contexts. As Breen observes, LLMs are "deeply, inherently textual" and rely on capabilities "directly linked to the skills and methods that we emphasize in university humanities classes." Specifically, humanities training develops capacity to "analyze the genre, cultural context, assumptions, and affordances of a primary source" and understand "the unspoken limits that shaped how, why, and for whom it was created, and what content it contains" (Breen, 2023).

These scholastic skills transfer systematically to AI contexts: evaluating model outputs requires the same critical assessment of bias, contextual limitations, and production constraints that humanities educators routinely teach students to apply to historical documents, literary texts, and cultural artefacts. Our classroom observations first demonstrated this transfer when students successfully applied historical source criticism techniques to assess AI-generated content for reliability, utility, and appropriate use contexts.

Educators require targeted professional development to bridge existing competencies with AI-specific applications rather than comprehensive retraining. Our classroom implementation revealed a clear gap between methodological competence and practical AI integration capabilities. Educators need support in understanding differential model capabilities, developing context-appropriate prompting strategies, and recognising when AI tools enhance rather than compromise learning objectives. This professional development differs fundamentally from technical training. Educators need guided experience applying their existing skills to technological contexts, supported reflection on AI integration decisions, and collaborative development of discipline-specific AI applications. The development requirements focus on application translation: helping educators recognise how their established expertise applies to technological contexts rather than acquiring entirely new skill sets.

Resistance to AI integration can stem from perceived threats to professional expertise rather than competency limitations. Our classroom discussions revealed that students, educators, and administrators fear AI as a challenge to university authority and professional validity rather than recognising technological enhancement opportunities. This anxiety manifests in defensive approaches that position AI as a threat to be contained rather than a tool for amplifying existing capabilities. Sharan and Romano (2020) demonstrate that individual beliefs about personal agency significantly influence trust in AI systems (even predating generative AI), suggesting that educators' external AI-LOC may parallel students' initial technological relationships. Successful AI integration requires reframing: positioning AI as a textual technology that extends rather than replaces humanities skills.

4.3 Implications for the University Teaching Context

Effective institutional AI integration requires individual educator capacity development before cultural transformation. Our classroom experience demonstrates that successful AI pedagogy depends on educator preparedness as well as institutional support, a successful change-management strategy for technological adoption (Petko et al., 2015). Bottom-up and combined change models prove more sustainable than exclusively top-down technological adoption when educators understand how their existing competencies apply to new contexts (Molla & Nolan, 2020). Surface-level adoption creates technological burden rather than educational enhancement. This burden without support and without demonstration of utiltiy, tends to be rejected by overworked educators (Rosenberg, 2023). Space and time to develop individual capacity, pedagogical approaches for AI, and faculty norms around AI use must precede institutional culture change because technological tools require human expertise for effective educational integration.

Effective AI literacy requires access to frontier models and experimental latitude that most institutions resist providing, because frontier models empower capabilities that cheap models simply do not possess (Mollick, 2024a, 2024b). Our classroom culture also primed students to expect uncertainty, and to become co-creators of their own education. These material conditions enabled students to experience capability differences between models, experiment with prompting strategies, and develop critical evaluation skills. Institutions that limit AI access to basic models or prohibit experimental use cannot support sophisticated literacy development. Educational AI integration requires financial commitment: providing frontier model access, protecting professional development time, and creating policy space for pedagogical innovation.

Our success required explicit permission to experiment without predetermined outcomes, freedom to adapt curriculum based on monthly technological changes, and assessment approaches that valued learning processes over demonstrable competencies. A. Edmondson (1999) demonstrates that psychological safety enables learning behaviour in organisational contexts, and our classroom confirmed this principle applies to educational innovation.

Current educational quality norms cannot adequately evaluate AI-integrated learning experiences. Outcome-based assessment models assume predictable learning progressions and measurable competency development (Wiggins & McTighe, 2005). Our assessment approaches challenged these norms by evaluating students' capacity to navigate technological uncertainty, develop critical evaluation skills, and apply disciplinary knowledge in novel contexts. Traditional learning outcome structures struggle to capture agency development, epistemic humility, and adaptive thinking capabilities that AI literacy requires. Universities need new standards for evaluating pedagogical innovation: recognising process-focused assessment validity, supporting experimental approaches to learning measurement, and developing quality indicators appropriate for technological uncertainty contexts.

Individual classroom innovations cannot address broader structural barriers to educational transformation. Our experimental context provided exceptional conditions including research funding, institutional permissions, and protected experimental space that most educators lack access to. Isolated pedagogical experiments remain vulnerable to institutional resistance, policy constraints, and resource limitations that prevent systematic adoption. Henriksen et al. (2021) observe that educational risk-taking requires supportive institutional environments that many universities cannot or will not provide. We acknowledge these limitations while providing foundation for broader institutional conversations about AI integration. Systematic change requires coordinated attention to individual capacity development, resource allocation, cultural adaptation, and policy reform that extends beyond individual classroom innovations.

