INTERNATIONAL JOURNAL OF INFORMATION TECHNOLOGY RESEARCH STUDIES (IJITRS)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: 3139-3209



Bridging the Affective Gap: A Conceptual Framework for Multimodal Emotion-Aware Adaptive Learning in Online Education

Afnitha K M

Assistant Professor, Department of Computer Science, Jai Bharath Arts and Science College, Arackappady, Perumbavoor, Kerala, India

Article information

Received: 23rd January 2026
Received in revised form: 24th February 2026
Accepted: 28th March 2026
Available online: 30th April 2026

Volume: 2
Issue: 2
DOI: https://doi.org/10.63090/IJITRS/3139.3209.0028

Abstract

Online learning has scaled rapidly, yet it strips away the nonverbal channel that lets teachers read confusion, frustration, or disengagement and adjust on the spot. This loss is the affective gap. Advances in facial, vocal, and textual emotion recognition make it technically feasible to sense learner affect at a distance, but the field still lacks an integrative design that connects sensing to pedagogy and to responsible governance. Most prior work either reports recognition accuracy on benchmark corpora or reviews that literature descriptively, leaving a gap between detection and classroom action. This paper contributes a conceptual framework, not an empirical system. We propose a four-layer architecture for emotion-aware adaptive learning that spans multimodal sensing, fusion and learning-centred emotion inference, an affective-pedagogical decision layer, and an analytics layer with a teacher in the loop, all wrapped in a cross-cutting governance boundary. We ground the design in a working taxonomy of learning-centred emotions, compare fusion strategies, formalise an uncertainty-gated mapping from inferred states to pedagogical interventions, and surface the privacy, consent, and fairness obligations the design must honour. The framework is offered as a blueprint and a research agenda; we close by outlining how each layer can be validated empirically.

Keywords:- Adaptive Learning, Affective Computing, AI Ethics, Educational Technology, Emotion Recognition, Learning Analytics, Multimodal Fusion, Online Learning.

I. INTRODUCTION

Online and blended education has moved from a supplement to a primary mode of instruction for millions of learners. The convenience is real, but so is a persistent problem: completion rates in large open courses often fall below half, and disengagement is a leading reason learners drift away. In a physical classroom a teacher constantly reads faces, posture, and tone, then slows down, re-explains, or encourages without being asked. A video tile or a discussion thread carries almost none of that signal back to the instructor. Researchers call this missing emotional channel the affective gap, and closing it is now a central concern for educational technology.

Emotion is not a side effect of learning; it shapes attention, memory, and persistence. Pekrun's control-value theory explains how achievement emotions such as enjoyment, anxiety, and boredom arise from a learner's sense of control and value, and how they feed back into effort [4]. Empirical work on tutoring systems shows that confusion, frustration, and boredom are common during difficult learning and that they predict both engagement

and dropout [6], [7]. Confusion in particular can be productive when it is resolved, and harmful when it is left to curdle into frustration [8]. If a system could notice these states and respond well, it might recover some of what the affective gap takes away.

The enabling technologies have matured. Facial expression recognition built on deep convolutional and transformer models reaches strong accuracy on benchmark corpora [11]-[13]. Speech emotion recognition learns affect directly from audio, [14] and transformer language models capture emotional nuance in short text [15], [20]. Multimodal learning offers principled ways to combine these noisy channels into a more stable estimate [18], [19]. Engagement detectors have already been demonstrated inside real online sessions [35]. The pieces exist.

What is missing is the design that joins them. Existing reviews catalogue methods, datasets, and reported accuracies, [26]-[28] which is useful, but a catalogue is not a blueprint. Few contributions specify how a detected emotion should translate into a pedagogical action, how uncertainty in that detection should restrain the action, or how the whole loop should respect a learner's privacy and the validity limits of emotion inference [10]. This paper addresses that design gap directly. It is a conceptual contribution: we propose an architecture and the reasoning behind it rather than a trained system, so the work can be assessed on its coherence and its research value rather than on benchmark numbers.

The contributions are as follows:

- A working taxonomy of learning-centred emotions placed in a valence-arousal space, with the observable cues each emotion leaves across modalities.
- A four-layer conceptual framework for emotion-aware adaptive learning that runs from sensing to action to oversight.
- A comparison of multimodal fusion strategies and the ecological-validity risks of relying on laboratory corpora.
- An uncertainty-gated mapping from inferred affective states to pedagogical interventions, with a human kept in the loop.
- A governance layer that treats privacy, consent, and fairness as first-class design constraints.
- A research agenda describing how each layer can be validated.

