ASCILITE 2026

The Future is Connection:

Collaboration, Creativity and Compassion in a Digital Society

From short-term gains to deep learning: Designing AI learning agents through productive failure

Polly K. Lai
Southern Cross University and Western Sydney University

Vinh Bui, Thanh Nguyen
Southern Cross University

This conceptual paper proposes a conceptual framework for designing Productive Failure-informed AI learning agents that function as pedagogical facilitators rather than answer-providing tools. Grounded in Productive Failure (PF), the framework structures AI-supported learning across three interconnected phases: idea generation and exploration, consolidation, and knowledge assembly. Within each phase, it specifies the role of the AI facilitator, the structure of AI–learner interaction, and the cognitive processes intended to support learning. The framework positions AI agents as mediators that activate prior knowledge, encourage multiple solution attempts, support comparison with canonical knowledge, promote recognition of knowledge gaps, and guide the organisation of emerging understanding into generalisable principles. Theoretically, the framework extends PF into AI-mediated learning contexts by translating its underlying mechanisms into interaction design principles. Practically, it provides educators and developers with a pedagogically grounded approach to designing AI-supported environments that prioritise cognitive engagement, conceptual understanding, and knowledge transfer. Future research should empirically examine the framework across disciplinary, learner, and institutional contexts.

Keywords: Generative AI, productive failure, AI learning agent, learning design, conceptual paper

Introduction

Generative artificial intelligence (GenAI) is rapidly reshaping higher education. Its capacity to generate human-like text, respond to open-ended questions, and simulate expert dialogue creates opportunities for personalised feedback, learner support, and self-directed learning. Students increasingly use GenAI to clarify concepts, generate ideas, summarise information, draft assessments, conduct research, explore complex topics, and seek feedback (Chan & Hu, 2023). Although these applications may improve task efficiency and immediate performance, current GenAI tools are not always aligned with pedagogical goals or designed to promote enduring learning outcomes (Bauer et al., 2025).

Recent research has examined whether GenAI enhances learning or, conversely, undermines creativity and critical thinking (Mei et al., 2025). However, comparatively less attention has been given to how GenAI systems can be designed to support deep, durable, and transferable learning rather than short-term task completion. Addressing this gap requires moving beyond the prevailing view of GenAI as primarily a content generator or productivity tool. Instead, GenAI can be conceptualised as an AI learning agent that plans and undertakes pedagogically purposeful actions to help learners achieve defined learning goals. Through carefully structured AI–learner interactions, such agents may support idea generation, knowledge construction, reflection, and metacognitive regulation (Ouyang & Jiao, 2021). This form of human–AI collaboration may foster co-intelligence by combining the complementary capabilities of learners and AI systems (Doroudi, 2023).

This conceptual paper proposes a framework for designing AI learning agents in higher education, grounded in Productive Failure (PF) (Kapur, 2008; Kapur & Bielaczyc, 2012). PF suggests that learners can deepen conceptual understanding when they first engage with complex problems and generate initial solutions before receiving structured instruction and consolidation (Cao et al., 2023; Lai et al., 2017). These processes can help

learners activate prior knowledge, recognise gaps and inconsistencies in their understanding, and reorganise emerging ideas into more coherent knowledge structures (Kapur, 2016).

Building on these principles, the proposed framework positions AI agents not as answer-providing tools, but as pedagogical facilitators and learning partners within intentionally designed learning environments. It specifies how AI agents can support learners across the phases of idea generation and exploration, consolidation, and knowledge assembly. In doing so, the paper shifts attention from technological novelty to pedagogical intentionality and offers a theoretically grounded basis for designing AI-supported interactions that prioritise deep and transferable learning over immediate task performance.

