This paper was presented at the XXXVII ISPIM Innovation Conference, held in Granada, Spain on 07 June to 10 June 2026. ISBN 978-952-65771-2-8.

Responsible AI Co-Creation: A Framework for Adaptive Systems

Barsha Goswami

Quantile AS, Stavanger, Norway.
E-mail: barsha@quantile.no

Antorweep Chakravorty

University of Stavanger, Stavanger, Norway.
E-mail: antorweep.chakravorty@uis.no

Tatiana Aleksandrovna Iakovleva*

University of Stavanger, Stavanger, Norway.
E-mail: tatiana.a.iakovleva@uis.no
* Corresponding author

Abstract: Digital transformation is reshaping the role of users from passive recipients of technology to active participants in its development. The integration of AI, further extends this shift as such systems start to learn and adapt based on user interactions. Co-creation, in this setting, becomes a continuous and evolving process rather than a one-time activity. Responsible Innovation (RI) offers an established approach to align technological development with societal values. However, existing RI frameworks do not adequately account for the adaptive and continuous nature of AI systems. RI provides limited guidance on ethical and regulatory considerations integrated throughout the AI development lifecycle. This paper proposes a conceptual framework that builds upon key principles from RI, user innovation and AI governance in a complementary manner, to support responsible co-creation in adaptive AI systems. The framework is applied to an AI-integrated educational platform as an illustrative case study to assess its practical relevance.

Keywords: Responsible Innovation; AI Governance; User Co-Creation; Adaptive AI Systems; Ethical AI Development; Ethics-by-Design; EU AI Act; Co-Creation Lifecycle; Third-party AI Governance; Educational AI.

1 Introduction

Artificial intelligence (AI) is no longer confined to specialist technical domains. It has permeated professional practice and everyday life, reshaping how individuals engage with technology and, more fundamentally, how they participate in its ongoing development. Unlike earlier generations of technology that operated within relatively fixed parameters following deployment, AI systems possess a distinctive capacity for continuous learning and adaptation through user interaction. As (Bessant et al., 2024) observe, this makes users not merely end-consumers of a finished product, but active contributors to a system that evolves in response to their behaviour. Co-creation in the context of AI, therefore, is

not a discrete phase within a bounded design process, but an iterative and open-ended dynamic that persists across the complete lifecycle of the system.

This characteristic does not fit well within existing frameworks for governing technological innovation. Responsible Innovation (RI) has emerged as an influential paradigm for aligning innovation processes with broader societal values, structured around the dimensions of anticipation, reflexivity, inclusiveness and responsiveness (Stilgoe et al., 2013). However, despite its appeal, a persistent gap has been noted in the literature between its principled commitments and actionable guidance for practitioners operating within specific technological contexts (Iakovleva et al., 2021; Buhmann and Fieseler, 2022). This limitation is particularly relevant for adaptive AI systems, where the continuous and consequential nature of user involvement demands a more responsive governance logic.

Parallel developments in AI governance and ethics have gone some way toward filling this void. The EU AI Act represents a significant regulatory intervention, establishing binding requirements around transparency, risk classification and accountability across the AI value chain. Complementary scholarly work on ethics-by-design and governance-by-design has sought to embed such considerations directly into development processes rather than treating them as post-hoc compliance obligations (Šekrst et al., 2024; Brey and Dainow, 2023). At the same time, user innovation theory demonstrates the value of recognising users as substantive contributors to innovation outcomes, rather than passive recipients of technological solutions (Huang et al., 2024; Gago and Rubalcaba, 2024).

Taken together, responsible innovation, AI governance and ethics, and user innovation theory each address part of the problem. However, they have largely evolved in isolation, and their intersection, particularly in relation to adaptive AI systems characterised by continuous co-creation, remains insufficiently addressed. Recent work by (Herrmann, 2023) gestures toward this convergence, but a unifying perspective has yet to be developed.

This paper, through a systematic review of literature across these three domains, examines how their respective contributions can be brought into productive dialogue. On this basis, we propose a conceptual framework that synthesises key principles from responsible innovation, user innovation and AI governance to support responsible co-creation in adaptive AI systems. To assess its practical relevance, the framework is applied to an AI-integrated educational platform as an illustrative case study.

The rest of the paper is structured as follows. Section 2 reviews the relevant literature across the three domains. Section 3 presents the proposed framework. Section 4 applies the framework to the case study. Section 5 concludes the paper with directions for future research.

2 Related Works

A structured literature search was conducted across responsible innovation, user innovation and AI governance using Google Scholar, with keyword combinations including "responsible innovation framework", "user co-creation innovation", "AI governance ethics by design" and "EU AI Act compliance". The aim was not exhaustive

coverage, but targeted identification of key contributions and gaps relevant to developing a unified framework for responsible AI co-creation.

