CriSTaL
CRITICAL STUDIES IN TEACHING & LEARNING

2026, Volume 14, Issue 1

The feedback box:
What is educational in the age of artificial intelligence?

Wayne Hugo
University of KwaZulu-Natal
Corresponding author: hugow@ukzn.ac.za

(Submitted: 9 February 2026; Accepted: 1 June 2026)

Abstract

Feedback is a neutral mechanism. The same response loop drives laboratory animal conditioning, commercial digital platforms engineered to capture attention, and classroom instruction. Because it is indifferent to its outcome, feedback does not automatically educate; it can just as easily condition or manipulate. Drawing on R.S. Peters, this paper argues that a feedback system becomes educational only when it meets a demanding set of conditions. It must engage the learner intentionally, with the learner conscious of what they are trying to master. Its methods must be morally defensible, ruling out coercion, manipulation, and indoctrination as means. It must aim at knowledge and understanding rather than only trained performance, giving learners the reason why things are as they are. What is learned must transform the learner's judgement and outlook rather than remaining inert. It must support the all-round development of the person rather than narrowing them to instrumental function and it must aim at autonomy, so that the scaffolding can eventually be withdrawn and the learner can think, judge, and act for themselves. These conditions are under direct pressure from contemporary AI systems, which can capture attention through addictive design and replace reasoned feedback with token economies that reward compliance rather than understanding. The school classroom is in a stronger position against these pressures than is usually recognised, because its boundaries are visible, its standards are publicly contestable, and its feedback is anchored in accountable human interaction that cannot easily be displaced by generative output. Higher education is far more exposed, precisely because of the freedoms that have always defined it: students working alone, constructing their own paths through reading and inquiry, and submitting artefacts that generative AI can now produce directly. The task is therefore not to adopt or reject AI but to configure feedback systems, in classrooms and university courses alike, so they remain answerable to educational purposes rather than to whatever the technology most easily affords.

Keywords: Artificial intelligence, backpropagation, classroom, educational, feedback, feedback box, manipulation, reinforcement

Opening the box

In 1898, Edward Thorndike placed a hungry cat in a wooden crate. The crate was rough, with slats through which the cat could see and smell food placed just outside. Inside, a loop of string hung from the ceiling; if the cat pulled the string, the door would open. The cat did not examine the mechanism, form a hypothesis, and test it; it thrashed. It bit at the slats, clawed at the walls, and pushed its paws through gaps. Amid this chaos, it eventually, by accident, pulled the string. The door opened, and the cat ate. On successive trials, these useless behaviours gradually dropped away, and the successful action emerged more quickly. Thorndike plotted the results as a smooth curve indicating the incremental strengthening of a connection between a situation and a response.

From these observations he derived the Law of Effect: actions followed by satisfying consequences are stamped in and become more likely, while those followed by discomfort are stamped out (Thorndike, 1898: 1911). The claim was radical because it demonstrated that learning could be explained mechanistically, without reference to understanding, reasoning, or consciousness.

The puzzle box was a methodological reduction, stripping learning to its barest measurable elements: a situation, a response, and a consequence. In doing so it also isolated the loop from everything that makes learning genuinely educational in a richer sense like the reasons behind an action, the understanding that transfers, the teacher–learner relationship, and the learner's own autonomy. The puzzle box was never designed to educate the cat.

What Thorndike released into the wild was a "neutral" mechanism. Under the formalisation of Sutton and Barto (2018), reinforcement learning has become one of the most successful paradigms in machine learning, producing systems that master games, discover strategies, and shape what billions of people encounter on screens. The same principles Thorndike identified in cats, and that B.F. Skinner refined in pigeons, now operate within huge data centres where human behaviour is tracked and optimised for profit.

This paper shows how these identical feedback mechanisms can be configured for education rather than manipulation. I use the term feedback box for any bounded arrangement in which a learner's actions are repeatedly met with designed signals, corrections, consequences, and explanations that shape subsequent action over time. The term is intentionally value-neutral. The definition lets Thorndike's puzzle box, Skinner's operant chamber, Montessori's materials, Pressey's teaching machine, a slot machine, a social media feed, a language-learning app, and an AI agent sit under the same genus. They are all, fundamentally, feedback boxes. What differs is the configuration: who controls the box, what it optimises for, whether reasons are in view, and whether the trajectory leads toward independent autonomy or perpetual dependence.

The mechanism isolated

Evolutionary feedback

Long before its deliberate use in teaching, feedback was already a basic principle in biological and mechanical systems, where it governed regulation and adaptation (Wiener, 1961; Ashby, 1956). Darwin's theory of natural selection describes a powerful feedback loop (Darwin, 1859/2003). Environmental pressures function as a selection mechanism across generations. Organisms with traits better suited for survival and reproduction pass those traits on, gradually shaping the population. But natural selection is a feedback loop without an agent. No one designed it, no one controls it, no one benefits from it by intention. Its feedback mechanism is survival or death.

Skinner's systematisation

Where Thorndike had used rough wooden crates, Skinner designed a standardised apparatus: a small chamber with a lever the animal could press and a mechanism for delivering food pellets (Skinner, 1938). This operant conditioning chamber, which others called the Skinner box though Skinner himself disliked the term (Latour, 1999), became the standard instrument for studying behaviour–consequence relationships.

Skinner stripped the mechanism of mentalistic language. A reinforcer is simply anything that increases the probability of the behaviour it follows (Skinner, 1953: 72–75). This made the mechanism portable. Working with Ferster, Skinner found that the pattern and timing of reinforcement matter as much as the reinforcement itself (Ferster & Skinner, 1957). The variable ratio schedule is key for this paper. When reinforcement comes after an unpredictable number of responses, the organism responds at a persistently high rate and the behaviour resists extinction long after reinforcement stops. Skinner noted that this is precisely how slot machines work (Skinner, 1953: 104). The observation was descriptive. The gambling industry would later treat it as a design specification.

Skinner also developed shaping: building complex behaviours by reinforcing successive approximations to a target. A pigeon could be trained to turn in a circle by first reinforcing any movement to the left, then quarter-turns, then half-turns, then full turns (Skinner, 1953: 91–95).

Power and danger

The Skinner box, like Thorndike's puzzle box, was a methodological reduction. By stripping away competing stimuli, it made the relationship between behaviour and consequence maximally clear, allowing researchers to discover genuine regularities. But what was studied was not education. The rat pressing a lever cannot say why pressing it produces food, cannot transfer that understanding, and cannot decide that lever-pressing is not, on reflection, a worthwhile activity. It has been conditioned, not educated.

So, the same mechanism that explains how organisms learn also explains how they can be manipulated. Variable ratio reinforcement produces persistent, compulsive responding in rats, and will produce similar effects in humans. Shaping can construct arbitrary behaviours in pigeons, and can influence the behaviour of children, consumers, and citizens. The mechanism is indifferent to whether it serves the interests of the organism or the interests of whoever controls the reinforcement.

Educational feedback boxes

The history of education contains feedback boxes that look nothing like Skinner's chamber, even though they share the underlying loop. Three traditions are briefly described to bring out what they have in common.

Montessori's self-correcting materials

Maria Montessori developed a method centred on specially designed materials that children manipulate in structured environments (Montessori, 1912, 1914). Many are stored in and unpacked from specific boxes or trays and, when played with, provide immediate, intrinsic feedback that the child can interpret without adult intervention.