5 Conclusion

Our theoretical frameworks provided essential interpretive tools for understanding this new pedagogical territory. Transformative Learning Theory supports how AI education involves sustained cognitive disruption, while Rumsfeld's matrix maps the expanded epistemic landscape that technological uncertainty creates. Our pedagogical design operationalised these insights through four interconnected pillars: risk-embracing assessment structures that decoupled task success from grading outcomes, intentional cultural development through vulnerability and shared experimentation, systematic navigation of technological uncertainty as a learning dimension, and facilitation of transformative learning through structured disorienting dilemmas. These pillars created conditions where students could develop agency through productive failure while educators modelled epistemic humility. Rather than requiring entirely new pedagogical approaches, our findings suggest that textual analysis, contextual evaluation, and assumption examination transfer systematically to AI educational contexts.

We acknowledge the exceptional conditions of our experimental context: research permissions, frontier model access, and voluntary participation. This work represents exploratory investigation rather than prescriptive solution. Future research should examine systematic implementation across diverse institutional contexts and investigate the cultural change requirements for broader AI integration. This framework provides colleagues with theoretical foundations and practical strategies for engaging with AI-mediated educational transformation while recognising the ongoing adaptation that technological change demands.

AI Use Disclaimer

This paper used Claude 3.7 Sonnet and 4 Opus extensively, with discussion transcription by Gemini Pro 2.5 Pro 03-25 and with some literature searching, argumentation analysis, and copy-editing with OpenAI's GPT 4.5. Full prompt logs of every interaction are available at . The prompt log repository also contains all transcribed discussions Ballsun-Stanton and Torrington used as input to Claude 3.7 Sonnet for purposes of outline design, abstract generation, and writing. This paper was generated using AI auto-interview techniques Ballsun-Stanton is currently exploring to support another paper: *An Absence of Judgement: AI's Limitations in Deep Research tasks* by Ballsun-Stanton and Ross (preprint links forthcoming).

Acknowledgements

We would like to acknowledge the British Educational Research Association's British Journal of Educational Technology for the format of the practitioner notes section.

References

Baele, S. J., Bukhari, I., Whyte, C., Cuomo, S., Jensen, B., Payne, K., & Garcia, E. V. (2024). AI IR: Charting international relations in the age of artificial intelligence. *International Studies Review*, *26*(2). https://doi.org/10.1093/isr/viae013

Ballsun-Stanton, B., & Khalid, M. (2025, June 2). *The emperor's new clothes: A manifesto for universities in an AI-haunted world* [Publisher: Zenodo]. https://doi.org/10.5281/zenodo.15573395

Breen, B. (2023, September 12). *Simulating history with ChatGPT* [Res obscura]. Retrieved May 22, 2025, from https://resobscura.substack.com/p/simulating-history-with-chatgpt

Chaudhary, G. (2024). Unveiling the black box: Bringing algorithmic transparency to AI. *Masaryk University Journal of Law and Technology*, *18*(1), 93–122. https://doi.org/10.5817/MUJLT2024-1-4

Council on Foreign Relations. (2021, March 5). *Opposing communism in 1947* [CFR education from the council on foreign relations]. Retrieved May 20, 2025, from https://education.cfr.org/teach/mini-simulation/opposing-communism-1947

Daase, C., & Kessler, O. (2007). Knowns and unknowns in the 'war on terror': Uncertainty and the political construction of danger. *Security Dialogue*, *38*(4), 411–434. Retrieved May 5, 2025, from https://www.jstor.org/stable/26299636

Darabi, A., Arrington, T. L., & Sayilir, E. (2018). Learning from failure: A meta-analysis of the empirical studies. *Educational Technology Research and Development*, *66*(5), 1101–1118. https://doi.org/10.1007/s11423-018-9579-9

Dinsmore, D. L., Alexander, P. A., & Loughlin, S. M. (2008). Focusing the conceptual lens on metacognition, self-regulation, and self-regulated learning. *Educational Psychology Review*, *20*(4), 391–409. https://doi.org/10.1007/s10648-008-9083-6

Edmondson, A. (1999). Psychological safety and learning behavior in work teams. *Administrative Science Quarterly*, *44*(2), 350–383. https://doi.org/10.2307/2666999

Edmondson, A. C. (2018). *The fearless organization: Creating psychological safety in the workplace for learning, innovation, and growth*. John Wiley & Sons, Incorporated. Retrieved May 5, 2025, from http://ebookcentral.proquest.com/lib/mqu/detail.action?docID=5596894