II. BACKGROUND AND RELATED WORK

A. Models of emotion

Two traditions dominate. Discrete models, following Ekman, treat emotion as a small set of categories such as happiness, sadness, anger, fear, surprise, and disgust [2]. Dimensional models instead place affect on continuous axes; Russell's circumplex describes any state by its valence and arousal [3]. Learning rarely produces the basic six in their textbook form. The states that matter in a course are confusion, frustration, boredom, curiosity, and flow, which are better described as regions of the valence-arousal plane than as discrete labels [6], [8]. Our taxonomy in Section III adopts this dimensional view while keeping category labels for communication with teachers.

B. Affective computing and recognition technologies

Affective computing, framed by Picard, is the study of systems that recognise, interpret, and respond to human emotion [1]. Calvo and D'Mello's interdisciplinary review mapped its methods and its early educational uses, and argued that combining modalities yields more robust detection than any single channel [5]. Since then, facial recognition has been driven by large in-the-wild corpora such as AffectNet and by deep architectures, from convolutional networks to vision transformers, surveyed by Li and Deng [11], [13], [17]. Speech emotion recognition moved to end-to-end learning from raw audio, [14] supported by acted multimodal corpora [21], [22]. Text emotion analysis was transformed by transformer language models [16] and fine-grained datasets such as GoEmotions [20]. Surveys of multimodal machine learning give a taxonomy of fusion and the alignment problems it must solve [18], [19].

C. Affect in educational systems

Intelligent tutoring systems showed early that adapting to a learner improves outcomes, though the first generation modelled knowledge and ignored affect [24]. Later work folded affective states into student models and found that detecting boredom and frustration sharpened predictions of disengagement, [7], [9] including in authentic in-the-wild settings. Learning analytics provided the data infrastructure for measuring and reporting on learners at scale, [23] and recent systems detect engagement inside live online classes [35]. Systematic reviews confirm momentum but repeatedly note the same shortfalls: validation in controlled settings rather than authentic

ones, attention to basic rather than learning-centred emotions, and thin treatment of ethics [26]-[28]. The framework below is organised around closing exactly those shortfalls.

D. Positioning against recent frameworks

A few conceptual frameworks have appeared very recently, so precise positioning is needed. The Multimodal Affective Tutoring System couples real-time facial, vocal, and code signals to a large language model that delivers empathic feedback, but it is scoped to programming instruction and to an LLM-centred tutor [36]. The E-MOTE framework applies facial action coding with AI and virtual reality to train teachers in emotional awareness rather than to adapt instruction for learners [37]. An earlier decision framework concentrates on selecting actions that shape learners' emotions in adaptive e-learning [38]. Our proposal differs on three counts. It is general to online learning rather than tied to a single subject or model; it is organised around a learning-centred emotion taxonomy and an uncertainty-gated decision layer, so action is restrained by confidence; and it treats ethical governance as a boundary that crosses every layer rather than as a closing remark.

III. A WORKING TAXONOMY OF LEARNING-CENTRED EMOTIONS

A design is only as good as the targets it aims at. Detecting a polished smile of happiness is of little pedagogical use; detecting the moment curiosity tips into frustration is the whole point. We therefore organise the emotions that matter for learning on the valence-arousal plane, as shown in Fig. 1. Positive, activating states such as curiosity and engaged delight sit in the upper right. Frustration and anxiety are negative and activating, in the upper left. Boredom and fatigue are negative and deactivating, in the lower left. Flow, the absorbed state of productive challenge, sits near the centre with mild positive valence and balanced arousal [3], [6].