My personal favourite is the wooden box that contains ten cylindrical holes of graduated sizes, varying in diameter, or height, or both. Corresponding wooden cylinders fit precisely into the holes, one cylinder to each hole. The child's task is to match cylinders to holes. When a cylinder is placed in the correct hole, it fits flush. When placed in the wrong hole, it either protrudes or rattles loosely, or a later cylinder has no home (Montessori, 1912: 167–170).

Notice what is absent. There is no external reinforcer, no candy, no points, no praise from an adult for correct performance. The feedback comes from the material itself, from the physical properties of the objects. The cylinder that does not fit is not an arbitrary punishment; it indicates a mismatch between the learner's action and the structure of reality. The child can see why this cylinder does not go here. It is too big, or too small, or the wrong shape. The feedback explains.

Montessori designed her materials to isolate specific concepts, what she called the 'isolation of difficulty' (Montessori, 1914: 80). The cylinder blocks isolate visual discrimination of dimension. The pink tower isolates graduated size in three dimensions. The colour tablets isolate chromatic discrimination. Each material presents a single conceptual challenge, with feedback built into the activity rather than delivered by an external agent. And the self-correcting nature of the materials means the child does not depend on adult approval. Over time the child internalises criteria for correctness and can work independently. The materials are designed to make themselves unnecessary: once the perceptual and conceptual capacities are developed, the particular cylinders are no longer needed.

Teaching machines

Sidney Pressey developed what is often regarded as the first automated teaching machine in the 1920s, presenting questions and providing immediate confirmation or correction (Pressey, 1926, 1927). The device enforced a tight loop: response, feedback, adjustment, and continued work. Skinner later framed teaching machines as programmed instruction, with small sequenced frames that require active responding and rapid feedback understood as reinforcement (Skinner, 1958, 1968). These machines sit between the operant chamber and classroom practice. They can train reliable responses without understanding, or they can support learning when the feedback explains and the programme has a genuine endpoint of competence rather than indefinite engagement (Benjamin, 1988). Power and danger lie in the same mechanism.

Formative assessment

Formative assessment names classroom practices that use feedback to improve learning rather than record performance (Black and Wiliam, 1998; Black, et al., 2003). Sadler specified the mechanism: to improve, learners must grasp the standard or goal, receive information that enables comparison between current performance and that standard, and be able to act to close the gap (Sadler, 1989). Feedback functions as information that supports self-regulation. Learners should come to hold a conception of quality roughly similar to that held by the teacher so they can monitor and evaluate their own work, an observation Vygotsky would have been proud of. Hattie and Timperley sharpen this by distinguishing task-level correctness, process-level guidance, and self-regulation support (Hattie & Timperley, 2007), with the consistent implication that feedback confined to verdicts or praise tends to be weaker, while feedback that addresses processes and self-regulation more reliably supports learning because it can be understood and acted on.

The shared configuration

Montessori's materials, the early teaching machines, and the formative assessment tradition differ in their settings and their methods, but they share two features that distinguish them from Skinner's chamber. First, the feedback teaches and explains rather than just marks. The learner can sense what is being evaluated and why. Second, the trajectory aims at independence: the scaffolding is designed to be withdrawn as competence grows. R.S. Peters worked out what makes these features educational rather than just effective.

Criteria for the educational feedback box

Peters argued, at length and against serious opponents, that education is not a neutral label for any effective learning. It is a normative concept with built-in standards for what counts as being educated, and those standards rule out certain processes even when they are efficient. When we move from classrooms to optimised feedback systems, the temptation is to treat learning as whatever produces behavioural change or improved performance on a metric. Peters gives us a disciplined way to resist that slide, because he draws a line between education and nearby practices such as training, conditioning, and indoctrination (Peters, 1966; Hirst & Peters, 1970). His position has been criticised and reworked (Biesta, 2009; White & White, 2022; Winch, 2017), but his central insistence on intentionality, understanding, and autonomy remains a strong test for any feedback box, especially where manipulation is a live possibility.

Peters's normative criteria

The core criteria can be stated in a sharpened form. Education requires intentional engagement with what is to be learned: the learner 'must know what he is doing, must be conscious of something that he is trying to master' (Peters, 2010: 7). Its methods must be morally defensible and respect the learner as a person; coercion, manipulation, and indoctrination are excluded as means (Peters, 2010: 2–3). It aims at knowledge and understanding, not only trained performance, and involves grasping principles and the 'reason why' of things (Peters, 2010: 4). What is learned should not remain inert; it should shape the learner's judgement and outlook (Peters, 2010: 5). Education supports all-round development rather than narrowing the learner to a limited instrumental function. External rewards and punishments may be used as aids, but education cannot be built on them because their link to the learning is purely extrinsic rather than intrinsic (Peters, 2010: 7).

Taken together, these criteria define a specific conception of a learning environment, one grounded in ethical practice, respect for autonomy, and the aim of fostering deep, transformative understanding that is intrinsically rewarding. This ideal stands in sharp contrast to algorithmic systems whose feedback loops prioritise user engagement, data extraction, or profit (Hyland, 2025).

Structural components of the feedback box

Peters supplies the "what is educational" test. To apply it across very different feedback boxes we also need a descriptive framework that identifies the basic operating principles of any feedback box. I propose four structural components that every feedback box possesses, whatever its material form. Whether a given configuration is educational is the separate evaluative question that Peters answers.

The shell is the boundary that sets the terms of entry and participation. It decides who is included, what counts as legitimate activity, what information is visible, and who has authority to change the rules. In a classroom the shell is largely public and legible: the room, the timetable, enrolment, school rules, and the teacher's professional authority. In digital systems the shell is layered. The device itself is a shell; inside it sits the platform shell of interface, notification design, and the algorithms that select what appears next. A slot machine has a shell engineered for immersion: no clocks, no windows, no natural stopping points. The question for any feedback box is who controls the shell, whose purposes it serves, and whether its terms can be inspected by those inside it.

The task space is what the participant is asked to do inside the shell. In a classroom it includes problems set, texts read, questions posed, experiments conducted. In a digital system it includes items presented, pathways offered, and the order in which tasks appear. Every task space carries a path to run, whether anyone names it or not (Matias & Wright, 2022).

The feedback engine is how the box answers the learner. It can operate as bare signal (yes, no, right, wrong, points, streaks, a food pellet) where success and failure are marked but the learner need not understand what is being judged or why. Or it can operate as reasons, making standards visible, explaining what counts as better, and allowing evaluation to be discussed (Peters, 1966). The distinction between signal and reasons does most of the diagnostic work later in this paper.

The trajectory logic describes how the learner's relationship to the box changes over time. Some trajectories aim at exit: competence compounds, evaluative capacity transfers, and eventually the scaffolding is no longer needed. Other trajectories aim at continuation, or put less politely, at entrapment. The system is designed so that the learner keeps returning, whether through subscription, habit, or escalating dependency. The difference is often invisible from inside the box.

The four components can be configured in different ways. The table below sets out three characteristic configurations to make the framework visible.

Table 1: Three configurations of the four-component framework

Component	**Classroom (professionally governed)**	**AI tutoring system (educationally aligned)**	**Commercial digital platform**
Shell	Public, legible, professionally bounded; rules contestable in public.	Personalised and portable; inspectable in principle, governed by educational purposes.	Engineered for immersion; defaults serve retention; terms opaque to those inside.
Task space	Curriculum-shaped by disciplinary standards; sequenced toward powerful knowledge.	Curriculum-shaped despite infinite generation; held to disciplinary structure by educational expertise.	Atomised into short payout items; structured by what is easy to generate and measure.
Feedback engine	Reasons-based; criteria explicit, contestable, carried in human dialogue.	Reasons-based and diagnostic; explanations tied to domain standards.	Signal-based; tuned for persistence through tokens, streaks, near-misses.
Trajectory logic	Aimed at autonomy; scaffolding withdrawn as competence compounds; exit is success.	Aimed at autonomy; memory serves cumulative independence; designed to make itself unnecessary.	Aimed at continuation; exit is failure; dependence is the business model.