Feng, X., Sundman, J., Aarnio, H., Taka, M., Keskinen, M., & Varis, O. (2025). Towards transformative learning: Students' disorienting dilemmas and coping strategies in interdisciplinary problem-based learning. *European Journal of Engineering Education*, *50*(2), 428–450. https://doi.org/10.1080/03043797.2024.2424197

Fleming, T. (2018). Critical theory and transformative learning: Rethinking the radical intent of mezirow's theory. *International Journal of Adult Vocational Education and Technology*, *9*(3), 1–13. https://doi.org/10.4018/IJAVET.2018070101

Gkintoni, E., Antonopoulou, H., Sortwell, A., & Halkiopoulos, C. (2025). Challenging cognitive load theory: The role of educational neuroscience and artificial intelligence in redefining learning efficacy. *Brain Sciences*, *15*(2), 203. https://doi.org/10.3390/brainsci15020203

Green, G., Rodgers, J., & Tainsh, C. (2024, October 28). Caligula's madness, an annotated bibliography: 1856–2024. https://doi.org/10.5281/zenodo.13999404

Henderson, M., Cosmos, A., Bangerter, M., Chen, M., D Souza, I., Fulcher, J., Halupka, V., Hook, J., Horton, C., Macfarlan, B., Mackay, R., Nagy, K., Schliephake, K., Trebilco, J., & Vu, T. (2022, June 10). Chapter 2: Educational design and productive failure: The need for a culture of creative risk taking [Section: Handbook of Digital Higher Education]. Retrieved May 27, 2025, from https://www.elgaronline.com/edcollchap/book/9781800888494/book-part-9781800888494-11.xml

Henriksen, D., Henderson, M., Creely, E., Carvalho, A. A., Cernochova, M., Dash, D., Davis, T., & Mishra, P. (2021). Creativity and risk-taking in teaching and learning settings: Insights from six international narratives. *International Journal of Educational Research Open*, *2*, 100024. https://doi.org/10.1016/j.ijedro.2020.100024

Hicks, B., & Kitto, K. (2025). Game theoretic models of intangible learning data. *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 970–976. https://doi.org/10.1145/3706468.3706557

Hoidn, S., & Kärkkäinen, K. (2014, January). *Promoting skills for innovation in higher education : A literature review on the efectiveness of problem-based learning and of teaching behaviours* (Technical Report) (Accepted: 3/25/2014 9:43). OECD. Retrieved May 15, 2025, from https://repositorio.minedu.gob.pe/handle/20.500.12799/2482

Huang, Q., Willems, T., Kaur, A., Poon, K. W., Samarakoon, B., & Elara, M. R. (2023). A pedagogical approach of "learning from failure" for engineering students: Observation and reflection on a robotics competition (RoboRoarZ-edition 2). *2023 IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE)*, 1–5. https://doi.org/10.1109/TALE56641.2023.10398407

Khalaf, B. K., & Zin, Z. B. M. (2018). Traditional and inquiry-based learning pedagogy: A systematic critical review. *International Journal of Instruction*, *11*(4), 545–564. https://doi.org/10.12973/iji.2018.11434a

Krathwohl, D. R. (2002). A revision of bloom's taxonomy: An overview. *Theory Into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Kuhn, T. S. (1994). *The structure of scientific revolutions* (2. ed., enlarged, 21. print). Univ. of Chicago Press.

Laros, A. (2017). Disorienting dilemmas as a catalyst for transformative learning. In A. Laros, T. Fuhr, & E. W. Taylor (Eds.), *Transformative learning meets bildung* (pp. 85–95). SensePublishers. https://doi.org/10.1007/978-94-6300-797-9_7

Lave, J. (1991). Situating learning in communities of practice. In L. B. Resnick, J. M. Levine, & S. D. Teasley (Eds.), *Perspectives on socially shared cognition.* (pp. 63–82). American Psychological Association. https://doi.org/10.1037/10096-003

Liu, D. Y. T., & Bates, S. (2025, January). *Generative AI in higher education: Current practices and ways forward* (Whitepaper). Association of Pacific Rim Universities. Retrieved May 15, 2025, from https://www.apru.org/wp-content/uploads/2025/01/APRU-Generative-AI-in-Higher-Education-Whitepaper_Jan-2025.pdf

Lodge, J. M., Howard, S., Bearman, M., Dawson, P., & Agostinho, S. (2023, November). *Assessment reform for the age of artificial intelligence* (TEQSA Resource). Tertiary Education Quality and Standards Agency. Melbourne, Australia. Retrieved December 15, 2023, from https://www.teqsa.gov.au/sites/default/files/2023-09/assessment-reform-age-artificial-intelligence-discussion-paper.pdf

Manz, E., & Suárez, E. (2018). Supporting teachers to negotiate uncertainty for science, students, and teaching [_eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/sce.21343]. *Science Education*, *102*(4), 771–795. https://doi.org/10.1002/sce.21343

Margetson, D. (1994). Current educational reform and the significance of problem-based learning. *Studies in Higher Education*, *19*(1), 5–19. https://doi.org/10.1080/03075079412331382103

Mezirow, J. (1978). Perspective transformation. *Adult Education*, *28*(2), 100–110. https://doi.org/10.1177/074171367802800202

Mezirow, J. (1991). *Transformative dimensions of adult learning* (1st ed). Jossey-Bass.