RI is a governance paradigm oriented toward making research and innovation processes transparent, inclusive and socially responsive. (Owen et al., 2012) offered an early conceptualisation, framing RI as a reorientation from science in society toward science for and with society. (Stilgoe et al., 2013) subsequently formalised this into the AREA framework, built around anticipation, reflexivity, engagement and action, as a structured reference point for governing emerging technologies. However, though widely cited, the framework was developed largely in a reflective manner rather than empirically. (Fraaije and Flipse, 2019) addressed this operationalisation gap by proposing a practical RRI implementation structure drawing on policy documents and academic literature. Further, (Ulnicane et al., 2022) argued that legal compliance continues to dominate the RI agenda in practice, and that moving beyond it requires integrating the AREA dimensions iteratively and experimentally, rather than as a sequential checklist.

Literature on user involvement in innovation demonstrates a well-established shift from passive recipient to active co-creator (Bessant et al., 2024; Iakovleva et al., 2021). (Huang et al., 2024) examined user engagement within living labs, identifying four key phases: recruitment, motivation, co-creation and relation. (Gago and Rubalcaba, 2024) further provided empirical evidence that active co-creation methodologies produce outcomes more closely aligned with end-user needs than passive involvement approaches. (Catarci et al., 2020) illustrated how participatory design approaches from ICT and product design can function as triggers for innovation, while (Laitio and Nätti, 2023) examined embedded lead users as contributors across development stages through a literature review.

The governance of AI systems has become a major focus of both regulatory and academic attention. The EU AI Act establishes a risk-based framework with binding requirements around transparency, accountability and conformity assessment (Cancela-Outeda, 2024; Canavese et al., 2024). (Canavese et al., 2024) translated its technical and organisational requirements into architectural guidelines for compliance-by-design. However, their analysis is narrowly compliance-focused and does not engage with user involvement or RI principles. (Prifti et al., 2024) offered a broader synthesis of regulation-by-design, distinguishing compliance, value creation and optimisation orientations, while noting persistent challenges around contextual applicability and methodological uncertainty. (Brey and Dainow, 2023) proposed an Ethics by Design approach for AI (EbD-AI), mapping six pre-selected ethical values onto design requirements across development phases; an approach adopted by the European Commission, though criticised for its rigidity and practical burden on designers. (Šekrst et al., 2024) addressed this through a customisable guardrail framework. They also acknowledge the need for deeper user involvement in ethical governance. (Buhmann and Fieseler, 2022) connected RI and AI governance through a deliberative framework spanning public, private and civil society actors, rather than a top-down regulatory approach. However, their treatment of user engagement under the purview of civil society participation does not adequately account for individual users whose ongoing interactions directly shape how adaptive AI systems evolve.

(Herrmann, 2023) demonstrated through a systematic review that responsible AI has developed largely independently of the RI tradition, with limited exchange of ideas between the two. The larger pattern still is that RI, user innovation and AI governance have each advanced on their own terms, but have not been brought together in a

complementary manner. RI offers principled governance dimensions, but lacks mechanisms suited to continuously adaptive technologies. User innovation research establishes the value of active participation, but has not theorised the distinctive dynamics of AI co-creation, where the system itself is transformed through use. Finally, AI governance embeds safeguards into design, but tends to position users as subjects of protection rather than as co-creators whose sustained involvement requires structural support.

3 Framework

Traditional innovation follows a path of design, user involvement and deployment. However, AI-based systems disrupt this process flow. Deployment is not an endpoint, but a threshold beyond which users continue to interact with the system, and those interactions shape its behaviour in ways that cannot be fully anticipated in advance. Co-creation, in this context, is open-ended and persists across the complete lifecycle of the system. Addressing this requires a framework that integrates RI principles of anticipation, reflexivity, inclusiveness and responsiveness (Stilgoe et al., 2013; Owen et al., 2012), with structured user involvement mechanisms covering recruitment, motivation, co-creation and relation (Huang et al., 2024), and AI governance safeguards encompassing ethical boundaries, regulatory compliance and conflict resolution (Brey and Dainow, 2023; Canavese et al., 2024; Buhmann and Fieseler, 2022).

The proposed framework, illustrated in Figure 1, comprises three layers operating simultaneously and continuously across the AI system's lifecycle. They are not sequential stages; each layer informs and receives feedback from the others throughout.