The educational configuration

A feedback box is educational when its shell is transparent and justifiable, its task space is structured by worthwhile knowledge, its feedback engine provides reasons, and its trajectory produces cumulative growth toward autonomy.

Feedback in Artificial Intelligence

AI systems track human choices, rank options, and recommend content in real time, drawing on digitised knowledge at a scale unimaginable when Peters wrote. Two computational mechanisms power them: backpropagation and reinforcement learning.

Backpropagation

Backpropagation is the key algorithm for training artificial neural networks (Rumelhart, et al., 1986). The mechanism operates as a strict mathematical loop. Input data travels forward through layers of connected nodes until the network generates an output. This output is compared to a known correct answer, and the difference between the two is calculated as the error. The algorithm then propagates this error backward through the network, layer by layer, adjusting the connection weights so that the nodes contributing most to the mistake are corrected most. Repeated millions of times, the loop minimises the system's errors. It is a pure instance of pattern-matching against a known target without conceptual understanding; the network modifies its weights to optimise a mathematical metric, nothing more.

Reinforcement learning

Reinforcement learning computationally replicates the trial-and-error process Thorndike and Skinner isolated and theorised. A digital agent interacts directly with an environment, executes an action, and receives a reward or penalty based on the outcome. Over time, the agent learns a strategy to maximise its cumulative reward (Sutton & Barto, 2018). Unlike backpropagation, the system does not begin with a dataset of correct answers; it must discover successful behaviours through exploration. It is the Law of Effect rendered in code.

Combined systems

In advanced AI systems, the two are frequently combined (Arulkumaran, et al., 2017). Reinforcement learning explores the environment to generate experience; backpropagation trains the underlying network to predict which actions will yield the highest future rewards. As the predictions improve, the exploration becomes informed and efficient (Mnih, et al., 2015). DeepMind's agents learned to play Atari games at superhuman levels from raw pixel data and a score signal alone (Mnih, et al., 2015), and AlphaGo Zero mastered Go from scratch without any human data, through pure self-play and reinforcement (Silver, et al., 2016, 2017). The systems validate at scale the mechanism Thorndike first observed in cats. There is a recursive quality worth noting: the very large language models now mediating education (and this paper) were themselves trained inside feedback boxes, shaped by human evaluators providing reward signals through reinforcement learning from human feedback (Christiano, et al., 2017). The systems are both products of the feedback box and builders of new ones.

Pedagogic potential

When such systems are turned toward human learning, they speak to what Benjamin Bloom identified as the two sigma problem (Bloom, 1984). Bloom's research demonstrated that students receiving one-on-one tutoring consistently performed two standard deviations higher than peers in conventional classroom settings. The finding highlighted the powerful effect of individualised instruction, but widespread implementation was deemed impractical due to prohibitive costs. Neural networks trained through backpropagation and reinforcement learning can now be embedded in applications that deliver responsive, individualised feedback across subjects, simulating aspects of expert tutoring at scale. Tools such as Khan Academy's Khanmigo are early examples (Khan, 2024). When designed thoughtfully, these systems offer a pathway toward realising Bloom's vision on an unprecedented scale. The same features that make AI tutors responsive also make them capable of capturing attention and conditioning behaviour for non-educational ends. Which way the technology turns depends on how the box is configured.

Pathological configurations

Digital systems sit under strong and continuous profit demands, and there is a long, well-documented history of designing feedback loops to capture attention, steer behaviour, and extract value (Zuboff, 2019). When those incentives govern, the box is optimised for retention and compulsion, packaged as monetisable engagement. The scale is genuinely staggering. Taking Alphabet and Meta alone, these two giants generate over one and a half billion dollars in revenue every single day by keeping around 5 billion human users inside their feedback boxes (Alphabet Inc., 2026; Meta Platforms, 2026; DataReportal, 2026).

The predatory feedback box: Trajectory captured

The predatory feedback box arises when the design objective is capture, retention, and addiction. Its closest relatives are slot machines, social media feeds, and mobile games (Yang, 2025). Natasha Dow Schüll's ethnography of machine gambling shows how reinforcement principles were industrialised for capture (Schüll, 2012). Modern slot machines are reinforcement-delivery systems calibrated through lights, sound, pacing, and micro-intervals. Their interface and rhythm of play cultivate what Schüll calls the 'machine zone', a state of absorbed repetition in which external concerns recede and attention collapses into the loop itself (Schüll, 2012: 2–14). The mechanism is Skinner's variable ratio schedule, refined through testing and deployed for maximum profit. Nir Eyal's *Hooked* provides the manual for turning the logic into product design (Eyal, 2014). Adam Alter traces the same techniques across social media and games, including the deliberate removal of stopping cues: the natural points where a human being might pause, reflect, and decide to stop. Remove the stopping cues and you remove the decision; the loop runs on momentum, not judgement (Alter, 2017). The smartphone with its many apps becomes slot machines in our pockets (Harris, 2016).

Now take that grammar and smuggle it into education under the word *engagement*. The signature failure is in the trajectory: continuation replaces exit as the goal of the system. The other components follow. The shell becomes an interface optimised for return visits. The task space is atomised into short items that pay out rapidly. The feedback engine becomes persistence machinery: you are close, you are improving, the next attempt might pay off. Two of Peters's conditions are structurally violated. Non-manipulation is breached because the system exploits reliable features of human attention and reward learning for ends external to the learner's good. Autonomy is undermined because the arc points toward dependence and information extraction. Zuboff's analysis makes the underlying economics clear: platforms that harvest behavioural data and monetise prediction do not treat prolonged engagement as an accidental by-product but as the product itself (Zuboff, 2019; Lehdonvirta, 2022). Once that incentive enters educational technology, engagement stops being a pedagogic virtue and becomes a revenue category.

The gamified feedback box: Engine captured

The gamified feedback box substitutes tokens for reasons (Mekler, et al., 2017). A sequence of tasks becomes a route through a reward economy of points and levels, with leaderboards hovering like a social weather system. It can be bright, friendly, even wholesome. On the surface the task space may remain recognisably educational. What changes is the framing. Actions are converted into scores and ranks; the interface foregrounds the game layer; the underlying content becomes the occasion for earning tokens rather than the object of study.

Deci, et al.'s (1999) meta-analysis on motivation found that if the environment repeatedly answers a why do this question with an external payoff, the external payoff becomes the focus, even if the activity was intrinsically rewarding in the first place. Recent empirical work confirms the mechanism operates through cognitive pathways: while personalisation and adaptive nudges can support motivation, overtly persuasive platform design simultaneously increases cognitive overload and decreases perceived autonomy, creating competing effects that undermine learning even as they appear to support it (Balaskas, et al., 2025). Alfie Kohn (1999) shows how these rewards push toward an attitude of doing what is minimally sufficient to trigger the reward. Reviews of educational gamification report increases in measurable activity and short-term engagement, with inconsistent effects on learning outcomes (Hamari, et al., 2014).