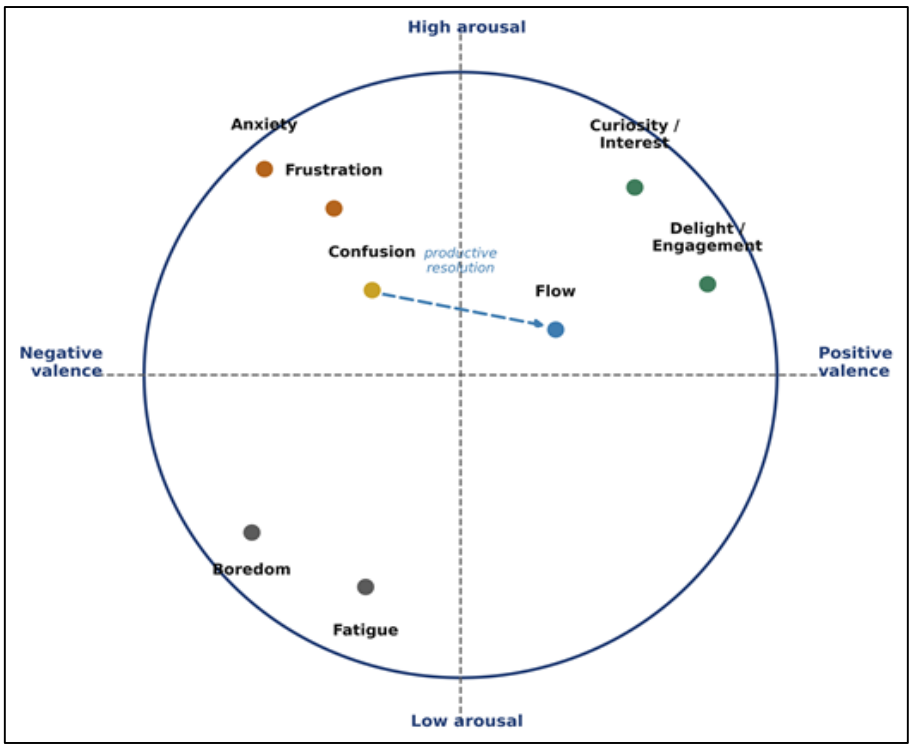


Fig 1: Learning-centred emotions on the valence-arousal plane. The dashed arrow marks the desirable trajectory from confusion to flow through productive resolution.

Two design implications follow. First, the same valence can carry opposite pedagogical meaning depending on arousal: bored disengagement and anxious overload are both negative yet call for very different responses. Second, transitions matter more than snapshots. Confusion that resolves into flow is a sign of healthy struggle, while confusion that decays into frustration and then boredom is the path to dropout [8]. A useful system tracks trajectories, not instants. Table 1 summarises each state, the cues it tends to leave in each modality, its pedagogical significance, and a first-line response.

Table 1. Learning-centred emotions, observable cues, and indicative responses.

Learning emotion	Typical multimodal cues	Pedagogical significance	First-line response
Curiosity / interest	Forward lean, raised brows, exploratory clicks, questioning text	Window for deeper challenge	Offer enrichment, harder problems
Engaged delight	Smiling, animated prosody, rapid on-task actions	Confidence and momentum	Maintain pace, acknowledge progress
Flow	Steady gaze, low fidget, consistent on-task pace	Optimal learning state	Protect it; avoid interruption
Confusion	Brow furrow, pauses, re-reading, hedging language	Productive if resolved soon [8]	Hint or scaffold, then recheck
Frustration	Tightened expression, tense voice, repeated errors, negative text	Risk of giving up	Reduce difficulty, encourage, offer help
Anxiety	Restlessness, shaky voice, worried phrasing	Impairs working memory [4]	Reassure, lower stakes, slow pacing
Boredom	Slumped posture, flat prosody, off-task latency	Strong dropout predictor [7]	Increase relevance, novelty, interactivity
Fatigue	Eye closure, yawning, slowing response time	Degraded performance	Suggest a break or session end

These cues are indicative, not definitive. Barrett and colleagues warn that facial configurations map to emotional states only loosely and that context heavily conditions meaning, so inferring an internal feeling from an outward sign is inherently uncertain [10]. This caution is built into the framework: detection feeds a probability with a confidence estimate, never a verdict, and the decision layer is gated accordingly.

IV. FRAMEWORK OVERVIEW

Fig. 2 presents the framework as four stacked layers wrapped by a governance boundary. Signals flow up from the learner through sensing and inference; decisions and adaptations flow back down and out to the learner, closing the loop. Four principles shape the design. Multimodality: no single channel is reliable across the conditions of real online study, so the architecture assumes several noisy channels and fuses them. Uncertainty-awareness: every inference carries a confidence, and low confidence suppresses action. Human-in-the-loop: the system advises and automates routine support, but a teacher retains oversight of consequential decisions [25]. Privacy-by-design: governance is not an appendix but a boundary that crosses every layer.