Productive Failure as a theoretical foundation

Central to this pedagogical shift is the need to consider how AI tools structure and guide learning interactions, particularly through questions and prompts that encourage cognitive engagement rather than merely providing efficient or correct answers. This emphasis reflects the distinction between performance and learning. Although AI tools may improve immediate task performance, such improvements do not necessarily result in deeper understanding or the ability to transfer knowledge to new contexts (Soderstrom & Bjork, 2015). Addressing this distinction requires a design perspective that deliberately creates opportunities for productive cognitive struggle, as conceptualised in the PF framework (Kapur, 2008).

PF is based on the premise that learners can develop deeper conceptual understanding when they first attempt to generate solutions to complex problems before receiving structured instruction. Three interrelated mechanisms help explain how this process supports learning. First, PF activates prior knowledge and promotes generative processing. Prior knowledge is organised into schemas that shape how learners interpret problems and integrate new information (Anderson, 1983). When learners encounter a complex problem, they draw on these existing knowledge structures to construct initial representations of the problem space and generate possible solutions. These attempts may be incomplete or incorrect; however, they make learners' existing ideas, assumptions, and reasoning processes available for examination. In this sense, prior knowledge provides the basis for learners' internal generative models and enables meaningful engagement with the problem. Without its activation, learners have limited resources from which to generate, compare, and revise potential solutions.

Second, learning is supported through the recognition and resolution of knowledge gaps. Although prior knowledge enables learners to generate initial predictions and solution attempts, conceptual development is driven by their awareness of discrepancies between these attempts and the requirements of the task. When such discrepancies remain unrecognised, learners may retain incomplete or inaccurate mental models. By contrast, awareness of impasses, contradictions, or inconsistencies can prompt learners to explain failures, reconsider assumptions, and adopt strategies for revising their understanding (Chi, 2000). Comparisons among learners' solutions, alternative approaches, and canonical explanations can therefore help make critical knowledge gaps visible and create a need to acquire or reorganise knowledge.

Third, deep learning requires the organisation and integration of knowledge. Learning does not result from the correction of errors alone; it requires learners to reorganise existing knowledge structures in ways that accommodate new conceptual understanding (Kapur & Bielaczyc, 2012). Within PF, this process involves differentiating productive from unproductive features of learners' initial solutions and integrating fragmented or partially correct ideas into more coherent representations. This mechanism is consistent with conceptual change perspectives and with accounts of knowledge as comprising context-sensitive elements that must be progressively coordinated and organised (diSessa, 2014). Such knowledge assembly is essential for enabling learners to recognise underlying principles and apply them across contexts.

Together, these mechanisms constitute a coherent yet dynamic learning architecture. Learning begins with the activation of prior knowledge and the generation of initial solution attempts, progresses through the identification and examination of knowledge gaps, and culminates in the reorganisation and integration of knowledge. This process is not strictly linear; learners may move iteratively between generating, comparing,

reflecting, and revising as their understanding develops. Moreover, the quality of engagement at each stage influences the learning opportunities available at subsequent stages.

Viewed through this lens, PF offers more than an account of how learning occurs. Its learning mechanisms provide a theoretical foundation for determining when and how AI agents should intervene in the learning process. Rather than immediately resolving difficulties or supplying correct answers, AI agents can elicit learners' prior knowledge, support the generation and comparison of multiple solutions, make inconsistencies visible, and facilitate reflection and knowledge integration. Building on these principles, the following section develops a conceptual framework for designing a PF-informed AI learning agent.

A conceptual framework for designing a PF-informed AI learning agent

The framework conceptualises AI not as an answer-providing tool, but as a facilitator that structures learning interactions to promote generative thinking, comparison and consolidation, and knowledge assembly. It is organised around three phases of PF: (1) idea generation and exploration, (2) consolidation, and (3) knowledge assembly. Within each phase, the framework specifies three interconnected dimensions: the pedagogical role of the AI facilitator, the structure of AI–learner interaction, and the cognitive processes that these interactions are intended to activate. The overall framework is presented in Figure 1.

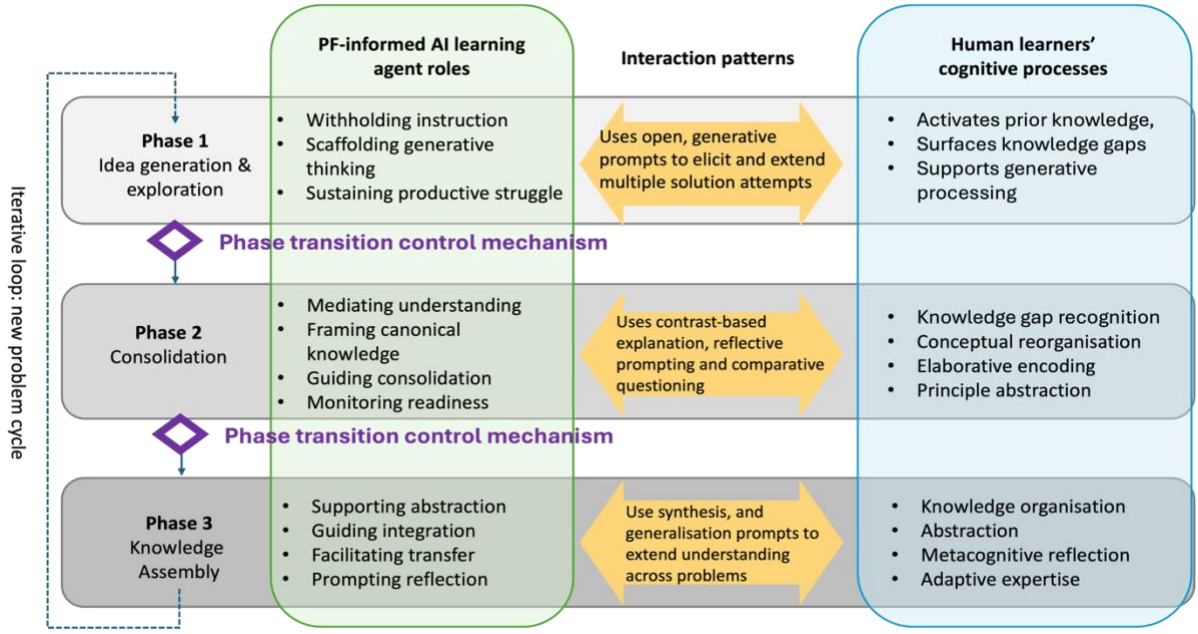


Figure 1. A conceptual framework of PF-informed AI learning agents across three phases with transition decisions guiding progression, continuation, or a new problem cycle

Phase 1: Idea generation and exploration

In the first phase, learners engage with complex, ill-structured problems for which solution pathways are not immediately apparent. The problem is introduced by the educator, while learners approach it with varying levels of prior knowledge, intuitive understandings, and potentially unarticulated misconceptions. The role of the PF-informed AI learning agent is not to explain the problem, provide solution procedures, or direct learners towards a correct answer. Instead, it seeks to activate learners' existing knowledge and support the generation of multiple solution attempts before explicit instruction is provided.

The deliberate withholding of explanations and correct answers represents a key departure from the conventional use of GenAI, which often prioritises rapid task completion and response accuracy. Within a PF approach, premature assistance may reduce opportunities for learners to explore the problem space, identify the limitations of their existing understanding, and develop a need for subsequent instruction. The AI agent

must therefore provide sufficient support to sustain productive engagement without removing the cognitive challenge that gives rise to learning.

AI and learner interaction during this phase is structured through open-ended and generative prompts that encourage learners to externalise and elaborate their thinking rather than converge prematurely on a single solution. Learners are invited to propose initial approaches, articulate their assumptions, explain their reasoning, and explore alternative possibilities, even when their ideas are incomplete or incorrect. The AI agent should transition to the next phase when the learner has demonstrated sufficient exploration of the problem. This occurs when the learner has provided at least two meaningfully distinct representations of solution methods (RSMs) and has explained the reasoning and assumptions associated with them. The transition would be based on evidence that the learner has considered multiple approaches.

These interactions are intended to activate prior knowledge, promote generative processing, and sustain divergent exploration. By eliciting multiple RSMs, the AI agent helps make learners' existing conceptions visible and available for subsequent comparison and refinement (Kapur & Bielaczyc, 2012). The goal of this phase is therefore not the production of a correct solution, but the creation of a rich set of learner-generated ideas that can serve as resources for consolidation and knowledge development in later phases.

Phase 2: Consolidation

In the consolidation phase, the focus shifts from generating ideas to organising, comparing, and refining learners' solution approaches. Central to this phase is the structured comparison of learner-generated and canonical solutions, which helps learners identify key conceptual features, recognise limitations in their initial reasoning, and develop more coherent understanding. Having generated multiple attempts and encountered the boundaries of their prior knowledge, learners are positioned to engage more meaningfully with formal concepts and strategies.

The role of the PF-informed AI learning agent correspondingly shifts from facilitating exploration to mediating between learners' initial ideas and canonical knowledge. This mediation remains deliberately non-evaluative. Rather than simply presenting the correct answer, the agent frames canonical solutions as extensions, refinements, or reorganisations of learners' existing thinking. In doing so, it preserves learner agency while directing attention to conceptual gaps, productive elements, and points of divergence.

The AI agent should transition to the next phase only after it has explained the canonical knowledge, explicitly compared it with the learner's previous solution methods, and supported the learner in recognising how their ideas relate to the canonical approach. Before progressing, the agent should obtain evidence that the learner has sufficiently consolidated the new knowledge. This occurs when the learner can explain the key principles of the canonical approach in their own words and identify important similarities and differences between the canonical approach and their initial attempts. A key pattern is contrast-based explanation, in which the agent juxtaposes learner-generated approaches with canonical solutions to highlight similarities, omissions, and critical distinctions (Jacobson et al., 2020). These interactions support learners in revising their mental models and developing more integrated conceptual understanding.

Phase 3: Knowledge assembly

In the knowledge assembly phase, the emphasis shifts from understanding a specific solution to constructing more generalisable and transferable knowledge structures. Learners integrate their refined ideas with canonical strategies to identify underlying principles and develop understanding that can be applied across contexts (Loibl et al., 2017). Having recognised the limitations of their initial approaches and consolidated key concepts, they are positioned to move beyond context-specific reasoning towards more abstract and flexible knowledge representations (Jacobson et al., 2020).

The role of the PF-informed AI learning agent in this phase is to support abstraction, integration, and knowledge organisation while maintaining a non-evaluative stance. Rather than introducing new content or supplying direct solutions, the agent guides learners in assembling their emerging understanding into coherent

and principled forms. This involves helping them connect their initial attempts, consolidated insights, and canonical methods, and supporting the transition from specific cases to broader conceptual principles.

The agent also facilitates reflective closure by prompting learners to examine the trajectory of their learning, from initial exploration and comparison to conceptual refinement and abstraction. This process reinforces the value of productive struggle and prepares learners to apply their understanding in future learning contexts. Interaction patterns in this phase therefore prioritise synthesis, generalisation, transfer, and forward-looking reflection.

Conclusion and future directions

This paper proposes a PF-informed framework for designing AI learning agents that support deep learning in higher education. It reconceptualises AI as a learning facilitator and structures interaction across three phases: idea generation, consolidation, and knowledge assembly. The framework contributes theoretically by extending PF to AI-mediated learning and practically by offering design principles that prioritise cognitive engagement over immediate performance. However, it remains conceptual and requires empirical validation across disciplines and contexts. Future research should examine its effects on learning and implementation.

References

Anderson, J. R. (1983). *The architecture of cognition*. Harvard University Press.

Bauer, E., Greiff, S., Graesser, A. C., Scheiter, K., & Sailer, M. (2025). Looking beyond the hype: Understanding the effects of AI on learning. *Educational Psychology Review*, *37*(2), 45. https://doi.org/10.1007/s10648-025-10020-8

Cao, L., Lai, P. K., & Yang, H. (2023). Using productive failure to learn genetics in a game-based environment. *Instructional Science* 1–21. https://doi.org/10.1007/s11251-023-09644-6

Chan, C. K. Y., & Hu, W. (2023). Students' voices on generative AI: Perceptions, benefits, and challenges in higher education. *International Journal of Educational Technology in Higher Education*, *20*(1), 43. https://doi.org/10.1186/s41239-023-00411-8

Chi, M. T. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. In R. Glaser (Ed.), *Advances in instructional Psychology* (Vol. 5). Routledge. https://doi.org/10.4324/9780203774526

diSessa, A. A. (2014). A history of conceptual change research. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (pp. 88–108). Cambridge University Press. https://doi.org/10.1017/CBO9781139519526.007

Doroudi, S. (2023). What happened to the interdisciplinary study of learning in humans and machines? *Journal of the Learning Sciences*, *32*(4-5), 663–681. https://doi.org/10.1080/10508406.2023.2260159

Jacobson, M. J., Goldwater, M., Markauskaite, L., Lai, P. K., Kapur, M., Roberts, G., & Hilton, C. (2020). Schema abstraction with productive failure and analogical comparison: Learning designs for far across domain transfer. *Learning and Instruction*, *65*. https://doi.org/doi.org/10.1016/j.learninstruc.2019.101222

Kapur, M. (2008). Productive Failure. *Cognition and Instruction*, *26*(3), 379–424. https://doi.org/10.1080/07370000802212669

Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist*, *51*(2), 289–299. https://doi.org/10.1080/00461520.2016.1155457

Kapur, M., & Bielaczyc, K. (2012). Designing for productive failure. *Journal of the Learning Sciences*, *21*(1), 45–83. https://doi.org/10.1080/10508406.2011.591717

Lai, P. K., Portolese, A., & Jacobson, M. J. (2017). Does sequence matter? Productive failure and designing online authentic learning for process engineering. *British Journal of Educational Technology*, *48*(6), 1217–1227. https://doi.org/10.1111/bjet.12492

Loibl, K., Roll, I., & Rummel, N. (2017). Towards a theory of when and how problem solving followed by instruction supports learning. *Educational Psychology Review*, *29*(4), 693–715. https://doi.org/10.1007/s10648-016-9379-x

Mei, P., Brewis, D. N., Nwaiwu, F., Sumanathilaka, D., Alva-Manchego, F., & Demaree-Cotton, J. (2025). If ChatGPT can do it, where is my creativity? Generative AI boosts performance but diminishes

experience in creative writing. *Computers in Human Behavior: Artificial Humans*, *4*, 100140. https://doi.org/https://doi.org/10.1016/j.chbah.2025.100140

Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence*, *2*, 100020. https://doi.org/10.1016/j.caeai.2021.100020

Soderstrom, N. C., & Bjork, R. A. (2015). Learning versus performance: An integrative review. *Perspectives on Psychological Science*, *10*(2), 176–199. https://doi.org/10.1177/17456916155690

Lai, P. K., Bui, V., & Nguyen, D. T. (2026). From short-term gains to deep learning: Designing AI learning agents through productive failure. In Cochrane, T., Narayan, V., Poole, T., Lai, P., Lake, P., Moore, O., & Li, L. (Eds.), *The Future is Connection: Collaboration, Creativity and Compassion in a Digital Society*. Proceedings ASCILITE 2026. Brisbane (pp. 683–688). https://doi.org/10.65106/apubs.2026.3311

Note: All published papers are refereed, having undergone a double-blind peer-review process.
The author(s) assign a Creative Commons by attribution license enabling others to distribute, remix, tweak, and build upon their work, even commercially, as long as credit is given to the author(s) for the original creation.

© Lai, P. K., Bui, V., & Nguyen, D. T. 2026