In Peters's terms the failure is conceptual and ethical. The internal goods of educational practice — coming to see why a proof works, why an explanation is adequate, why a method is valid, why a distinction matters — are displaced by a token economy whose value is defined externally by the system. The learner is invited to treat reasons as optional decoration and the token count as the real story. The box still outputs progress, but it is progress in the game, not necessarily progress in the subject (Patiño Barriga, et al., 2025).

The simulation feedback box: Task space captured

The simulation feedback box emerges when generative AI is positioned as a stand-in agent for the learner's cognition. Visible academic outputs continue to be produced and assessed, but the learner no longer performs the internal sequence of attention, selection, modelling, and self-correction that education presupposes. The breakdown is in the task space: the task is still nominally there, but the learner is no longer the one doing it.

Easier cheating is the surface effect. The deeper shift is a change in the allocation of cognitive agency (Xu, et al., 2025). Current frontier models are increasingly being integrated into agentic systems rather than remaining simple chat interfaces. They can receive a high-level goal and decompose it into sub-tasks, browse the web, execute code, and interact with software, all without further instruction. They maintain persistent memory across sessions, building a cumulative model of the user. They make decisions about which sub-tasks to pursue, evaluate their own outputs, and revise. An agentic system set loose on an assignment can perform the entire cognitive workflow: it finds and evaluates sources, builds the argument, drafts, revises, and inserts the citations. The learner's role can collapse into a thin layer of prompting and oversight, and, even worse, into a single instruction.

Large language models lower the cost of producing plausible work to near zero. A learner can prompt a system and submit fluent output with minimal effort, eviscerating the struggle and friction that makes the process academically meaningful (Chan, 2023). Beyond individual skill loss, AI-assisted work can produce systematic homogenisation, with learners skipping critical developmental stages and converging toward similar solutions (Huo, et al., 2025). Effective resistance requires pedagogical friction, the deliberate design of activities that compel verification and reconstruction rather than passive acceptance (Rocco, 2025). Most assessment systems assumed that student submissions stand as proxies for cognitive work. To write an essay is to read, select, argue, draft, and revise. Generative AI shatters this proxy with slop.

Peters's criterion of transformation tied to reasons is directly undermined. The educated person can give and recognise reasons within a domain. The simulation box produces acceptable artefacts while hollowing out the capacities education exists to develop. As models become more capable and more agentic, this hollowing accelerates: the gap between what the learner appears to know and what the learner actually understands widens with each generation of the technology.

The common pattern

Across these three pathologies the same structural move repeats. Feedback remains highly active, often carefully engineered, but it is stripped of what makes it educational. It is optimised to secure behaviour in the system rather than to build capacities that travel beyond it (O'Neil, 2016). Standards become opaque or displaced, evaluation becomes a stream of signals, and the learner's relationship to the box shifts from apprenticeship toward dependence (OECD, 2026).

This is a predictable drift because the dominant incentives in contemporary digital environments favour retention, continuous measurement, and optimisation at scale. The easiest thing to optimise is what can be counted and captured: clicks, time-on-task, streaks, completion, submission, compliant performance. Educational goods are harder. The damage extends beyond measurable outcomes to encompass cognitive offloading, diminished agency, emotional disengagement, and surveillance concerns that together threaten intellectual autonomy (Favero, et al., 2026).

The educational configuration treats exit as success; the pathological configurations treat it as failure because the system's value depends on continued participation. The system cannot let you go, so algorithms prioritise high-arousal content, systematically amplifying anger, drama, and moral outrage because these emotions are the most effective drivers of viral engagement and platform retention (Brady, et al., 2021; Fan, et al., 2014). When platforms optimise for these hot emotions, they create emotional contagion: the platform's profit motive directly shapes the collective mood of its users. Reasons-based feedback moves users toward intellectual autonomy. Hot-emotion feedback moves them toward reactive impulsivity.

The classroom as ethical feedback box

In a landscape where extractive optimisation pushes feedback toward capture, proxies, dependence, and hot emotions, the classroom's ordinary features take on renewed significance. Its boundaries are visible. Its purposes can be stated. Its curriculum and pedagogy can be argued about in public between teachers, parents, curriculum experts and politicians under well-understood rules of engagement. And its feedback can be anchored in reasons that learners can hear, contest, and eventually carry for themselves. The classroom remains one of the most robust institutional forms for keeping feedback answerable to education rather than to extraction (Selwyn, 2025).

A designed environment

The classroom stabilises conditions that education repeatedly needs: shared attention, sustained talk, visible demonstration, and the slow work of practice and correction (Robson, 1874). Its acoustics, lighting, furniture, and focal points are part of a practical architecture that makes these activities easier to sustain and administer (Cubberley, 1916). This becomes vital in the current context because many digital environments are built around interruption, rapid switching, and constant pings for attention.

Within a school, the classroom belongs to a differentiated ecology of spaces: libraries, laboratories, workshops, studios. That ecology is an institutional claim in built form. It says that different kinds of knowledge require different conditions and tools, and that these conditions should be publicly provided rather than left to private optimisation. It is a social investment in the idea that education depends on structured environments. Access to information alone is not enough.

The classroom also locates feedback in accountable human interaction. Feedback is carried through embodied presence: eye contact, attention, gesture, tone, pacing, shared inspection of work, and the presence of others who can see what is being asked and what counts as a good response (Merleau-Ponty, 2012; Gallagher, 2005). Reasons can be given and challenged in real time. Misunderstanding can be noticed before it is disguised by fluent output.

Structured knowledge and critical resistance

The classroom provides a setting for initiation into structured disciplinary knowledge, and that matters because structured knowledge is one of the few reliable defences against manipulation, whether from propaganda or from algorithmic optimisation. Education initiates learners into what Young and Muller (2016) call powerful knowledge: knowledge that is systematic and principled, with clear concepts, precise terms, and agreed methods for deciding what counts as a good claim (Hirst, 1974). To grasp the reason why is to grasp how a field builds and tests knowledge, and what kinds of reasons are admissible within it.

For example, to learn science is to learn what counts as evidence, how to control observation, how to test hypotheses, and how claims remain open to revision under better evidence (Popper, 1959). To learn history is to learn how arguments are built from sources, how context constrains interpretation, and how claims are justified through disciplined inference rather than assertion. In each case the learner is being inducted into a practice with internal standards. Those standards can be stated, taught, and contested in public.

When this induction succeeds, it cultivates epistemic vigilance: sensitivity to inconsistency, weak inference, and arguments that fail the relevant tests of the domain (Woolcock, 1989). Learners learn to ask for justification and to recognise when the wrong kind of justification is being offered (Boltanski & Thévenot, 2006). Initiation into structured disciplinary knowledge is also training in resistance: a capacity to refuse the wrong feedback, especially when combined with critical AI literacy (Atef & Bali, 2025).

The human dimension

The classroom is also a site of social learning. Structured dialogue, collaborative tasks, and shared experience develop both knowledge and the social competencies and trust that enable critical engagement and intellectual risk (Dewey, 1916; Vygotsky, 1978; Lave & Wenger, 1991). Belonging and emotional security are conditions under which learners can challenge ideas, tolerate difficulty, and revise their thinking (Noddings, 1984; Hargreaves, 2003).

Much practical competence, as Winch (2017) argues, cannot be reduced to propositions or transmitted through disembodied interfaces. Know-how is learned through guided participation, observation, and situated feedback in settings where tools, materials, and expert presence matter (Lakoff & Johnson, 1999). Classrooms, workshops, laboratories, studios remain necessary for this kind of learning. Over-reliance on AI systems that process primarily language-based inputs risks marginalising the embodied and spatial dimensions central to skilled practice, and the intersubjective grain of working together (Jin & Wang, 2025).