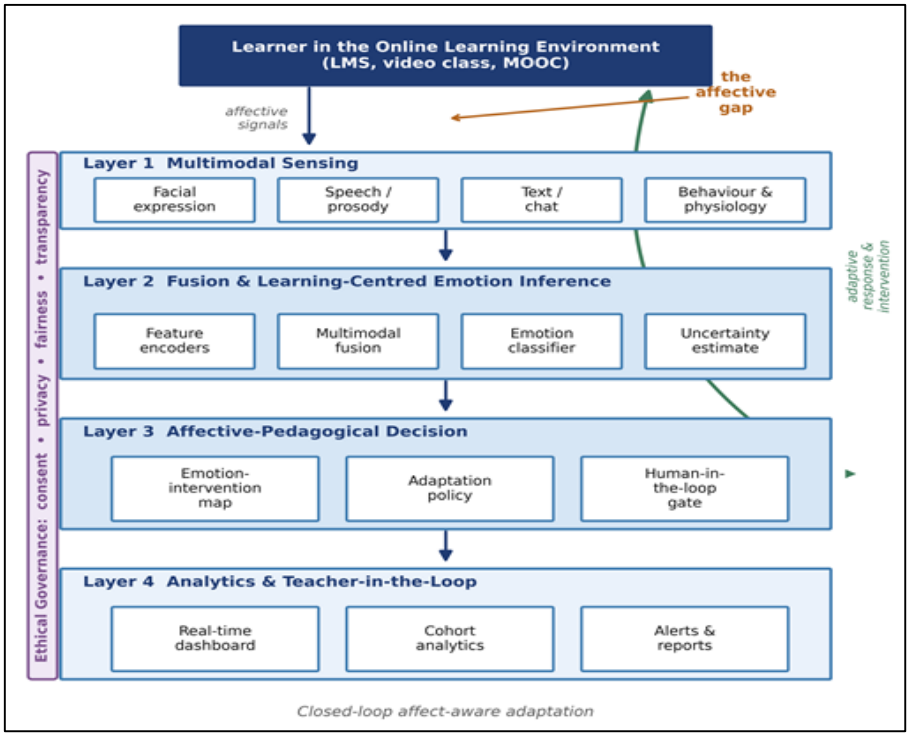


Fig 2: The four-layer conceptual framework for emotion-aware adaptive learning, with a cross-cutting ethical governance boundary and a closed adaptation loop.

Layer 1 senses the learner across facial, vocal, textual, and behavioural channels. Layer 2 encodes and fuses these signals and infers a learning-centred emotional state together with an uncertainty estimate. Layer 3 maps that state, when confidence permits, to a pedagogical intervention under a policy and a human-in-the-loop gate. Layer 4 turns the same stream into analytics for instructors and feeds alerts back into the loop. The remaining sections develop each layer.

V. MULTIMODAL SENSING AND FUSION LAYER

A. Channels and signals

Four channels are practical in an online setting. The facial channel uses the webcam to read expression and head pose [11], [13]. The vocal channel reads prosody from microphone audio during speech or discussion [14], [21]. The textual channel analyses chat, forum posts, and open responses [15], [20]. The behavioural channel needs no extra sensor: clickstreams, response latency, pauses, navigation, and performance already encode engagement and are the least intrusive source [35]. Physiological signals add accuracy but require wearables and are treated here as optional.

B. Fusion strategies

Because each channel is noisy and intermittently unavailable, the inference quality depends on how channels are combined. Fig. 3 sketches three families. Early fusion concatenates features before a single model, capturing low-level cross-modal interactions but coupling the modalities tightly. Late fusion runs a model per modality and combines decisions, which is robust to a missing channel but blind to interactions. Hybrid and attention-based fusion learn where to look across modalities and can weight each by its reliability, at the cost of complexity and data appetite [16], [18], [19]. Table 2 compares them on the criteria that matter for deployment.

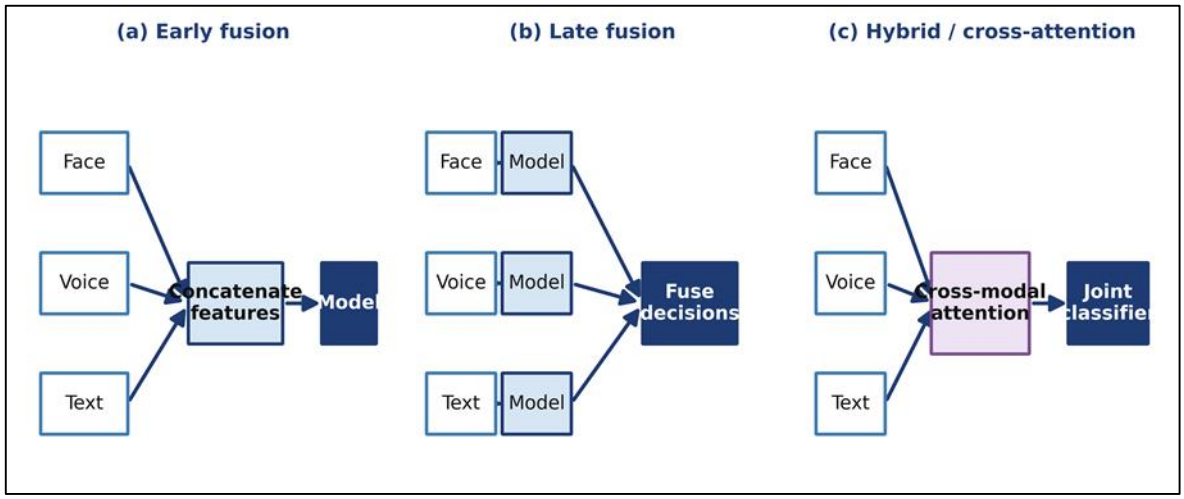


Fig 3: Multimodal fusion strategies: (a) early feature-level fusion, (b) late decision-level fusion, and (c) hybrid fusion with cross-modal attention.