Mezirow, J. (1997). Transformative learning: Theory to practice. *New Directions for Adult and Continuing Education*, *1997*(74), 5–12. https://doi.org/10.1002/ace.7401

Mezirow, J. (2000). Learning to think like an adult: Core concepts of transformation theory [Publisher: Jossey-Bass]. *Learning as Transformation : Critical Perspectives on a Theory in Progress.*

Molla, T., & Nolan, A. (2020). Teacher agency and professional practice [Publisher: Routledge _eprint: https://doi.org/10.1080/13540602.2020.1740196]. *Teachers and Teaching*, *26*(1), 67–87. https://doi.org/10.1080/13540602.2020.1740196

Mollick, E. (2023, August 20). *Now is the time for grimoires.* Retrieved May 19, 2025, from https://www.oneusefulthing.org/p/now-is-the-time-for-grimoires

Mollick, E. (2024a). *Co-intelligence: Living and working with AI.* WH Allen.

Mollick, E. (2024b, December). *What just happened.* Retrieved May 22, 2025, from https://www.oneusefulthing.org/p/what-just-happened

Moström Åberg, M. (2023). Contextual preconditions to foster transformative learning: A recursive process, activity, and core elements [Publisher: SAGE

Publications]. *Journal of Transformative Education*, *21*(2), 167–189. https://doi.org/10.1177/15413446221091769

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Pawson, R., Wong, G., & Owen, L. (2011). Known knowns, known unknowns, unknown unknowns: The predicament of evidence-based policy [Publisher: SAGE Publications Inc]. *American Journal of Evaluation*, *32*(4), 518–546. https://doi.org/10.1177/1098214011403831

Petko, D., Egger, N., Cantieni, A., & Wespi, B. (2015). Digital media adoption in schools: Bottom-up, top-down, complementary or optional? *Computers & Education*, *84*, 49–61. https://doi.org/10.1016/j.compedu.2014.12.019

Picasso, F., Atenas, J., Havemann, L., & Serbati, A. (2024). Advancing critical data and AI literacies through authentic and real-world assessment design using a data justice approach. *Open Praxis*, *16*(3), 291–310. https://doi.org/10.55982/openpraxis.16.3.667

Rosenberg, B. (2023). *"whatever it is, i'm against it": Resistance to change in higher education*. Harvard Education Press.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement [Place: US Publisher: American Psychological Association]. *Psychological Monographs: General and Applied*, *80*(1), 1–28. https://doi.org/10.1037/h0092976

Rumsfeld, D. (2002, February 12). *Defense department briefing, february 12*. Retrieved May 5, 2025, from https://usinfo.org/wf-archive/2002/020212/epf202.htm

Sharan, N. N., & Romano, D. M. (2020). The effects of personality and locus of control on trust in humans versus artificial intelligence [Publisher: Elsevier]. *Heliyon*, *6*(8). https://doi.org/10.1016/j.heliyon.2020.e04572

Smith, M. (2020). Educating risk: How fear of failure is stifling creative practice within higher education [Publisher: Manchester Animation Festival and University of Salford]. *EDUCATING ANIMATORS Academic Conference 2019 Teaching the World's most expressive Art Form*. Retrieved May 27, 2025, from https://lau.repository.guildhe.ac.uk/id/eprint/17684/

Teng, M. F., & Yue, M. (2023). Metacognitive writing strategies, critical thinking skills, and academic writing performance: A structural equation modeling approach. *Metacognition and Learning*, *18*(1), 237–260. https://doi.org/10.1007/s11409-022-09328-5

Torrington, J., Ballsun-Stanton, B., & Lai, J. (2025, February 21). Teaching students how to effectively interact with LLMs at university: Insights on the longitudinal development and plasticity of locus of control. https://doi.org/10.35542/osf.io/6mke5_v1

Torrington, J., Bower, M., & Burns, E. C. (2023). What self-regulation strategies do elementary students utilize while learning online? *Education and Information Technologies*, *28*(2), 1735–1762. https://doi.org/10.1007/s10639-022-11244-9

UNESCO. (2023). *Guidance for generative AI in education and research*. https://doi.org/10.54675/EWZM9535

Wiggins, G. P., & McTighe, J. (2005). *Understanding by design* [Google-Books-ID: N2EfKlyUN4QC]. ASCD.

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview [Publisher: Routledge _eprint: https://doi.org/10.1207/s15430421tip4102_2]. *Theory Into Practice*, *41*(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2