The classroom also involves ethical formation. Virtues such as honesty, fairness, and respect for disagreement are modelled in practice as well as stated as principles (Dewey, 1909; Peters, 1966). The teacher's role extends beyond subject instruction to encompass ethical modelling within a learning community whose norms can be made explicit and contested (Arendt, 1961; MacIntyre, 1981).

None of this is automatic. The institutional structure of schooling has been read as a disciplinary apparatus with surveillance-like features (Foucault, 1977), as a cult of business efficiency (Callahan, 1962), as a factory model of batching, pacing, standardisation, and throughput (Tyack & Cuban, 1995), as a vehicle for capitalist exploitation (Bowles & Gintis, 1976), and as an institution that can reproduce social hierarchies even while claiming neutrality (Bourdieu & Passeron, 1977). Those critiques matter here because they show that the classroom's ethical and educational status is not guaranteed by sentiment; it is an ongoing political and professional achievement.

The South African case makes this concrete. Christian National Education and Bantu Education demonstrate how classroom feedback mechanisms can be deliberately configured for oppression rather than liberation (Christie & Collins, 1984; Kallaway, 1984). The classroom's ethical status depends on who controls it, toward what ends, and under what system of public accountability. Its form alone guarantees nothing. Contemporary South African classrooms remain sites where this struggle continues, requiring continuous political, professional, and material commitment (Soudien, 2024). What the South African case also shows is that when classrooms work as ethical feedback boxes, they do so through functioning pedagogic mechanisms: tried and tested ways to secure shared attention, pace practice, make standards visible, stage explanation and correction with exemplars, and handle large numbers while keeping learning publicly accountable. The form is imperfect but robust.

AI tutors, agentic systems, and the next educational form

The classroom has been a durable way of keeping feedback answerable to reasons, standards, and public authority. But it was an answer to a configuration problem of its time, shaped by what was possible in print, in buildings, in salaried professional labour. Classrooms are historical artefacts. They emerged under particular constraints and they have changed many times (Hamilton, 1989). We are now entering another transition in educational form, driven by AI systems whose capabilities are changing faster than most educational discussion has registered (Selwyn, 2025). The question is what counts as educational when feedback is delivered through systems that can remember, adapt, plan, and act autonomously (Alonzo, et al., 2025).

The AI primer, once the realm of science fiction (Stephenson, 1995), is now visible: a tutor for every learner, able to track what has been mastered, what remains fragile, what misconceptions recur, and what kinds of explanation and practice work for that learner. Current frontier systems are agentic. They receive goals and decompose them into sub-tasks, use tools, browse the web, execute code, and interact with software autonomously. They maintain persistent memory, building cumulative models of users across sessions. They evaluate their own outputs and revise (Bornet, et al., 2025). This is a product direction now, not a distant prospect. Khan Academy frames Khanmigo as a Socratic tutor that guides learners towards their own inferences rather than supplying answers, aimed at keeping the learner doing the cognitive work (Khan Academy, 2026). Google's LearnLM initiative presents a similar ambition, making learning science an explicit target for improving generative systems (Google, 2024). UNESCO's guidance treats this promise as real but conditional: benefits depend on governance, privacy protections, and human-centred educational purposes rather than default market incentives (UNESCO, 2023).

Earlier educational forms concentrated the learning environment in a place and time. AI tutors concentrate it in a personalised interface that can follow the learner across settings and across the day. The box becomes persistent and portable. It can sit inside a phone, a platform account, a school system, or an employer tool. It can be private rather than public, continuous rather than periodic, and owned by actors whose purposes are not educational. The technology is a pharmakon in the classical sense (Derrida, 1981): the same structure that can support mastery learning can also poison through capture, dependency, and behavioural steering, all under the guise of personalised intimacy and continuous feedback.

Four design commitments follow from the framework we have been working on throughout this paper. To begin with, memory must serve cumulative independence rather than enclosure. A tutor that remembers can support the compounding of understanding and self-correction. The same memory can support fine-grained persuasion and dependence. Memory in educational systems must therefore be inspectable and correctable, limited to learning-relevant records, and governed in educational terms rather than by platform default. Memory and the controls around it are now presented as standard capabilities, which makes this an educational design question rather than an abstract worry (OpenAI, 2024).

Secondly, task spaces must remain curriculum-shaped even when generation is effectively infinite. Large models can produce endless questions and explanations. That is quantity, not structure. Without disciplinary organisation, the system drifts toward what is easy to generate and easy to score. Powerful disciplinary knowledge remains the target because the tutor's job is to induct learners into conceptual systems and differentiated standards that travel beyond the system, not to keep them busy (Hirst, 1974; Young & Muller, 2016; Muller & Young, 2019). AI agents can become disciplinary experts with algorithmic insights into signature pedagogies and evidence-based practice, but educational expertise is needed to hold the system to account (Winch, 2017).

Thirdly, feedback must persist with reasons and inference, not only signals (Brandom, 2000). Many systems can produce explanations, but explanation is not yet reason-giving in a practice unless standards are made intelligible, errors are diagnosed in domain-relevant terms, and learners are supported to take over the inference and evaluative functions for themselves. Formative assessment remains the best anchor for what educational feedback is supposed to accomplish (Sadler, 1989; Hattie & Timperley, 2007).

Finally, educational AI must preserve the learner as the agent of their own cognitive work. The agentic turn makes this a design problem we cannot defer. When a system can autonomously decompose a task, browse sources, construct arguments, and execute self-critical revision loops, it creates an educational simulacrum: the entire sequence of cognitive operations that education is meant to develop is performed by the system for the system. Peters's foundational criterion of intentional engagement with the reasons behind a task is not just violated; it is rendered structurally unnecessary. The pathology of the simulation feedback box, encountered earlier as a diagnosis, returns here as a design imperative.

The design commitments leave a question open that the science fiction writer Neal Stephenson had already seen in 1995. In *The Diamond Age*, Stephenson imagined the Young Lady's Illustrated Primer: a book that looks like a perfectly configured educational feedback box. Its memory serves the child, its task space is structured by deep cultural and disciplinary knowledge, its feedback explains, and its trajectory aims at autonomy and exit (Stephenson, 1995). The novel is often read as a vision of the ideal AI tutor. The Primer, though, is not a pure AI. Its voices, characters, and real-time responses are performed by paid human actors, called ractors, working remotely through a global network. Nell, the slum child who acquires a Primer by accident, is raised for years by a single ractor, Miranda, who never meets her. Wealthy children with commissioned Primers receive consistent, high-quality ractor attention. Pirated copies for poor children draw on cheaper, more rotational labour. Read this way, *The Diamond Age* is a fable about the human labour substrate underneath what looks like AI tutoring, and about how access to genuinely educational attention, dressed in the same personalised interface, tracks wealth.

That is the equity problem that the design commitments alone cannot resolve. Even a Primer that satisfies all four commitments raises the further question of who receives the well-configured version and who receives the cheap one. In the South African case, where structural inequality has shaped every prior educational form, this is not a speculative worry. It is the next instance of an old pattern. The agentic tutor inherits not only the design problem the four commitments address, but the distributional problem the classroom has never solved either. The Primer can be a tool for subversive autonomy, as it eventually is for Nell, or a high-tech Skinner box that automates the learner out of their own education. The configuration decides one. The political economy of access decides the other.