Table 2. Comparison of multimodal fusion strategies for online-learning conditions.

Strategy	Cross-modal interaction	Robust to missing modality	Cost / data need	Best fit
Early (feature)	High	Low	Moderate	Tightly coupled cues, complete data
Late (decision)	Low	High	Low	Intermittent channels, modular reuse
Hybrid / attention	High, adaptive	Moderate-High	High	Variable reliability, ample data

C. Ecological validity and uncertainty

A recurring weakness in the literature is that recognition models are trained and evaluated on acted or curated corpora gathered under good lighting with frontal faces and clean audio, such as AffectNet, RAVDESS, CREMA-D, and GoEmotions [11], [20]-[22]. Authentic online study looks nothing like that: side lighting, off-axis webcams, background noise, code-switching, and learners who suppress expression. Models therefore suffer domain shift, and reported laboratory accuracy overstates field performance [10]. The framework responds in two ways. It prefers the behavioural channel, which transfers better because it does not depend on appearance, and it

requires every inference to be calibrated and to expose an uncertainty estimate that the next layer can act on. Cross-domain validation, not benchmark accuracy, is the standard the sensing layer must eventually meet [13].

VI. AFFECTIVE-PEDAGOGICAL DECISION LAYER

Sensing without pedagogy is surveillance. The decision layer is where an inferred state becomes a teaching act. Fig. 4 shows its structure. The inferred state and its confidence enter an adaptation policy; if confidence is below a threshold the policy defers and takes no action, which protects learners from being acted on by a guess [10]. When confidence is sufficient, the policy selects an intervention proportionate to the state, and consequential actions pass through a human-in-the-loop gate before reaching the learner [25].

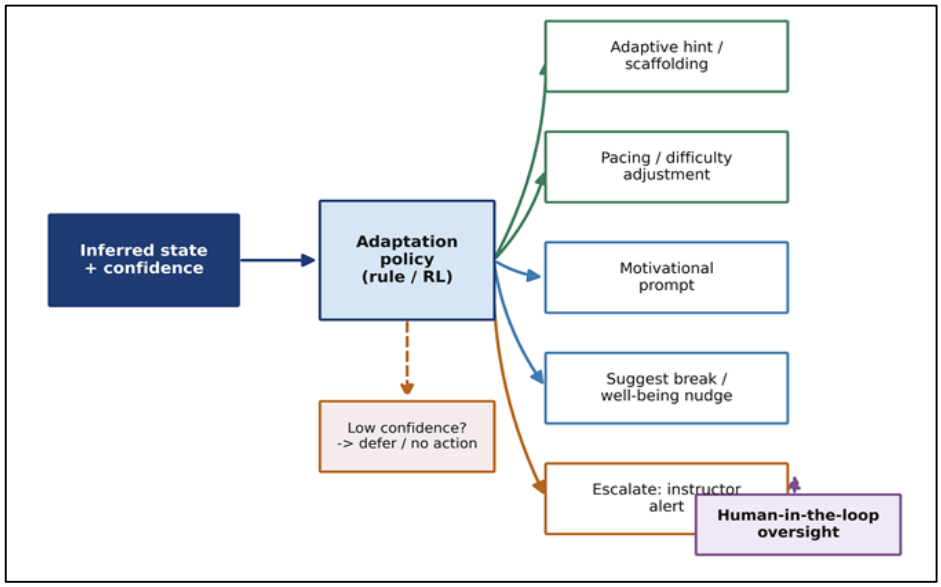


Fig 4: The affective-pedagogical decision layer: an uncertainty-gated mapping from inferred state to proportionate intervention, with human oversight on escalation.

The mapping itself should be pedagogically grounded rather than ad hoc. Control-value theory and the confusion-can-help finding suggest, for example, that brief confusion deserves a hint that preserves the struggle, whereas sustained frustration deserves a genuine reduction in difficulty. [4], [8] Table 3 gives an indicative mapping. The policy can be implemented as transparent rules, which are auditable and preferable early on, or learned with reinforcement learning once enough interaction data and a safe reward signal exist; the two can coexist, with rules as guardrails around a learned policy.

Table 3. Indicative emotion-to-intervention mapping used by the decision policy.

Inferred state (sustained)	**Pedagogical intervention**	**Rationale**
Confusion	Targeted hint or worked step; then recheck	Preserve productive struggle [8]
Frustration	Lower difficulty, offer help, encourage	Prevent giving up [7]
Boredom	Inject relevance, novelty, interactivity	Re-engage before disengagement [6]
Anxiety	Reassure, lower stakes, slow pacing	Free working memory [4]
Flow	No interruption; protect the state	Avoid harming optimal learning
Fatigue	Suggest a break or stopping point	Avoid degraded practice
Persistent negative + low control	Escalate to instructor with context	Human judgement and care [25]

Timing is as important as choice. An intervention delivered too early interrupts a learner who would have resolved the difficulty alone; delivered too late it arrives after disengagement has set in [6]. The policy therefore acts on trajectories and dwell time rather than single frames, which is why the framework treats emotional state as a tracked signal over a session.

VII. ANALYTICS, TEACHER-IN-THE-LOOP, AND ETHICAL GOVERNANCE

A. Analytics and the teacher in the loop

Not every response should be automated. Layer 4 aggregates the affective stream into analytics that keep a human informed and in control: a real-time dashboard of individual and cohort states, trajectories across a module, and alerts when a learner shows persistent negative affect [23]. Aggregated affective data also tells instructional designers which materials reliably produce confusion or boredom, supporting evidence-based revision. Fig. 5 shows the full cycle as a closed loop, sense, infer, decide, act, and the learner responds, all inside a governance boundary. The teacher is not outside this loop but a participant in it, receiving context and retaining authority over escalations [25].

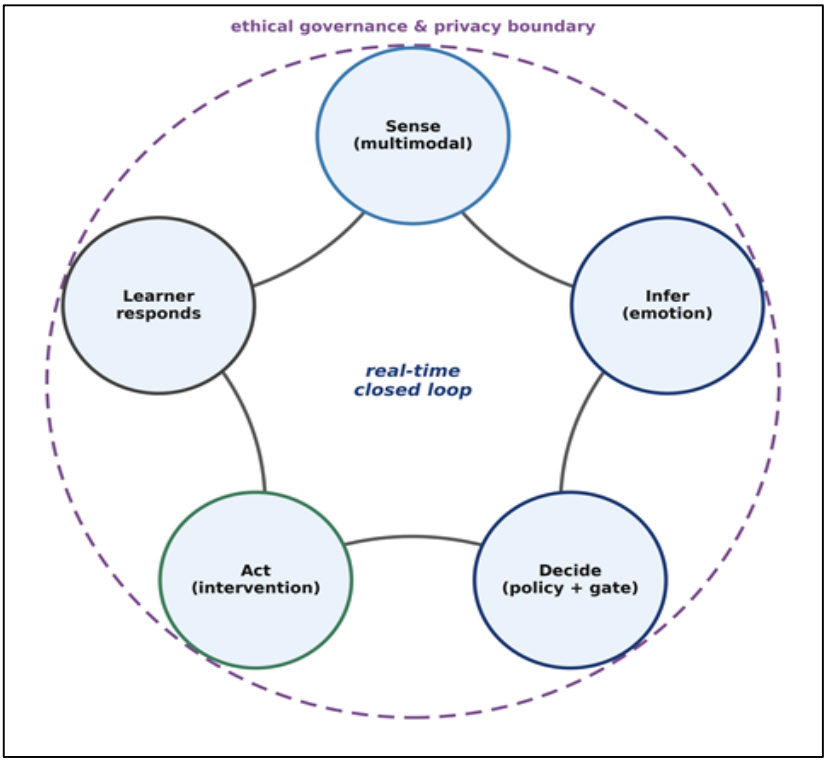


Fig 5: The real-time closed loop of affect-aware adaptation, enclosed by an ethical governance and privacy boundary.

B. Ethical governance

Emotion data is intimate, and inferring it can be wrong, so governance is not optional. Three obligations cut across every layer. Privacy and consent: capturing a learner's face, voice, and affect demands informed, revocable consent and data minimisation; processing on the device or training with federated learning keeps raw signals off central servers, and differential privacy bounds what aggregates can leak [29], [30]. Fairness: recognition models are known to perform unevenly across demographic groups, so the system must be audited for disparate accuracy and corrected before deployment, lest it deliver worse support to the learners who already face barriers [33]. Transparency: learners deserve to understand that affect is being sensed and how it is used, and instructors deserve explanations they can question, which is where saliency and feature-attribution methods enter. [31], [32] These duties align with emerging frameworks for the ethics of AI in education [34]. Table 4 pairs the principal risks with concrete safeguards.

Table 4. Ethical risks and the safeguards embedded across the framework layers.

Risk	Safeguard built into the framework
Surveillance and loss of privacy	Informed revocable consent, data minimisation, on-device or federated processing [29]
Leakage from stored affect data	Differential privacy on aggregates; short retention [30]
Demographic bias in recognition	Subgroup accuracy audits and bias mitigation before deployment [33]
Acting on wrong inferences	Confidence gating; emotion treated as probabilistic, not certain [10]
Opaque automated decisions	Explainable inferences and auditable rule-based policy [31], [32]
Erosion of learner autonomy	Human-in-the-loop oversight; learner can opt out [25], [34]

VIII. DISCUSSION AND RESEARCH AGENDA

The framework is a blueprint, and a blueprint earns its keep by directing research. Five open problems stand out. First, the validity of emotion inference itself: given that outward signs map loosely to inner states, how should confidence be calibrated and communicated so that action stays proportionate to certainty [10]. Second, robustness in the wild: closing the gap between benchmark accuracy and field performance under realistic lighting, audio, and learner behaviour [13]. Third, cross-cultural validity, since emotional expression and interpretation vary across cultures and models trained on one population may misread another [5]. Fourth, the scarcity of authentic, ethically collected datasets of learning-centred emotion, as distinct from acted corpora [20]-[22]. Fifth, the effectiveness question: whether emotion-aware adaptation actually improves engagement and outcomes, which only controlled field studies can answer.

These problems suggest a staged validation of the framework itself. The sensing and inference layers can be evaluated first through cross-corpus and cross-domain tests that measure how well recognition transfers from curated data to authentic sessions, reporting calibration alongside accuracy. The fusion design can be tested by ablation, removing channels to confirm that late and hybrid fusion degrade gracefully when a modality drops out. The decision layer can be examined in simulation and then in small randomised studies that compare emotion-aware adaptation against an unaware baseline on engagement, persistence, and learning gains. The governance layer can be assessed through subgroup fairness audits and through studies of learner trust and acceptance. This sequence mirrors a phased doctoral programme and lets each layer be confirmed before the next is built on it.

The contribution and its limits should be stated plainly. This is a conceptual and architectural paper; it does not report a trained system or empirical results, and the indicative mappings in Tables 1 and 3 are hypotheses to be tested, not validated rules. Its value is in organising a fragmented field into a coherent, ethically bounded design and in making the research questions concrete. We regard that as a necessary step before, not a substitute for, the empirical work it frames.

IX. CONCLUSION

The affective gap is the price online learning pays for scale and distance: the emotional signals a teacher reads in a room rarely survive the screen. Affective computing can recover some of those signals, but only if sensing is connected to pedagogy and to responsible governance rather than left as a recognition benchmark. We have proposed a four-layer conceptual framework that makes that connection, grounded in a taxonomy of learning-centred emotions, structured around uncertainty-aware decisions and a teacher kept in the loop, and bounded throughout by privacy, fairness, and transparency. The design is offered as a blueprint and a research agenda. Realising emotion-aware learning that genuinely helps learners, and does so fairly, will take the staged empirical validation this framework is built to guide.

REFERENCES

[1] R. W. Picard, Affective Computing. Cambridge, MA, USA: MIT Press, 1997.
[2] P. Ekman, "An argument for basic emotions," Cognition and Emotion, vol. 6, no. 3-4, pp. 169-200, 1992.
[3] J. A. Russell, "A circumplex model of affect," J. Personality and Social Psychology, vol. 39, no. 6, pp. 1161-1178, 1980.
[4] R. Pekrun, "The control-value theory of achievement emotions: Assumptions, corollaries, and implications for educational research and practice," Educational Psychology Review, vol. 18, no. 4, pp. 315-341, 2006.
[5] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," IEEE Trans. Affective Computing, vol. 1, no. 1, pp. 18-37, 2010.
[6] S. D'Mello and A. Graesser, "Dynamics of affective states during complex learning," Learning and Instruction, vol. 22, no. 2, pp. 145-157, 2012.
[7] R. S. J. d. Baker, S. K. D'Mello, M. M. T. Rodrigo, and A. C. Graesser, "Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive-affective states during interactions with three different computer-based learning environments," Int. J. Human-Computer Studies, vol. 68, no. 4, pp. 223-241, 2010.
[8] S. D'Mello, B. Lehman, R. Pekrun, and A. Graesser, "Confusion can be beneficial for learning," Learning and Instruction, vol. 29, pp. 153-170, 2014.
[9] N. Bosch et al., "Automatic detection of learning-centered affective states in the wild," in Proc. 20th Int. Conf. Intelligent User Interfaces (IUI), 2015, pp. 379-388.
[10] L. F. Barrett, R. Adolphs, S. Marsella, A. M. Martinez, and S. D. Pollak, "Emotional expressions reconsidered: Challenges to inferring emotion from human facial movements," Psychological Science in the Public Interest, vol. 20, no. 1, pp. 1-68, 2019.
[11] Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," IEEE Trans. Affective Computing, vol. 10, no. 1, pp. 18-31, 2019.
[12] J. Goodfellow et al., "Challenges in representation learning: A report on three machine learning contests," Neural Networks, vol. 64, pp. 59-63, 2015.
[13] S. Li and W. Deng, "Deep facial expression recognition: A survey," IEEE Trans. Affective Computing, vol. 13, no. 3, pp. 1195-1215, 2022.

[14] G. Trigeorgis et al., "Adieu features? End-to-end speech emotion recognition using a deep convolutional recurrent network," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP), 2016, pp. 5200-5204.
[15] Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in Proc. NAACL-HLT, 2019, pp. 4171-4186.
[16] Vaswani et al., "Attention is all you need," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017, pp. 5998-6008.
[17] Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in Proc. Int. Conf. Learning Representations (ICLR), 2021.
[18] T. Baltrusaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 41, no. 2, pp. 423-443, 2019.
[19] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," Information Fusion, vol. 37, pp. 98-125, 2017.
[20] D. Demszky et al., "GoEmotions: A dataset of fine-grained emotions," in Proc. 58th Annu. Meeting Assoc. Computational Linguistics (ACL), 2020, pp. 4040-4054.
[21] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," PLoS ONE, vol. 13, no. 5, e0196391, 2018.
[22] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "CREMA-D: Crowd-sourced emotional multimodal actors dataset," IEEE Trans. Affective Computing, vol. 5, no. 4, pp. 377-390, 2014.
[23] G. Siemens and P. Long, "Penetrating the fog: Analytics in learning and education," EDUCAUSE Review, vol. 46, no. 5, pp. 30-40, 2011.
[24] R. Anderson, A. T. Corbett, K. R. Koedinger, and R. Pelletier, "Cognitive tutors: Lessons learned," J. Learning Sciences, vol. 4, no. 2, pp. 167-207, 1995.
[25] D. Richards and V. Dignum, "Supporting and challenging learners through pedagogical agents: Addressing ethical issues through designing for values," British J. Educational Technology, vol. 50, no. 6, pp. 2885-2901, 2019.
[26] S. K. Khare, V. Blanes-Vidal, E. S. Nadimi, and U. R. Acharya, "Emotion recognition and artificial intelligence: A systematic review (2014-2023) and research recommendations," Information Fusion, vol. 102, art. 102019, 2024.
[27] O. R. Vistorte et al., "Integrating artificial intelligence to assess emotions in learning environments: A systematic literature review," Frontiers in Psychology, vol. 15, art. 1387089, 2024.
[28] S. A. Salloum et al., "Emotion recognition for enhanced learning: Using AI to detect students' emotions and adjust teaching methods," Smart Learning Environments, vol. 12, art. 21, 2025.
[29] McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in Proc. 20th Int. Conf. Artificial Intelligence and Statistics (AISTATS), 2017, pp. 1273-1282.
[30] Dwork, F. McSherry, K. Nissim, and A. Smith, "Calibrating noise to sensitivity in private data analysis," in Proc. Theory of Cryptography Conf. (TCC), 2006, pp. 265-284.
[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in Proc. IEEE Int. Conf. Computer Vision (ICCV), 2017, pp. 618-626.
[32] T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016, pp. 1135-1144.
[33] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in Proc. Conf. Fairness, Accountability and Transparency (FAccT), 2018, pp. 77-91.
[34] W. Holmes et al., "Ethics of AI in education: Towards a community-wide framework," Int. J. Artificial Intelligence in Education, vol. 32, pp. 504-526, 2022.
[35] S. Gupta, P. Kumar, and R. K. Tekchandani, "Facial emotion recognition based real-time learner engagement detection system in online learning context using deep learning models," Multimedia Tools and Applications, vol. 82, pp. 11365-11394, 2023.
[36] Author(s), "A conceptual framework for an LLM-powered multimodal affective tutoring system for programming education," in Proc. 3rd Int. Conf. Educational Knowledge and Informatization (ICEKI), 2025, doi: 10.1145/3765325.3765345.
[37] Author(s), "E-MOTE: A conceptual framework for emotion-aware teacher training integrating FACS, AI and VR," Vision (MDPI), vol. 10, no. 1, art. 5, 2026.
[38] Author(s), "Toward an artificial intelligence-based decision framework for developing adaptive e-learning systems to impact learners' emotions," Interactive Learning Environments, 2023, doi: 10.1080/10494820.2023.2188398.