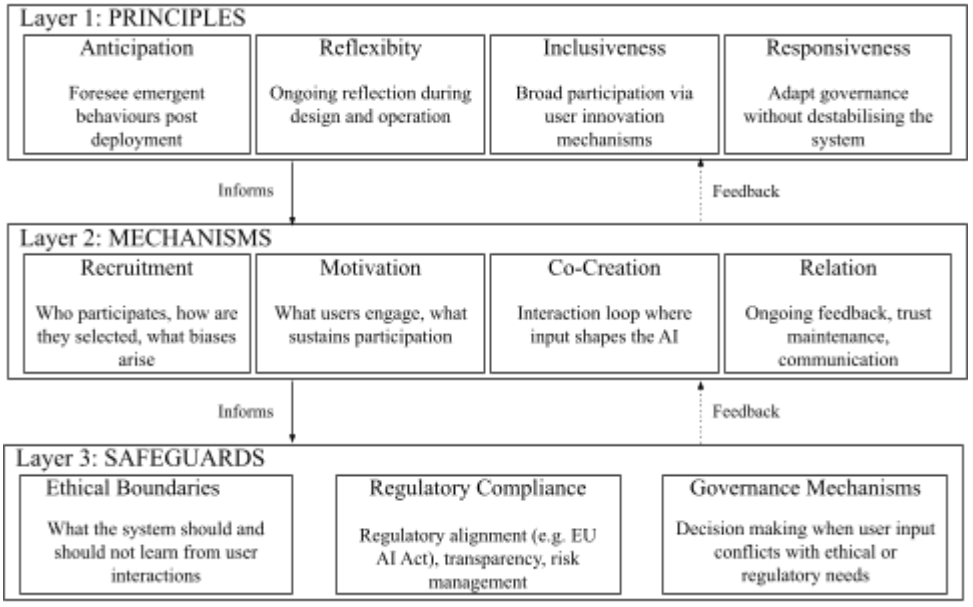

Figure 1 Three-layered responsible AI co-creation framework

Layer 1: Responsible Innovation Principles adapts the AREA dimensions to the adaptive AI context. Anticipation extends beyond pre-deployment risk assessment to encompass

emergent behaviours arising from unanticipated patterns of use, when users begin interacting with the system in ways that were not originally planned for. Reflexivity is framed as an ongoing practice of revisiting design assumptions as the system evolves, rather than a one-time evaluation. It necessitates developers and stakeholders to regularly revisit the assumptions built into the system as it evolves. Inclusiveness, through the mechanisms in Layer 2, can translate what would otherwise remain an abstract commitment into concrete participation practices. Finally, responsiveness refers to the capacity of both the system and the organisation to act on what is learned, such as adjusting governance and system behaviour without introducing instability.

Layer 2: User Involvement Mechanisms structures co-creation across the system's lifecycle. Recruitment concerns who participates, how they are identified and what selection biases may result; in adaptive AI systems, the boundary between general user and co-creator is inherently blurred and must be explicitly managed. Motivation addresses what sustains engagement in the absence of defined project timelines, since open-ended co-creation cannot rely on project-based incentives alone. Co-creation refers to the interaction loop through which user input, whether feedback, usage patterns or generated content, shapes system behaviour. Finally, relation captures the ongoing relationship between users and developers, including transparency about how input is used and the maintenance of trust over time.

Layer 3: Governance Safeguards ensure co-creation remains within ethical and regulatory bounds. Ethical boundaries define what the system may and may not learn or reproduce from user interactions, drawing on ethics-by-design approaches (Brey and Dainow, 2023) and customisable guardrail frameworks (Šekrst et al., 2024). Regulatory compliance addresses alignment with legal requirements such as those under the EU AI Act, including transparency obligations, risk categorisation and accountability structures (Canavese et al., 2024; Cancela-Outeda, 2024). Governance mechanisms address what happens when user input or system behaviour conflicts with ethical or regulatory requirements, specifying who decides, through what process and with what recourse for users (Buhmann and Fieseler, 2022; Prifti et al., 2024).

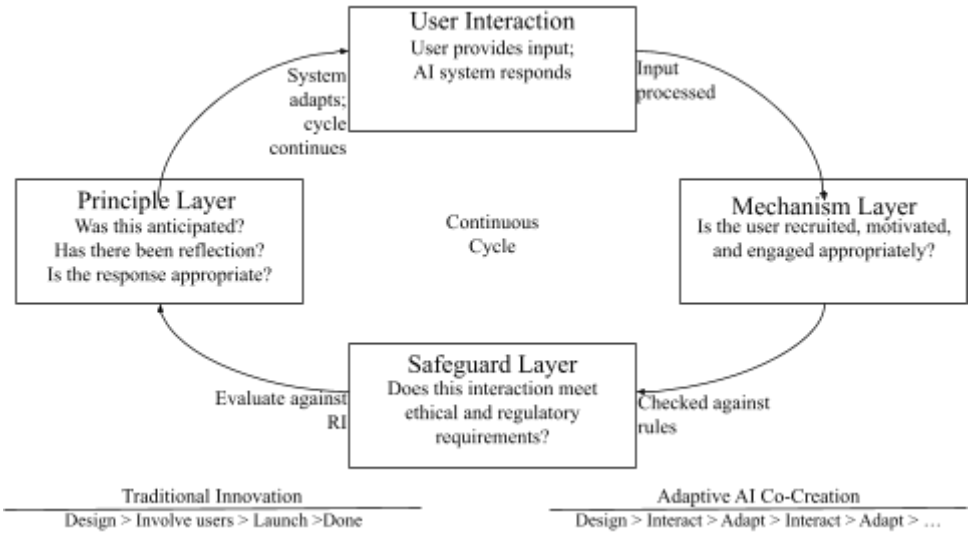

Figure 2 Continuous co-creation cycle

Figure 2 illustrates how the three layers interact in practice. Each user interaction triggers parallel checks across all three layers. The mechanism layer verifies whether participation criteria are being met, for example through a checklist that confirms users are appropriately recruited, engaged and related to. The safeguard layer assesses whether the interaction falls within ethical and regulatory boundaries, through automated guardrails or compliance checks. Finally, the principle layer evaluates whether the interaction was anticipated, whether its implications have been sufficiently reflected upon and whether the system's response is appropriate.

The nature of any response depends on which layer is triggered. Safeguard violations should trigger immediate action, such as blocking a non-compliant response or flagging prohibited content. Mechanism-layer issues, such as the systematic exclusion of a user group, are logged for structural review rather than immediate intervention. Principle-layer gaps, such as unanticipated system behaviours, are similarly recorded and escalated to stakeholders for periodic review. The periodic review aggregates findings across all three layers to identify emerging patterns and determine whether adjustments are warranted. For instance, repeated safeguard blocks on a particular query type might indicate either overly restrictive guardrails or unanticipated user behaviour, each requiring a different corrective response at the principle or mechanism layer.

Applying the framework begins with mapping an AI system's existing practices against each layer: whether design incorporated anticipation of emergent behaviours and mechanisms for ongoing reflexivity (Layer 1); how users are recruited, engaged and related to over time (Layer 2); and whether ethical boundaries, regulatory requirements and conflict resolution procedures are in place (Layer 3). The framework's objective is to provide a coherent process structure for organising the core concerns of responsible AI co-creation. However, it is not a prescriptive evaluation instrument for each layer. The development of layer-specific tools such as detailed checklists, metrics or audit procedures remains an area for future work.

4 Case Application

To assess the framework's practical relevance, an illustrative case study platform was selected against four criteria derived from the framework's scope. The platform must be an adaptive AI system shaped by continuous user interaction, it must involve co-creation in a meaningful sense, it must be sufficiently documented through publicly available sources to permit analysis without proprietary access, and it must present an identifiable ethical or governance dimension relevant to the safeguard layer.

Khan Academy's Khanmigo[1] satisfied these criteria and was selected. Khanmigo is an AI-powered tutoring assistant built on a large language model, integrated into the Khan Academy platform to provide personalised support to students and teachers. Users interact with it continuously, and the system adapts its responses accordingly. Several of its design features are directly relevant to the framework: a Socratic constraint that withholds direct answers in favour of prompts and hints, content moderation calibrated to age-appropriate interaction, and differentiated roles for students and teachers. Extensive public documentation, including official guidelines, blog posts, media coverage and public statements by its founder, supports analysis without requiring internal access.

[1] https://www.khanmigo.ai/

The selection of Khanmigo for this case study does not imply that it is the most representative platform for this purpose. It was chosen because it meets the stated criteria and enables a meaningful application of all three framework layers. Comparative evaluation across multiple platforms or domains would further strengthen the relevance of the framework and is identified as a direction for future work.

4.1 Data Collection

Given that the objective is to evaluate the framework's utility rather than conduct an empirical study, a document-based approach was adopted. No internal design documents, proprietary data, user logs or staff interviews were accessed. The analysis is therefore limited to what is observable from publicly available sources, organised across three source categories corresponding to the framework's layers.

The first category comprises Khan Academy's official documentation. (Khan Academy Blog, 2025a) sets out the platform's responsible AI principles, risk management process and the guardrails implemented for Khanmigo, developed by adapting the nine tenets of the Institute for Ethical AI in Education's framework alongside the AI Risk Management Framework (AI RMF). (Khan Academy Help Center, 2025) provides further detail on safety features, including content moderation, daily usage limits, conversation flagging and parental notification systems. Technical updates on model evaluation and accuracy improvements were also reviewed (Khan Academy Blog, 2025b).

The second category comprises independent third-party evaluations. (Common Sense Media, 2024) assessed Khanmigo against its AI Principles, classifying the platform as low risk and scoring it highly on transparency, safety, trust, learning and privacy. However, the evaluation also identified gaps, including the absence of published transparency reports and instances of unintentionally reinforced bias. A complementary analysis by (Public Services Alliance, 2025) examined Khanmigo's data privacy and security practices, noting a reliance on contractual rather than technical enforcement mechanisms. Further, (Shetye, 2024) evaluated Khanmigo as a language-learning tool using an applied linguistics framework, finding it promising on some criteria but uneven across assessed dimensions, highlighting both the value and the limitations of domain-specific evaluation approaches.

The third category comprises published works and public discourse by the platform's founder. (Khan, 2024), in Brave New Words: How AI Will Revolutionize Education, provides an account of the design decisions, pedagogical philosophy and ethical commitments underpinning Khanmigo. Public talks and media interviews offered supplementary context on the platform's development trajectory and safety architecture (Freethink, 2025).

4.2 Framework Application

This section applies the framework to Khanmigo, evaluating the platform against each layer in terms of what is present, what is partially addressed and where gaps remain.

Layer 1: Principles

On anticipation, Khan Academy demonstrates a structured pre-deployment risk assessment process. Its Responsible AI Framework documents likelihood and impact

evaluations for identified risks, with corresponding mitigation strategies (Khan Academy Blog, 2025a). The risk of harmful use, for instance, was rated high and addressed through moderation APIs and conversation flagging. However, this anticipatory logic is primarily focused on launch-stage safety risks. There is limited publicly available evidence of structured mechanisms for anticipating emergent behaviours arising after deployment, for example novel patterns of student interaction not captured in the initial risk assessment.

Reflexivity is partially present. A Responsible AI Steering Group and Extended Working Group continuously evaluate launched features through demos and feedback loops, indicating an organisational commitment to revisiting assumptions over time (Khan Academy Blog, 2025a). (Common Sense Media, 2024) noted that Khan Academy has been receptive to feedback on identified biases. However, the absence of published transparency reports limits the ability of external stakeholders to assess how reflexive practices are applied in practice.

Inclusiveness, in terms of user participation in shaping the system, is limited. The platform distinguishes between student and teacher roles, and teachers can provide feedback where they review and edit AI-generated content (Freethink, 2025). Students interact with the system continuously, but their role is primarily as learners rather than as co-creators who actively shape system behaviour. There is no publicly documented mechanism through which students or parents contribute to the design or revision of the platform's AI features.

Responsiveness is present at the technical level. Daily usage limits were introduced after extended sessions produced repetitive or unrelated conversations (Khan Academy Help Center, 2025), and the platform has migrated between LLM versions based on performance assessments (Khan Academy Blog, 2025b). However, responsiveness at the governance level, which includes formal processes for adjusting ethical boundaries based on accumulated findings, is not explicitly documented in the publicly available sources reviewed.

Layer 2: Mechanisms

Recruitment is based on access controls that determine who participates. Users must be 18 or older to register independently, and minors access the platform through parental accounts or school district partnerships (Khan Academy Help Center, 2026). While this structures participation, it is oriented toward access management rather than the active recruitment of diverse user groups into co-creation. No publicly documented strategy exists for ensuring representativeness or identifying selection-related biases.

Motivation is supported through pedagogical design. The Socratic constraint encourages active engagement by withholding direct answers (Khan, 2024), and teacher-facing tools reduce administrative workload (Freethink, 2025). However, these mechanisms are designed to sustain learning engagement, not to motivate users to contribute to system improvement or governance.

Co-creation is present in a constrained form. Students shape their immediate learning experience through interaction, and teachers can review and modify AI-generated content. However, Khan Academy has confirmed that neither student nor teacher data is used by its third-party partner OpenAI to train its language models (Common Sense Media, 2024). Co-creation in Khanmigo is therefore better characterised as experiential co-creation, where users shape the immediate interaction, rather than developmental co-creation, where users shape the underlying behaviour of the system.

The relation between users and the platform is maintained through parental and teacher access to chat histories, moderation alerts, feedback features and appeal processes (Khan Academy Help Center, 2025). In-product messaging acknowledges the AI's limitations, including the possibility of inaccurate responses. However, these relational mechanisms are primarily oriented toward oversight and safety, rather than toward building a collaborative relationship in which user feedback systematically informs system development.

Layer 3: Safeguards

Ethical boundaries are clearly defined. The Socratic constraint prevents direct answers in favour of guided questioning (Khan, 2024). Content moderation detects harmful interactions and automatically notifies connected adults when flagged (Khan Academy Help Center, 2025). Further, daily usage limits prevent sessions from drifting from educational purposes, and in-product messaging explicitly cautions that AI responses may be incorrect and should not substitute teacher or parental guidance.

The issue of hallucination is relevant to both this layer and the principle layer. Khan Academy has acknowledged it as one of its most significant ongoing challenges (Khan, 2024). While benchmark datasets and model migrations have been deployed to improve accuracy (Khan Academy Blog, 2025b), there is no mechanism that signals to students when the AI is uncertain about a response. For a platform serving learners as young as grade 3, this represents a notable gap at the intersection of ethical boundaries and anticipation.

Regulatory compliance is addressed through data privacy practices aligned with educational standards. Khan Academy anonymises data sent to OpenAI and prohibits its use for external model training (Common Sense Media, 2024). Access controls, parental oversight mechanisms and educational-use terms of service constitute further compliance measures. However, explicit alignment with instruments such as the EU AI Act is not documented.

A further structural concern at this layer is that Khanmigo is built on OpenAI's GPT infrastructure, meaning student interactions are processed through a third-party system. As (Public Services Alliance, 2025) observed, most of Khanmigo's safety measures rest on contractual agreements and external oversight rather than technical enforcement. This dependency may produce governance gaps as usage scales and regulatory requirements evolve.

Finally, governance mechanisms for handling conflicts between user input and system constraints are partially in place. Moderation triggers follow a defined protocol: the conversation is halted, the user is redirected to community standards, and an adult is notified; persistent violations can result in account suspension (Khan Academy Blog, 2025a). Internal governance structures, such as the Steering Group and Extended Working Group, provide oversight of new features and capabilities. However, no publicly documented process exists through which users can formally challenge governance decisions beyond built-in feedback and appeal features, and the criteria informing those decisions are not transparent to external stakeholders.

This paper was presented at the XXXVII ISPIM Innovation Conference, held in Granada, Spain on 07 June to 10 June 2026. ISBN 978-952-65771-2-8.

4.3 Discussion

Table 1 summarises the evaluation of the Khanmigo platform through the lens of the proposed framework. Khanmigo is strong where the framework asks for protective safeguards, and weak where it asks for participatory mechanisms. Content moderation, usage limits, role separation and the Socratic constraint are present and documented. However, recruitment into co-creation, motivation to contribute beyond learning, and relational channels that feed user input back into system development demonstrate clear gaps. This asymmetry goes beyond Layer 2 and cascades into Layer 1, where inclusiveness remains an abstract commitment, and governance-level responsiveness is similarly limited because there is no structured user voice for governance to respond to. The observation matters because it is unlikely to be specific to Khanmigo. Adaptive AI systems built on third-party LLM infrastructure tend to struggle with similar issues. Safety is enforceable through APIs, filters and contractual clauses; co-creation, on the other hand, requires organisational commitment, user-facing structures and feedback loops that most deployers have no incentive to build.

A distinction implicit in the Khanmigo assessment worth highlighting is that the platform supports what may be called experiential co-creation, where users shape their immediate interaction with the system, but lacks developmental co-creation, where users shape the underlying behaviour of the system. This can be argued as a defensible safety decision for a platform serving minors. At the same time, it is also a decision that removes users from any developmental role in the system they use. The framework does not argue that one form of co-creation is preferable to the other. Rather, its contribution here is to make the trade-off visible. Once named, the absence of developmental co-creation becomes something designers can justify, rather than something that follows by default from the choice of infrastructure.

Third-party model dependency deserves attention for the same reason. When the adaptive layer of a system sits on a model the deploying organisation does not control, Layer 3 degrades in character. Ethical boundaries and regulatory compliance move from technical enforcement to contractual assurance. Khan Academy's safeguards against training-data reuse rest on its agreement with OpenAI, rather than on architectural separation it can audit. This is a category of risk that neither the EU AI Act nor classical RI is currently positioned to address. The Act is structured around deployer obligations and does not fully reach the provider-deployer relationship. Further, classical RI assumes a bounded innovation system in which the actors shaping the technology are identifiable and accountable to one another. Adaptive AI systems built on third-party models satisfy neither assumption. The framework adds an analytical view here by treating third-party dependency as a structural feature of the safeguard layer, rather than an implementation detail.

Taken together, these findings also address the need for co-ordination between RI, user innovation, and AI governance to be more useful in combination than in isolation. Responsible innovation alone would have surfaced the inclusiveness gap, but not the third-party dependency issue, which sits outside its traditional frame of reference. AI governance alone would have surfaced the compliance gap, but not the absence of developmental co-creation, which it tends to treat as a safety feature rather than a design choice. User innovation alone would have surfaced the co-creation asymmetry, but not the regulatory and ethical constraints that make it, in this case, a reasonable one.

Finally, the analysis is based on publicly available sources. No internal design documents, user logs or staff interviews were accessed, and it is possible that practices exist which are not externally visible. A single illustrative case cannot validate a framework; it can only demonstrate that the framework produces coherent and non-trivial findings when applied. Khanmigo also occupies a specific position in the risk landscape. It serves a K-12 educational context with clearly defined ethical commitments and a cautious approach to data use. The framework's behaviour in higher-risk domains such as healthcare, finance and employment, where the trade-offs between safeguards and participation are sharper, remains to be evaluated. These are limitations of the present application rather than of the framework itself, and are taken up as directions for future works.

Table 1 Khanmigo assessment summary

Layer	*Component*	*Assessment*	*Key Finding*
Principles	Anticipation	Partially present	Structured risk assessment at launch; limited evidence of anticipation of post-deployment emergent behaviours
	Reflexivity	Partially present	Internal steering and working groups conduct ongoing evaluation; no published transparency reports for external assessment
	Inclusiveness	Limited	Distinct user roles exist; no documented mechanism for users to participate in shaping AI features or governance
	Responsiveness	Partially present	Technical adaptations based on observed behaviour; governance-level responsiveness not explicitly documented
Mechanisms	Recruitment	Limited	Access controls determine participation; no strategy for ensuring representative or diverse co-creation involvement
	Motivation	Partially present	Pedagogical design sustains learning engagement; motivation to contribute to system improvement is not addressed
	Co-creation	Limited	Experiential co-creation through interaction; no developmental co-creation where users shape underlying system behaviour
	Relation	Partially present	Oversight and safety channels exist; relational mechanisms do not systematically feed user input into system development
Safeguards	Ethical boundaries	Present	Socratic method constraint, content moderation, usage limits, and AI limitation messaging are in place
	AI accuracy	Partially present	General messages on possibilities for AI accuracies; students uninformed about confidence of the AI systems on its answers

Regulatory compliance	Partially present	Data privacy practices aligned with educational standards; alignment with EU AI Act or other related regulations not documented
Third-party LLM dependency	Gap identified	Data commitments rely on contractual rather than technical enforcement; governance extends beyond Khan Academy's direct control
Governance mechanisms	Partially present	Internal governance structures and moderation protocols exist; decision criteria and formal user appeal processes are not transparent

5 Conclusion and Future Work

This paper addresses a gap at the intersection of responsible innovation, user innovation and AI governance, each of which addresses part of the challenges emerging from adaptive AI systems. Responsible innovation offers principles but lacks mechanisms suited to post-deployment adaptation. AI governance embeds safeguards but tends to treat users as subjects of protection rather than as co-creators. User innovation recognises participation but has not extended to systems that evolve through use. The proposed framework brings these three perspectives together in a three-layer structure: RI principles adapted for continuous adaptation, user involvement mechanisms structured around the co-creation lifecycle, and governance safeguards covering ethical boundaries, regulatory compliance and conflict resolution.

Applied to Khanmigo as an illustrative case, the framework identified strengths in content moderation and pedagogical design, and surfaced gaps in anticipating emergent behaviours, enabling developmental co-creation and governing third-party infrastructure dependency. The contribution of this work is threefold. First, the framework offers a coherent structure for organising concerns that have so far been treated in isolation. Second, it introduces the distinction between experiential and developmental co-creation as an analytical lens for adaptive AI. Finally, it identifies third-party model dependency as a structural feature of the safeguard layer, rather than an implementation detail.

The framework is conceptual in nature and requires further extension in future works. The first is operationalisation, through layer-specific indicators, checklists and audit procedures that move the framework from analytical structure toward practitioner use. The second is empirical validation across domains that differ in regulatory environment, user expectations and risk profile, with healthcare, finance and employment as potential areas given their position in the EU AI Act's risk tiers. The third is further development of the experiential and developmental co-creation distinction, which this paper introduces but does not expand upon. Finally, the fourth is governance of third-party model dependencies, where organisations lack control over the underlying infrastructure and existing regulatory instruments provide limited guidance. Taken together, these steps can move the framework from a conceptual contribution toward practitioner implementation.

References

Bessant, J., Oftedal, E.M., Iakovleva, T., 2024. Introduction: Meeting the inclusion

challenge in innovation, in: Meeting the Inclusion Challenge in Innovation. De Gruyter, pp. 1–20.

Brey, P., Dainow, B., 2023. Ethics by design for artificial intelligence. AI and Ethics 4, 1265–1277. https://doi.org/10.1007/s43681-023-00330-4

Buhmann, A., Fieseler, C., 2022. Deep Learning Meets Deep Democracy: Deliberative Governance and Responsible Innovation in Artificial Intelligence. Business Ethics Quarterly 33, 146–179. https://doi.org/10.1017/beq.2021.42

Canavese, D., Ferreira, A., Laborde, R., Kandi, M.A., 2024. Artificial Intelligence Systems in the European Union: Guidelines and Architectures for Compliance-by-Design. Elsevier BV.

Cancela-Outeda, C., 2024. The EU's AI act: A framework for collaborative governance. Internet of Things 27, 101291. https://doi.org/10.1016/j.iot.2024.101291

Catarci, T., Marrella, A., Santucci, G., Sharf, M., Vitaletti, A., Di Lucchio, L., Imbesi, L., Malakuczi, V., 2020. From Consensus to Innovation. Evolving Towards Crowd-based User-Centered Design. International Journal of Human–Computer Interaction 36, 1460–1475. https://doi.org/10.1080/10447318.2020.1753333

Common Sense Media, 2024. Khanmigo AI Product Review [WWW Document]. Common Sense Media. URL https://www.commonsensemedia.org/ai-ratings/khanmigo (accessed 4.4.26).

Fraaije, A., Flipse, S.M., 2019. Synthesizing an implementation framework for responsible research and innovation. Journal of Responsible Innovation 7, 113–137. https://doi.org/10.1080/23299460.2019.1676685

Freethink, 2025. Sal Khan wants to give every student on Earth a personal AI tutor. Freethink Media.

Gago, D., Rubalcaba, L., 2024. Co-creation for public innovation: The role of living labs and user-engagement methodologies. Public Money & Management 45, 604–614. https://doi.org/10.1080/09540962.2024.2425423

Herrmann, H., 2023. What's next for responsible artificial intelligence: a way forward through responsible innovation. Heliyon 9, e14379. https://doi.org/10.1016/j.heliyon.2023.e14379

Huang, J.H., Iakovleva, T.A., Bessant, J., 2024. FOSTERING USER INVOLVEMENT IN COLLABORATIVE INNOVATION SPACES: INSIGHTS FROM LIVING LABS. International Journal of Innovation Management 28.

https://doi.org/10.1142/s1363919624500324

Iakovleva, T., Oftedal, E., Bessant, J., 2021. Changing Role of Users—Innovating Responsibly in Digital Health. Sustainability 13, 1616. https://doi.org/10.3390/su13041616

Khan Academy Blog, 2025a. Khan Academy's Framework for Responsible AI in Education [WWW Document]. Khan Academy Blog. URL https://blog.khanacademy.org/khan-academys-framework-for-responsible-ai-in-education/ (accessed 4.4.26).

Khan Academy Blog, 2025b. Khanmigo Math Computation and Tutoring Updates [WWW Document]. Khan Academy Blog. URL https://blog.khanacademy.org/khanmigo-math-computation-and-tutoring-updates/ (accessed 4.4.26).

Khan Academy Help Center, 2026. I meet all of the Khanmigo requirements, but it's saying that I'm under 18 years old. What should I do? [WWW Document]. Khan Academy Help Center. URL https://support.khanacademy.org/hc/en-us/articles/19682906328461-I-meet-all-of-the-Khanmigo-requirements-but-it-s-saying-that-I-m-under-18-years-old-What-should-I-do (accessed 26).

Khan Academy Help Center, 2025. What safety features does Khanmigo have? [WWW Document]. Khan Academy Help Center. URL https://support.khanacademy.org/hc/en-us/articles/14394814244365-What-safety-features-does-Khanmigo-have (accessed 4.4.26).

Khan, S., 2024. Brave New Words: How AI Will Revolutionize Education (and Why That's a Good Thing). Random House.

Laitio, A., Nätti, S., 2023. How embedded lead users can contribute to innovation process: A systematic literature review. Journal of Innovation Management 11, 157–172. https://doi.org/10.24840/2183-0606_011.002_0006

Owen, R., Macnaghten, P., Stilgoe, J., 2012. Responsible research and innovation: From science in society to science for society, with society. Science and Public Policy 39, 751–760. https://doi.org/10.1093/scipol/scs093

Prifti, K., Morley, J., Novelli, C., Floridi, L., 2024. Regulation by Design: Features, Practices, Limitations, and Governance Implications. Minds and Machines 34. https://doi.org/10.1007/s11023-024-09675-z

Public Services Alliance, 2025. Student data privacy and security: Khan Academy's Khanmigo [WWW Document]. Public Services Alliance. URL https://publicservicesalliance.org/2025/10/22/student-data-privacy-and-security-khan-academys-khanmigo/ (accessed 4.4.26).

Šekrst, K., McHugh, J., Cefalù, J.R., 2024. AI Ethics by Design: Implementing Customizable Guardrails for Responsible AI Development. Qeios Ltd.

Shetye, S., 2024. An Evaluation of Khanmigo, a Generative AI Tool, as a Computer-Assisted Language Learning App. Studies in Applied Linguistics and TESOL 24. https://doi.org/10.52214/salt.v24i1.12869

Stilgoe, J., Owen, R., Macnaghten, P., 2013. Developing a framework for responsible innovation. Research Policy 42, 1568–1580. https://doi.org/10.1016/j.respol.2013.05.008

Ulnicane, I., Mahfoud, T., Salles, A., 2022. Experimentation, learning, and dialogue: an RRI-inspired approach to dual-use of concern. Journal of Responsible Innovation 10. https://doi.org/10.1080/23299460.2022.2094071