Higher education as a weaker shell

The school classroom is a thick feedback box: 30 hours a week of compulsory attendance, weekly content prescribed, continuous teacher contact, parental accountability, assessment punctuating the day. All four components sit under sustained external control, which is what lets the classroom resist pathological configurations.

The university is a different box. The shell is thin: 12 to 16 hours a week of formal contact, with most cognitive work expected to happen alone, in private, on the student's own device, at hours the institution does not see. The task space is more open, specified at the level of outcomes and reading lists, with the student constructing the path from outcome to artefact. The feedback engine is sparse, delivered a handful of times per module, usually after the work is done, with the bulk of formative judgement supposed to come from the students themselves against disciplinary standards they are still learning. The trajectory is self-administered: the institution holds the standard but the student must choose, repeatedly, to do the cognitive work.

Every component is where generative AI is most efficient and most seductive. The student is alone with a model, in private, with no presence to interrupt, in a task space the model can saturate, with feedback too sparse to detect when the cognitive work has been displaced. Gerlich (2025) found a significant negative correlation between frequent AI tool use and critical thinking, mediated by cognitive offloading, with the effect strongest among younger users. Tian and Zhang (2025) report the same pattern in 580 university students. Kosmyna, et al. (2025) document measurable changes in neural engagement when essay writing is delegated to a large language model. The effects are partially moderated by prior educational attainment: those who already have evaluative capacities use AI without losing them; those still developing them lose them faster. An anecdote from a colleague's daughter illustrates the point. She loves writing essays as part of thinking, but has found her enjoyment shrinking as friends gain higher marks by offloading the task to AI.

The higher education assessment literature has converged on a structural response, led by the Centre for Research in Assessment and Digital Learning (CRADLE) at Deakin University in Melbourne. Dawson, et al. (2024) reframe the question around validity: the educational concern is not whether the work is the student's, but whether the assessment still indicates that the student has the capacities the qualification claims. Bearman, et al. (2024) place evaluative judgement at the centre of the response, defining it as the learner's capacity to judge the quality of work against disciplinary standards, including their own work and the work an AI produces for them. Corbin, et al. (2025), bluntly titled "talk is cheap", argue that policy adjustments and detection tools will not hold, and that the assessment system itself has to be structurally redesigned.

The three components most under pressure can be sorted in turn. AI-supported marking sits in the feedback engine. The test for any AI marker is whether it generates reasons the student can contest, or scoring signals dressed in explanatory language. Bearman and Ajjawi (2023) frame the prior problem in "Learning to Work with the Black Box": students cannot develop evaluative judgement against systems whose criteria are themselves hidden from them. Generative writing tools sit in the task space, where the university is structurally weakest because its thin shell offers no countervailing presence. The CRADLE response is to redesign the task space so the cognitive operations become visible: oral defence of written work, interim drafts as submission, comparative analysis where students assess AI outputs against disciplinary standards, and assessment organised around process rather than artefact. AI literacy sits in the trajectory logic. The dominant institutional framing of literacy as competent prompting is trajectory-flat: it builds graduates whose evaluative capacities are co-extensive with what the current tools afford. Critical AI literacy, in the sense Atef and Bali (2025) develop, cultivates the evaluative capacities that let students recognise when an output is wrong, when a system is designed to manipulate them, and when their own thinking has been displaced rather than supported. Competent prompting is training. Critical literacy is education. Most current university initiatives stop at the first.

Higher education cannot respond by tightening the shell as a school classroom would, because the university shell was never tight to begin with. Its educational claim has always rested on the student doing cognitive work the institution does not directly observe, evaluated through artefacts treated as proxies. Generative AI has broken the proxy. The response has to be in the task space and the feedback engine: assessment redesigned so the cognitive operations become visible, feedback redesigned so disciplinary reasoning is carried forward by the student, trajectory redesigned so evaluative judgement is the explicit graduate capacity rather than an assumed by-product.

Conclusion

The argument of this paper has been about the structure of educational mechanisms. Feedback is a neutral cybernetic loop. The same mechanism that can educate can condition, capture, and manipulate. Peters supplied the normative test; the four components supplied the descriptive grammar. Together they let us ask what is educational of any feedback box, in any historical period, and get a defensible answer.

The classroom emerges from this analysis as a robust institutional form. Its strength lies in the thickness of its boundaries, the public accountability of its standards, and the durability of its pedagogic mechanisms. Higher education has never operated as a thick shell, and the response to generative AI there has to be different and is more challenging.

We are entering a period in which feedback is agentic and memory-bearing at a scale earlier educational forms never had to confront. This creates real possibilities for the kind of distributed mastery support Bloom glimpsed thirty years ago. It also creates material possibilities for extraction and manipulation under educational branding, and for a deepening of the distributional question of who has access to a well-configured feedback box and who is left with a cheap one. The educational task is, through intentional engagement and morally defensible means, to induct learners into knowledge and understanding that transforms judgement and supports all-round development. It aims at a person who can think, judge, and act for themselves once the scaffolding is gone.

Declaration of Generative AI and AI-assisted technologies in the writing process

The author used Claude Opus 4 and ChatGPT 5 as generative AI tools as research and editorial assistants during drafting and revision. Put more strongly, AI tools acted as thought partners throughout the process. The author reviewed, revised, verified, and takes full responsibility for all arguments, citations, interpretations, and final wording.

Author Biography

Wayne Hugo is an Associate Professor at the University of KwaZulu-Natal's School of Education, and an educational theorist occupied with the question of how teaching actually produces learning? His thinking draws on Basil Bernstein's sociology of education, cybernetics, information theory, and mechanistic explanation. He works across both the long history of pedagogy and its current forms, with hierarchy, professional judgement, causal mechanisms, and the elusive moment of good teaching his habitual points of return. His current intellectual energy centres on how artificial intelligence and human pedagogy work together.

References

Alonzo, D., Abril, J.M.V. & Oo, C.Z. 2025. The use of Artificial Intelligence in English language assessment: Empirical evidence and future directions. *Critical Studies in Teaching and Learning*, 13(2): 102–127.

Alphabet Inc. 2026. *Alphabet Announces Fourth Quarter and Fiscal Year 2025 Results.* Google Investor Relations. Available at: https://abc.xyz/investor/ (Accessed: 6 February 2026).

Alter, A. 2017. *Irresistible: The Rise of Addictive Technology and The Business of Keeping Us Hooked*. London: Penguin Press.

Arendt, H. 1961. *Between Past and Future: Eight Exercises in Political Thought*. London: Viking Press.

Arulkumaran, K., Deisenroth, M.P., Brundage, M. & Bharath, A.A. 2017. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6): 26-38.

Ashby, W.R. 1956. *An Introduction to Cybernetics*. London: Chapman & Hall.

Atef, Y. & Bali, M. 2025. Whose metaphor is it anyway? Analysing AI metaphors from positionality and values of speaker and recipient. *Critical Studies in Teaching and Learning*, 13(2): 128-155.

Balaskas, S., Yfantidou, I., Nikolopoulos, T. & Komis, K. 2025. The psychology of EdTech nudging: Persuasion, cognitive load, and intrinsic motivation. *European Journal of Investigation in Health, Psychology and Education*, 15(9): 179.

Bearman, M. & Ajjawi, R. 2023. Learning to work with the black box: Pedagogy for a world with artificial intelligence. *British Journal of Educational Technology*, 54(5): 1160–1173.

Bearman, M., Tai, J., Dawson, P., Boud, D. & Ajjawi, R. 2024. Developing evaluative judgement for a time of generative artificial intelligence. *Assessment and Evaluation in Higher Education*, 49(6): 893–905.

Benjamin, L.T. 1988. A history of teaching machines. *American Psychologist*, 43(9): 703–712.

Biesta, G. 2009. Good education in an age of measurement: On the need to reconnect with the question of purpose in education. *Educational Assessment, Evaluation and Accountability*, 21(1): 33-46.

Black, P., Harrison, C., Lee, C., Marshall, B. & Wiliam, D. 2003. *Assessment for Learning: Putting it into Practice*. London: Open University Press.

Black, P. & Wiliam, D. 1998. Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1): 7-74.

Bloom, B.S. 1984. The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher*, 13(6): 4-16.

Boltanski, L. & Thévenot, L. 2006 [1991]. *On justification: Economies of worth*. Princeton University Press. (Original work published 1991)

Bornet, P., Wirtz, J., Davenport, T.H., De Cremer, D., Evergreen, B., Fersht, P., Gohel, R. & Khiyara, S. 2025. *Agentic Artificial Intelligence: Harnessing AI Agents to Reinvent Business, Work, and Life*. World Scientific Books.

Bourdieu, P. & Passeron, J.-C. 1977 [1970]. *Reproduction in Education, Society and Culture*. Translated by R. Nice. London: Sage.

Bowles, S. & Gintis, H. 1976. *Schooling in Capitalist America: Educational Reform and The Contradictions of Economic Life*. London: Basic Books.

Brady, W.J., McLoughlin, K., Doan, T.N. & Crockett, M.J. 2021. How social learning amplifies moral outrage expressions in online social networks. *Science Advances*, 7(33): 1-10.

Brandom, R. 2000. *Articulating Reasons: An Introduction to Inferentialism*. Cambridge: Harvard University Press.

Callahan, R.E. 1962. *Education and The Cult of Efficiency*. Chicago: University of Chicago Press.

Chan, C.K.Y. 2023. Is AI changing the rules of academic misconduct? An in-depth look at students' perceptions of 'AI-giarism'. arXiv preprint, arXiv: 2306.03358

Christiano, P.F., Leike, J., Brown, T., Martic, M., Legg, S. & Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems (NeurIPS)*, 30.

Christie, P. & Collins, C. 1984. Bantu education: Apartheid ideology and labour reproduction. In Kallaway, P. (Ed.). *Apartheid and Education*. Johannesburg: Ravan Press: 160-183.

Corbin, T., Dawson, P. & Liu, D. 2025. Talk is cheap: Why structural assessment changes are needed for a time of GenAI. *Assessment & Evaluation in Higher Education,* 50(7): 1087-1097.

Cubberley, E.P. 1916. *Public School Administration*. London: Houghton Mifflin.

Darwin, C. 2003 [1859]. *On The Origin of Species*. London: Signet Classics.

DataReportal. 2026. *Digital 2026: Global Overview Report.* Kepios. Available at: https://datareportal.com/reports/digital-2026-global-overview-report (Accessed: 6 February 2026).

Dawson, P., Bearman, M. & Dollinger, M. 2024. Validity matters more than cheating: Re-centering assessment design in the generative AI era. *Assessment & Evaluation in Higher Education*, 49(7): 945–960.

Deci, E.L., Koestner, R. & Ryan, R.M. 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, 125(6): 627-668.

Derrida, J. 1981 *Dissemination*. Transl. by B. Johnson. Chicago: University of Chicago Press.

Dewey, J. 1909. *Moral Principles in Education*. London: Houghton Mifflin.

Dewey, J. 1916. *Democracy and Education*. London: Macmillan.

Eyal, N. 2014. *Hooked: How to Build Habit-Forming Products*. London: Portfolio.

Fan, R., Zhao, J., Chen, Y. & Xu, K. 2014. Anger is more influential than joy: Sentiment correlation in Weibo. *PLOS ONE*, 9(10): e110184.

Favero, L., Pérez-Ortiz, J.A., Käser, T. & Oliver, N. 2026. AI in education beyond learning outcomes: cognition, agency, emotion, and ethics. arXiv preprint, arXiv:2602.04598.

Ferster, C. B. & Skinner, B. F. 1957. *Schedules of Reinforcement*. Appleton-Century-Crofts.

Foucault, M. 1977 [1975]. *Discipline and Punish: The Birth of the Prison*. Transl. A. Sheridan. New York: Pantheon Books.

Gallagher, S. 2005. *How The Body Shapes The Mind*. Oxford: Oxford University Press.

Gerlich, M. 2025. AI tools in society: Impacts on cognitive offloading and the future of critical thinking. *Societies*, 15(1): 6.

Google. 2024. LearnLM: Improving Gemini for learning. Google Cloud. Available at: https://cloud.google.com/solutions/learnlm (Accessed: 6 February 2026).

Hamari, J., Koivisto, J. & Sarsa, H. 2014. Does gamification work? A literature review of empirical studies on gamification. In *Proceedings of the 47th Hawaii International Conference on System Sciences*. IEEE: 3025-3034.

Hamilton, D. 1989. *Towards a Theory of Schooling*. London: Falmer Press.

Hargreaves, A. 2003. *Teaching in the Knowledge Society: Education in The Age of Insecurity*. New York: Teachers College Press.

Harris, T. 2016. The slot machine in your pocket, 27 July. *Der Spiegel*. Available at: https://www.spiegel.de/international/zeitgeist/smartphone-addiction-is-part-of-the-design-a-1104237.html (Accessed: 6 February 2026).

Hattie, J. & Timperley, H. 2007. The power of feedback. *Review of Educational Research*, 77(1): 81-112.

Hirst, P.H. 1974. *Knowledge and the Curriculum: A Collection of Philosophical Papers*. London: Routledge & Kegan Paul.

Hirst, P.H. & Peters, R.S. 1970. *The Logic of Education*. London: Routledge & Kegan Paul.

Huo, H., Aziz, F. & Kunasegaran, M. 2025. The association between generative AI use and homogenization in entrepreneurship education: Manifestations, potential mechanisms, and implications. *Journal of Advances in Social Sciences*, 1(2): 25-50.

Hyland, T. 2025. Artificial Intelligence (AI), de-growth philosophy and education for sustainability. *Journal of Advances in Education and Philosophy*, 9(11): 496-507.

Jin, H. & Wang, S. 2025. Why not replace people with machines? Exploring Artificial Intelligence's (in)capacity for education as subjectification. *Journal of Philosophy of Education*. https://doi.org/10.1093/jopedu/qhaf087

Kallaway, P. (Ed.). 1984 *Apartheid and Education: The Education of Black South Africans*. Johannesburg: Ravan Press.

Khan, S. 2024. *Brave New Words: How AI will Revolutionize Education (and Why That's a Good Thing)*. London: Crown Publishing.

Khan Academy. 2026. Khanmigo: AI-powered personal tutor and teaching assistant. Khan Academy. Available at: https://www.khanmigo.ai/ (Accessed: 6 February 2026).

Kohn, A. 1999. *Punished by Rewards: The Trouble with Gold Stars, Incentive Plans, A's, Praise, and Other Bribes.* Rev. ed. London: Houghton Mifflin.

Kosmyna, N., Hauptmann, E., Yuan, Y.T., Situ, J., Liao, X.-H., Beresnitzky, A.V., Braunstein, I. & Maes, P. 2025. *Your brain on ChatGPT: Accumulation of cognitive debt when using an AI assistant for essay writing tasks*. arXiv preprint, arXiv:2506.08872.

Lakoff, G. & Johnson, M. 1999. *Philosophy in the Flesh: The Embodied Mind and its Challenge to Western Thought*. New York: Basic Books.

Latour, B. 1999. *Pandora's Hope: An Essay on the Reality of Science Studies*. Cambridge: Harvard University Press.

Lave, J. & Wenger, E. 1991. *Situated Learning: Legitimate Peripheral Participation*. Cambridge: Cambridge University Press.

Lehdonvirta, V. 2022. *Cloud Empires: How Digital Platforms are Overtaking the State and How We Can Regain Control*. Cambridge: MIT Press.

MacIntyre, A. 1981. *After Virtue: A Study in Moral Theory*. Notre Dame, Ill: University of Notre Dame Press.

Matias, N. & Wright, L. 2022. *Impact Assessment of Human-Algorithm Feedback Loops*. Just Tech, Social Science Research Council.

Mekler, E.D., Brühlmann, F., Tuch, A.N. & Opwis, K. 2017. Towards understanding the effects of individual gamification elements on intrinsic motivation and performance. *Computers in Human Behavior*, 71: 525-534.

Merleau-Ponty, M. 2012[1945]. *Phenomenology of Perception.* Transl. D.A. Landes. London: Routledge..

Meta Platforms, Inc. 2026. *Meta Reports Fourth Quarter and Full Year 2025 Results.* Meta Investor Relations. Available at: https://investor.atmeta.com/ (Accessed: 6 February 2026)

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., , Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, ., Kumaran, D., Wierstra, D., Legg, S. 7 Hassabis, D. 2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540): 529-533.

Montessori, M. 1912. *The Montessori Method*. Transl. A.E. George. New York: Frederick A. Stokes.

Montessori, M. 1914. *Dr. Montessori's Own Handbook*. New York: Frederick A. Stokes.

Muller, J. & Young, M. 2019. Knowledge, power and powerful knowledge re-visited. *The Curriculum Journal*, 30(2): 196-214.

Noddings, N. 1984. *Caring: A Relational Approach to Ethics and Moral Education*. Chapel Hill: University of California Press.

OECD 2026. *OECD Digital Education Outlook 2026: Exploring Effective Uses of Generative AI in Education*, OECD Publishing, Paris. Available at: https://doi.org/10.1787/062a7394-en (Accessed: 6 February 2026)

O'Neil, C. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. London: Crown.

OpenAI. 2024. Memory and new controls for ChatGPT. OpenAI. Available at: https://openai.com/index/memory-and-new-controls-for-chatgpt/ (Accessed: 6 February 2026).

Patiño Barriga, D.M., Vargas Sánchez, A.D. & Valdivia Vizarreta, P. 2025. Deep Gamification and Artificial Intelligence as Catalysts of Educational Transformation. *F1000Research*, 14: Article 1156.

Peters, R.S. 1966. *Ethics and Education*. London: Allen & Unwin.

Peters, R.S. 2010 [1967]. What is an educational process? In Peters, R.S. (Ed.). *The Concept of Education.* 2nd ed. London: Routledge, 1-16.

Popper, K.R. 1959. *The Logic of Scientific Discovery*. London: Hutchinson.

Pressey, S.L. 1926. A simple apparatus which gives tests and scores, and teaches. *School and Society*, 23(586): 373-376.

Pressey, S.L. 1927. A machine for automatic teaching of drill material. *School and Society*, 25(645): 549-552.

Robson, E.R. 1874. *School Architecture: Being Practical Remarks on the Planning, Designing, Building, and Furnishing Of School-Houses*. London: John Murray.

Rocco, L. 2025. The mechanical yes-man: Emancipatory AI pedagogy in higher education', *arXiv* [Preprint], arXiv:2510.10176.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J. 1986. Learning representations by back-propagating errors. *Nature*, 323(6088): 533-536.

Sadler, D.R. 1989. Formative assessment and the design of instructional systems. *Instructional Science*, 18(2): 119-144.

Schüll, N.D. 2012. *Addiction by Design: Machine Gambling in Las Vegas*. New jersey: Princeton University Press.

Selwyn, N. 2025. *Digital Degrowth: Radically Rethinking our Digital Futures*. Cambridge: Polity.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, Il, Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. & Hassabis, D. 2016. Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587): 484-489.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Liillicrap, T., Hui, F., Sifre, L., Van den Driessche, G., Graepel, T. & Hassabis, D. 2017. Mastering the game of Go without human knowledge. *Nature*, 550(7676): 354-359.

Skinner, B.F. 1938. *The Behavior of Organisms: An Experimental Analysis*. Appleton-Century-Crofts.

Skinner, B.F. 1953. *Science and Human Behavior*. London: Macmillan.

Skinner, B.F. 1958. Teaching machines. *Science*, 128(3330), 969-977.

Skinner, B. F. 1968. *The Technology of Teaching*. Appleton-Century-Crofts.

Soudien, C. 2024. The quest for educational equity in schools in South Africa. *Daedalus*, 153(4): 146-164.

Stephenson, N. 1995. *The Diamond Age: Or, a Young Lady's Illustrated Primer*. New York: Bantam Spectra.

Sutton, R.S. & Barto, A.G. 2018. *Reinforcement Learning: An Introduction*.2nd ed. Cambridge: MIT Press.

Thorndike, E.L. 1898. Animal intelligence: An experimental study of the associative processes in animals. *Psychological Review Monograph Supplements*, 2(4): i-109.

Thorndike, E.L. 1911. *Animal Intelligence: Experimental Studies*. London: Macmillan.

Tian, J. & Zhang, R. 2025. Learners' AI dependence and critical thinking: The psychological mechanism of fatigue and the social buffering role of AI literacy. *Acta Psychologica*, 260: 105725.

Tyack, D. & Cuban, L. 1995. *Tinkering toward Utopia: A Century of Public School Reform*. Cambridge, MA: Harvard University Press.

UNESCO. 2023. Guidance for generative AI in education and research. Paris: UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000386693 (Accessed: 6 February 2026).

Vygotsky, L.S. 1978. *Mind in Society: The dDvelopment of Higher Psychological Processes*. Ed. M. Cole, V. John-Steiner, S. Scribner and E. Souberman. Cambridge, MA: Harvard University Press.

White, J. & White, P. 2022. Richard Peters and his legacy. *London Review of Education*, 20(1), Article 20.

Wiener, N. 1961. *Cybernetics: Or Control and Communication in the Animal and the Machine.* 2nd ed. Cambridge: MIT Press.

Winch, C. 2017. *Teachers' Know-How: A Philosophical Investigation*. London: Wiley-Blackwell.

Woolcock, P.G. 1989. Forms of knowledge, teacher education and the manipulation of beliefs. *Australian Journal of Teacher Education*, 14(1).

Xu, W., Huang, C., Gao, S. & Shang, S. 2025. LLM-based agents for tool learning: A survey. *Data Science and Engineering*, 10(4): 533–563.

Yang, L. 2025. Innovative applications and challenges of Dobby slot machine algorithms in online gaming. *ITM Web of Conferences*, 73, Article 01023.

Young, M. & Muller, J. 2016. *Curriculum and the Specialization of Knowledge: Studies in the Sociology of education*. London: Routledge.

Zuboff, S. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs.