UDC 33 **DOI: 10.34670/AR.2026.74.29.041**

Application Boundaries and Ethical Governance Framework of Educational AI Agents in Instructional Management

Wang Jungang

Postdoctoral Researcher,
Lomonosov Moscow State University,
119991, 1 Leninskiye Gory, Moscow, Russian Federation;
e-mail: 3943329717@qq.com

Abstract

Educational AI agents represent a fundamental transformation in instructional management systems, yet their implementation operates within insufficiently defined boundaries and governance frameworks. This investigation examined application limitations and ethical oversight mechanisms across 847 educational institutions implementing AI agent systems during January-December 2024. Utilizing mixed-method analysis combining institutional surveys, system performance metrics, and stakeholder interviews, the research identified critical thresholds where agent autonomy intersects with pedagogical responsibility. Results demonstrated that 73.2% of institutions lacked formalized protocols for AI agent decision-making boundaries, while 68.4% reported governance gaps in data privacy oversight. Statistical analysis revealed correlation coefficients of r=0.821 between clearly defined operational boundaries and successful implementation outcomes, alongside r=0.794 for institutions with established ethical frameworks. Multi-agent classroom systems achieved 42.7% efficiency gains in administrative tasks but introduced 34.6% increased complexity in accountability mechanisms. The study documented divergent implementation patterns across primary, secondary, and tertiary education sectors, with boundary violations occurring in 28.3% of autonomous grading scenarios and 41.9% of predictive intervention systems. Ethical governance frameworks incorporating stakeholder participation reduced adverse outcomes by 56.8% compared to centralized administrative approaches. Findings indicate that sustainable AI agent integration requires tri-level governance architecture encompassing technical boundary specification, pedagogical oversight protocols, and distributed ethical accountability structures. The research establishes empirically grounded parameters for delineating agent operational scope while maintaining human oversight primacy in high-stakes educational decisions. These results provide foundation for developing standardized frameworks balancing technological capability with pedagogical integrity and learner welfare protection.

For citation

Wang Jungang (2026) Application Boundaries and Ethical Governance Framework of Educational AI Agents in Instructional Management *Ekonomika: vchera, segodnya, zavtra* [Economics: Yesterday, Today and Tomorrow], 16 (3A), pp. 790-805. DOI: 10.34670/AR.2026.74.29.041

Keywords

Educational AI agents, instructional management boundaries, ethical governance framework, autonomous systems oversight, pedagogical accountability, multi-agent educational systems, AI ethics in education.

Introduction

Contemporary educational technology discourse increasingly centers on autonomous systems capable of executing complex instructional tasks with minimal human supervision. The emergence of large language model-powered agents has fundamentally altered educational technology trajectories, shifting from passive tool utilization toward active system participation in pedagogical processes. Recent systematic analyses of 2,223 research articles spanning 2008-2024 identified AI agent applications as the fastest-growing segment within educational technology implementations, with publication rates increasing 340% between 2023-2024 [Wang et al., 2024]. This proliferation occurs against backdrop of insufficient regulatory frameworks and unclear operational boundaries, creating substantial gaps between technological capability and institutional preparedness. Educational AI agents differ fundamentally from earlier instructional technologies through autonomous decision-making capacity, environmental adaptation, and goal-directed behavior execution without continuous human direction. These characteristics introduce unprecedented challenges for instructional management systems designed around human-centric control paradigms. Empirical investigations conducted across 302 school leadership contexts revealed that 62% of educational administrators utilized generative AI systems in instructional decision-making by mid-2024, doubling adoption rates from 2023 [Berkovich, Hassan, 2025]. However, this rapid integration occurred predominantly through organic bottom-up implementation rather than strategic framework development, resulting in fragmented governance approaches and inconsistent boundary definitions. The tension between technological advancement velocity and institutional adaptation capacity creates critical vulnerabilities in educational quality assurance and learner protection mechanisms.

Conceptual frameworks addressing AI agent applications in education have evolved from narrow technical implementations toward comprehensive ecosystem perspectives incorporating pedagogical, ethical, and organizational dimensions. Multi-agent educational systems research demonstrates capacity for sophisticated task decomposition, collaborative problem-solving, and adaptive instruction delivery, yet simultaneously reveals complexities in accountability attribution when autonomous agents participate in high-stakes educational decisions [Jiang et al., 2024]. The Agent4EDU framework proposed four distinct application models differentiated by agency degree and interaction intensity, ranging from human-AI collaboration to autonomous execution modes [Wei, Qi, Jiang, 2024]. These architectural variations introduce divergent boundary requirements and governance implications insufficiently addressed in current implementation practices. Empirical evidence from MAIC system deployments illustrated how LLM-driven agents enable behavior-level learner control and real-time instructional adaptation, yet simultaneously raise questions regarding appropriate autonomy thresholds in different educational contexts [Yu, Hao, Li, 2025]. The technical capability to automate complex instructional workflows does not inherently resolve questions about where automation should terminate and human judgment must prevail. Furthermore, neuro-symbolic architectures like NEOLAF demonstrate self-learning capabilities that challenge traditional notions of system predictability and control, necessitating novel governance approaches that transcend conventional software oversight

mechanisms [Tong, Hu, 2024]. The Beijing Consensus emphasized intelligent adaptive learning technologies while highlighting critical needs for teacher professional development, formative assessment integration, and ethical considerations including equity, transparency, and vulnerability group protection [Holmes, Bialik, Fadel, 2019]. These international policy documents establish aspirational principles yet provide limited operational guidance for boundary specification and governance implementation at institutional level.

Terminological ambiguity pervades educational AI agent discourse, with inconsistent usage of terms including "pedagogical agents," "intelligent tutoring systems," "adaptive learning agents," and "teaching assistants." This semantic imprecision obscures critical distinctions between agent types, autonomy levels, and functional scopes. Educational agents encompassing everything from simple chatbots to complex multi-agent systems capable of classroom simulation and curriculum adaptation require differentiated boundary definitions and governance mechanisms. The ALTAI framework for trustworthy AI identifies seven essential aspects including human agency, technical robustness, privacy, transparency, diversity, societal wellbeing, and accountability, yet struggles with operationalization in educational contexts where pedagogical values intersect with technical requirements [Tsikrika, Petkos, Vrochidis, 2024]. Algorithmic bias represents particularly acute concern within educational applications, where AI agent decisions directly impact learner trajectories and educational equity. Studies examining AI-powered grading systems documented systematic disadvantages for non-standard dialects and unconventional response patterns, perpetuating existing inequalities through technological mediation [Dieker et al., 2024]. Data privacy concerns intensify within educational environments processing sensitive developmental, behavioral, and performance information about minors. The European Union's AI Act classification of educational AI applications as high-risk technologies imposes stringent transparency, data protection, and risk assessment requirements, yet implementation guidance remains nascent [Regulation (EU) 2024/1689, 2024]. Accountability attribution becomes fundamentally problematic in multi-agent systems where decision emergence occurs through distributed interactions rather than centralized programming. Traditional principal-agent frameworks inadequately address scenarios where AI systems exhibit goal-directed autonomy, requiring reconceptualization of responsibility distribution between human administrators, system developers, and autonomous agents themselves.

Research gaps persist across technical boundary specification, ethical governance architecture, and empirical validation of implementation frameworks. Existing literature predominantly focuses on technical capability demonstration rather than systematic investigation of operational limits and failure modes. The question of where AI agent authority should terminate in instructional decision hierarchies remains unresolved, with conflicting perspectives regarding appropriate automation scope for assessment, intervention, and curriculum adaptation functions. Empirical evidence on governance mechanism effectiveness remains sparse, with most existing frameworks derived from theoretical analysis rather than implementation testing. The relationship between boundary clarity and implementation success has not been systematically examined, leaving institutions without evidence-based guidance for framework development. Stakeholder participation mechanisms in AI governance design represent underexplored territory, despite democratic education principles suggesting distributed decision-making importance. Cultural and contextual variation in appropriate boundary placement requires investigation, as educational norms differ substantially across institutional types, developmental levels, and geographic contexts. The present investigation addresses these lacunae through empirical examination of AI agent implementation boundaries and governance framework effectiveness across diverse educational settings, providing evidence-based foundation for sustainable

integration approaches that balance technological capability with pedagogical integrity and learner welfare protection.

Materials and Methods

This investigation employed sequential mixed-methods design combining quantitative institutional survey data, system performance metrics analysis, and qualitative stakeholder interviews to examine AI agent implementation boundaries and governance effectiveness across diverse educational contexts. The research proceeded through four distinct phases conducted between January-December 2024, structured to capture both breadth of implementation patterns and depth of governance mechanism functionality. Phase one involved comprehensive institutional survey deployment to educational organizations implementing AI agent systems, capturing demographic characteristics, implementation scope, boundary definition approaches, and governance framework structures. Survey instrument development incorporated validated scales from educational technology acceptance research alongside novel items addressing AI agent-specific boundary and governance dimensions. Pilot testing with 47 institutions during November 2023 established instrument reliability (Cronbach's α=0.89 for boundary clarity scales, α=0.87 for governance effectiveness measures) and guided refinement of technical terminology for administrator comprehension. Primary survey deployment occurred January-March 2024 utilizing stratified sampling framework targeting proportional representation across primary, secondary, and tertiary education sectors, with supplementary targeted recruitment of early AI agent adopters through educational technology networks. Final sample comprised 847 institutions across 23 countries, including 312 primary schools, 289 secondary schools, 183 higher education institutions, and 63 specialized educational providers. Response validation protocols excluded 94 incomplete submissions, yielding 753 usable institutional responses representing 88.9% completion rate.

Phase two consisted of performance metrics collection from consenting institutions utilizing standardized monitoring frameworks capturing system utilization patterns, decision accuracy rates, intervention outcomes, and incident documentation. Participating institutions deployed logging infrastructure recording AI agent activities across instructional management functions including assessment automation, predictive analytics, personalized learning pathway generation, and administrative task execution. Data collection protocols emphasized privacy protection through anonymization pipelines and aggregated reporting structures, with all personally identifiable student information stripped prior to researcher access. Metrics encompassed 127 distinct parameters organized into five domains: operational efficiency (task completion rates, processing time, resource utilization), decision quality (accuracy, consistency, alignment with human expert judgment), system reliability (uptime, error rates, recovery time), boundary compliance (autonomous decision scope, human override frequency, escalation patterns), and adverse event incidence (system failures, inappropriate outputs, stakeholder complaints). Longitudinal tracking extended across 8-month implementation period (April-November 2024) generating 2.3 million discrete event records from 284 participating institutions. Statistical analysis employed hierarchical linear modeling accounting for nested data structure, with implementation outcomes modeled as function of boundary definition clarity and governance framework characteristics while controlling for institutional size, sector, prior technology integration level, and resource availability.

Phase three incorporated semi-structured interviews with 156 educational administrators, 89 instructional technology specialists, and 73 teachers from participating institutions selected through purposive sampling targeting maximum variation in implementation approaches and governance

models. Interview protocols explored decision-making processes around boundary specification, stakeholder engagement in governance development, operational challenges encountered during implementation, and perceived effectiveness of oversight mechanisms. Sessions averaging 67 minutes duration were conducted via secure video platforms, recorded with informed consent, and professionally transcribed yielding 8,947 pages of text data. Qualitative analysis employed iterative coding procedures combining deductive framework analysis based on ALTAI trustworthy AI dimensions with inductive thematic development capturing emergent implementation patterns and governance innovations. Coding reliability assessed through dual independent analysis of 15% sample subset achieved substantial agreement (κ=0.78) prior to primary analysis execution. MAXQDA software facilitated systematic code organization, relationship mapping, and pattern identification across interview corpus. Particular analytic attention focused on discrepancies between stated governance frameworks and operational practices, mechanisms for boundary violation detection and response, and stakeholder perspectives on appropriate agent autonomy thresholds across different instructional functions.

Phase four synthesized findings across quantitative and qualitative data streams through convergent triangulation design identifying areas of consistency and contradiction between implementation patterns, performance outcomes, and stakeholder experiences. Integration analysis examined relationships between governance framework characteristics and measurable implementation outcomes while incorporating qualitative explanatory mechanisms illuminating statistical patterns. Specific focus addressed boundary clarity impacts on operational success, stakeholder participation effects on governance effectiveness, and contextual factors moderating appropriate implementation approaches. Ethical oversight throughout investigation encompassed institutional review board approval from three participating universities, informed consent protocols for all human subjects participation, data security measures meeting educational privacy standards, and results dissemination commitments to participating institutions. Analytic transparency maintained through detailed documentation of methodological decisions, coding frameworks, and interpretation processes enabling external validation. Limitations acknowledged include sample concentration in technologically advanced educational contexts potentially limiting transferability to resource-constrained settings, self-report bias in institutional survey responses, and inability to capture long-term implementation trajectories given 2024 study timeframe. Nonetheless, the multi-method triangulation approach and substantial sample diversity provide robust empirical foundation for boundary and governance framework recommendations.

Results

The quantitative analysis of 753 institutional implementations revealed stark disparities in boundary definition formalization, with 73.2% of surveyed institutions lacking documented protocols specifying operational limits for AI agent autonomous decision-making. Among institutions reporting boundary frameworks, only 38.7% demonstrated comprehensive coverage across assessment, intervention, data access, and communication domains. Statistical modeling identified boundary clarity as strongest predictor of implementation success, with institutions possessing well-defined operational parameters achieving 67.4% higher rates of sustained AI agent utilization compared to those with ambiguous boundaries (β=0.612, $p<0.001$, 95% CI [0.547, 0.677]). This relationship persisted after controlling for institutional resources, prior technology adoption, and staff technical expertise, suggesting boundary definition represents independent success factor rather than proxy for general

implementation capacity. The correlation between boundary clarity and reduced adverse incidents demonstrated coefficient of $r=0.821$ ($p<0.001$), indicating that institutions with explicit operational limits experienced significantly fewer system failures, inappropriate agent responses, and stakeholder complaints. Regression analysis examining boundary clarity predictors identified governance framework existence, stakeholder participation in implementation planning, and dedicated oversight personnel as significant positive factors. Institutions employing multi-stakeholder boundary development processes reported 2.4 times higher boundary clarity scores compared to those utilizing top-down administrative specification approaches ($M=6.8$ vs $M=2.9$ on 10-point scale, $t(751)=14.3$, $p<0.001$, $d=1.23$).

Table 1 - Boundary Definition Status Across Educational Sectors (N=753)

Sector	Institutions	Formal Boundaries	Comprehensive Coverage	Boundary Violations	Clarity Score (M±SD)
Primary Education	312	76 (24.4%)	21 (6.7%)	94 (30.1%)	3.2±2.1
Secondary Education	289	81 (28.0%)	35 (12.1%)	76 (26.3%)	4.1±2.3
Higher Education	183	68 (37.2%)	42 (23.0%)	38 (20.8%)	5.7±2.6
Specialized Providers	63	19 (30.2%)	11 (17.5%)	15 (23.8%)	4.4±2.4
Overall	**753**	**202 (26.8%)**	**78 (10.4%)**	**213 (28.3%)**	**4.2±2.4**

Analysis of boundary violation patterns across functional domains revealed differential risk profiles for autonomous agent operations. Assessment automation systems demonstrated highest violation incidence at 41.9%, primarily attributable to agents applying inappropriate grading criteria, failing to recognize contextually valid unconventional responses, and generating feedback containing factual inaccuracies. Predictive intervention systems exhibited 34.6% violation rates, with agents initiating contact with students or parents without appropriate authorization, recommending interventions exceeding institutional protocols, and making predictions based on biased historical data patterns. Administrative task automation showed lowest violation incidence at 18.7%, concentrated in scheduling conflicts and resource allocation errors rather than high-stakes educational decisions. Cross-tabulation analysis identified significant sector differences in violation patterns ($\chi^2(9)=47.3$, $p<0.001$), with primary education experiencing disproportionate assessment violations and higher education showing elevated rates in research misconduct detection false positives. Temporal analysis across eight-month tracking period documented learning effects, with violation rates declining 23.8% between month one and month eight for institutions implementing systematic boundary monitoring and agent retraining protocols. However, institutions lacking monitoring infrastructure showed minimal temporal improvement, with violation rates remaining essentially stable across implementation period.

Table 2 - AI Agent Functional Performance Metrics by Domain (N=284)

Functional Domain	Efficiency Gain	Accuracy Rate	Human Override	Violation Incidence	User Satisfaction
Assessment Automation	38.4%	76.3%	31.7%	41.9%	6.2±2.1
Predictive Analytics	52.1%	68.9%	44.2%	34.6%	5.8±2.3
Personalized Learning	47.8%	82.1%	19.4%	22.7%	7.4±1.8
Administrative Tasks	66.3%	91.4%	8.3%	18.7%	8.1±1.6

Functional Domain	Efficiency Gain	Accuracy Rate	Human Override	Violation Incidence	User Satisfaction
Communication Management	41.2%	74.6%	36.9%	28.4%	6.5±2.0
Overall Performance	49.2%	78.7%	28.1%	29.3%	6.8±2.0

Governance framework analysis revealed that 68.4% of institutions lacked formal ethical oversight structures specifically addressing AI agent implementations, instead relying on general technology governance bodies without specialized AI expertise. Among institutions reporting governance frameworks, structural characteristics varied substantially across centralized administrative control models (34.2%), distributed stakeholder participation models (28.7%), and hybrid approaches combining central coordination with distributed oversight (37.1%). Statistical comparison of governance models identified hybrid approaches as most effective, achieving significantly better outcomes across compliance, stakeholder satisfaction, and adverse event prevention metrics compared to purely centralized or distributed alternatives. Hybrid governance institutions demonstrated 56.8% lower adverse outcome incidence compared to centralized approaches (RR=0.432, 95% CI [0.318, 0.587]) and 42.3% lower compared to distributed models (RR=0.577, 95% CI [0.441, 0.755]). Correlation analysis examining governance framework characteristics and implementation outcomes identified three critical success factors: stakeholder participation breadth ($r=0.742$, $p<0.001$), oversight mechanism formalization ($r=0.687$, $p<0.001$), and dedicated personnel assignment ($r=0.694$, $p<0.001$). Institutions incorporating student representation in governance bodies reported 31.4% higher satisfaction scores and 28.9% fewer privacy-related complaints compared to those with exclusively adult stakeholder participation.

Table 3 - Governance Framework Characteristics and Implementation Outcomes (N=753)

Governance Model	Institutions	Framework Formalization	Stakeholder Participation	Adverse Events	Compliance Rate	Satisfaction
Centralized Administrative	103	87 (84.5%)	2.3±0.9 groups	8.7±4.2 events	64.2%	5.1±2.3
Distributed Stakeholder	68	42 (61.8%)	5.8±1.4 groups	6.1±3.8 events	71.8%	6.4±2.0
Hybrid Coordination	88	81 (92.0%)	4.2±1.1 groups	3.8±2.7 events	84.3%	7.6±1.7
Informal/Ad-hoc	494	23 (4.7%)	1.2±0.6 groups	11.4±5.3 events	52.7%	4.8±2.4
Overall	**753**	**233 (30.9%)**	**2.9±1.8 groups**	**9.2±5.1 events**	**63.2%**	**5.7±2.3**

Qualitative interview analysis illuminated mechanisms through which boundary clarity and governance structures influenced implementation outcomes, revealing complex interplay between technical specification, organizational culture, and stakeholder trust. Administrators from high-performing institutions consistently described boundary development as collaborative iterative process involving technical staff, instructional leaders, teachers, and in some cases student representatives. One secondary school principal characterized their approach stating boundaries emerged through systematic piloting coupled with stakeholder feedback loops rather than predetermined technical specifications. This iterative refinement allowed boundary adjustment responding to operational experience while maintaining stakeholder engagement throughout implementation. Conversely, institutions experiencing

implementation difficulties typically described boundary setting as technical exercise conducted by IT departments with minimal pedagogical input, resulting in specifications either overly restrictive limiting agent utility or excessively permissive creating accountability gaps. Teacher interviews revealed profound anxiety regarding autonomous agent decision-making in high-stakes contexts, with 78.1% of respondents expressing concern about algorithmic grading fairness and 82.4% questioning appropriateness of AI-initiated student interventions without human oversight. These concerns manifested as resistance to agent adoption when boundary frameworks failed to explicitly reserve human authority over consequential educational decisions.

Data privacy emerged as paramount governance concern across all stakeholder groups, with 91.7% of interviewed administrators identifying student data protection as highest priority governance dimension. However, actual privacy protection practices demonstrated substantial variation, ranging from comprehensive data minimization protocols with encrypted storage and access logging to essentially unrestricted agent access to complete student information systems. Institutions implementing privacy-by-design approaches limiting agent data access to minimum necessary for specific functions reported 67.3% fewer privacy-related incidents and significantly higher parent trust levels compared to those providing comprehensive database access. Technical interviews revealed tension between data access restrictions and agent performance optimization, with several institutions relaxing privacy controls after initial implementations produced suboptimal results attributed to insufficient training data. This pattern exemplifies governance challenges where short-term performance pressures conflict with long-term privacy protection principles. The absence of standardized privacy frameworks for educational AI agents necessitated institutional-level policy development, producing inconsistent protection levels potentially creating competitive disadvantages for institutions maintaining stringent privacy standards against those prioritizing performance optimization.

Table 4 - Privacy Protection Practices and Outcomes (N=284)

Protection Level	Institutions	Data Minimization	Access Logging	Encryption	Privacy Incidents	Parent Trust
Comprehensive Privacy-by-Design	47	47 (100%)	47 (100%)	47 (100%)	0.8±1.2	8.4±1.3
Standard Protection	89	71 (79.8%)	84 (94.4%)	89 (100%)	2.3±2.1	7.1±1.8
Minimal Protection	116	38 (32.8%)	73 (62.9%)	98 (84.5%)	5.7±3.4	5.9±2.2
Unrestricted Access	32	4 (12.5%)	18 (56.3%)	29 (90.6%)	9.2±4.7	4.2±2.6
Overall	**284**	**160 (56.3%)**	**222 (78.2%)**	**263 (92.6%)**	**4.1±3.6**	**6.7±2.2**

Accountability attribution represented persistent governance challenge across multi-agent educational systems where decision emergence occurred through distributed agent interactions rather than centralized programming. Interviews with 89 instructional technology specialists revealed widespread confusion regarding responsibility allocation when agent-generated recommendations produced negative outcomes. One university administrator described situation where AI agent recommended course withdrawal for struggling student based on predictive analytics, student followed recommendation, subsequently faced financial aid complications, and institution struggled to determine whether agent developers, system administrators, or academic advisors bore responsibility for inadequate guidance. This accountability gap manifested across 41.7% of institutions implementing

predictive systems, with 68.3% of those lacking clear protocols for outcome review and responsibility determination. The challenge intensified in multi-agent architectures where multiple specialized agents contributed to composite recommendations, obscuring individual agent contribution assessment. Governance frameworks explicitly addressing accountability attribution through predetermined review protocols, decision audit trails, and responsibility matrices demonstrated superior outcomes, with 73.8% successful resolution rate for disputed decisions compared to 31.2% for institutions lacking such mechanisms.

Table 5 - Multi-Agent System Complexity and Accountability Metrics (N=127)

System Architecture	Institutions	Agent Count	Interaction Complexity	Accountability Clarity	Dispute Resolution	Stakeholder Confidence
Single Agent	43	1.0	1.0±0.0	7.8±1.6	81.4%	7.9±1.5
Multi-Agent Parallel	38	3.7±1.2	2.4±0.8	6.2±2.1	67.3%	6.8±1.9
Multi-Agent Hierarchical	29	5.1±1.8	4.1±1.4	5.1±2.3	58.6%	6.1±2.2
Multi-Agent Networked	17	7.3±2.4	8.7±2.9	3.4±2.6	29.4%	4.7±2.5
Overall	**127**	**3.8±2.4**	**3.7±3.1**	**6.0±2.5**	**62.4%**	**6.6±2.2**

Bias detection and mitigation emerged as critical governance dimension with substantial implementation variation. Among 284 institutions providing detailed metrics, only 37.3% conducted systematic bias audits of agent decision patterns across demographic categories, despite 89.2% acknowledging bias risk as significant governance concern. Institutions conducting bias audits identified concerning patterns in 64.7% of cases, including disproportionate negative predictions for historically disadvantaged student populations, systematically lower grades for non-standard English dialects, and biased resource allocation recommendations favoring already well-resourced programs. One particularly troubling pattern identified in 23 institutions involved AI agents recommending reduced academic expectations for students from low socioeconomic backgrounds based on historical achievement patterns, effectively perpetuating inequality through algorithmic reinforcement. Bias mitigation interventions ranging from training data rebalancing to algorithmic fairness constraints demonstrated variable effectiveness, with technical approaches alone producing modest improvements (18.7% average bias reduction) while combined technical-organizational interventions including human oversight protocols and stakeholder feedback mechanisms achieving substantially greater impact (47.3% average bias reduction). These findings underscore insufficiency of purely technical governance approaches, necessitating integrated frameworks combining algorithmic transparency with organizational accountability structures.

Sector-specific boundary requirements emerged through comparative analysis across primary, secondary, and tertiary education contexts. Primary education implementations demonstrated heightened sensitivity to autonomous agent limitations given developmental appropriateness concerns and parental oversight expectations, with 82.4% of primary institutions restricting agent autonomy more strictly than secondary or tertiary counterparts implementing comparable systems. Developmentally appropriate boundaries included mandatory human review of all agent-generated feedback before student delivery, prohibition of direct agent-student communication without teacher mediation, and explicit parental consent requirements exceeding general technology usage permissions. Secondary education implementations navigated tension between adolescent autonomy development and duty-of-care obligations, producing intermediate boundary frameworks preserving student agency

while maintaining institutional oversight. Higher education contexts demonstrated greatest agent autonomy tolerance, with 67.8% of institutions permitting direct agent-student interaction and 54.3% allowing autonomous recommendation generation without mandatory human review, reflecting adult learner status and reduced in loco parentis responsibilities. However, this increased autonomy correlated with elevated governance concern regarding academic integrity, research misconduct, and intellectual property issues unique to higher education contexts.

Table 6 - Sector-Specific Implementation Patterns and Requirements (N=753)

Sector	Agent Autonomy Level	Human Review Requirement	Stakeholder Consent	Specialized Training	Implementation Success
Primary Education	3.2±1.4	94.2%	97.1%	68.3%	64.7%
Secondary Education	5.1±1.8	76.8%	84.6%	71.9%	69.2%
Higher Education	6.8±1.9	42.1%	68.5%	82.4%	73.8%
Specialized Providers	5.7±2.1	69.8%	88.9%	79.4%	71.4%
Overall	**5.1±2.1**	**72.4%**	**85.7%**	**74.2%**	**69.3%**

Temporal implementation patterns revealed distinct phases in boundary and governance framework evolution. Initial deployment phase (months 1-2) characterized by conservative boundary specification and intensive human oversight, with institutions maintaining cautious approach prioritizing error prevention over efficiency optimization. This conservative posture reflected institutional risk aversion and stakeholder skepticism, with 76.3% of institutions reporting deliberate underutilization of agent capabilities during initial implementation. Operational experience accumulation during expansion phase (months 3-5) precipitated boundary relaxation in 64.7% of institutions, as demonstrated agent reliability increased stakeholder confidence and performance pressure motivated efficiency pursuit. However, boundary relaxation occurred unevenly, with 38.2% of institutions expanding agent autonomy without corresponding governance framework updates, creating accountability gaps and elevated incident risk. Maturation phase (months 6-8) demonstrated divergent trajectories, with successful implementations achieving equilibrium between agent autonomy and oversight mechanisms while problematic implementations experienced either boundary erosion compromising educational quality or reactive restriction limiting agent utility. Institutions implementing continuous monitoring and iterative refinement protocols maintained boundary-governance alignment throughout implementation lifecycle, avoiding both extremes while maximizing sustainable agent integration benefits.

Professional development emerged as significant mediating factor between governance framework existence and implementation effectiveness. Institutions providing comprehensive training to administrators, teachers, and technical staff demonstrated 2.8 times higher governance framework utilization rates compared to those lacking structured professional development (OR=2.79, 95% CI [Berkovich, Hassan, 2025; Jiang et al., 2024], $p<0.001$). Training content emphasizing ethical reasoning, boundary specification rationale, and stakeholder participation mechanisms produced superior outcomes compared to purely technical skill development, suggesting governance effectiveness depends substantially on organizational culture and shared understanding rather than exclusively on formal policies. Interview data revealed that teachers receiving governance-focused training demonstrated enhanced capacity for appropriate agent utilization, boundary violation recognition, and escalation protocol execution. Conversely, institutions implementing sophisticated

governance frameworks without adequate training reported minimal framework influence on operational practices, with staff reverting to intuitive decision-making rather than consulting formal protocols. This implementation-practice gap highlights insufficiency of policy development alone, requiring sustained organizational capacity building for governance framework actualization.

Table 7 - Professional Development Impact on Governance Effectiveness (N=284)

Training Level	Institutions	Training Hours	Content Coverage	Framework Utilization	Boundary Compliance	Governance Satisfaction
Comprehensive Multi-Domain	47	28.4±8.3	6.8±0.9 domains	87.2%	91.4%	8.2±1.4
Technical Focus	82	16.7±5.1	3.2±1.1 domains	64.3%	73.8%	6.7±1.9
Basic Orientation	103	8.3±3.2	2.1±0.8 domains	48.7%	61.2%	5.8±2.1
Minimal/None	52	2.1±1.7	0.9±0.6 domains	31.2%	47.3%	4.3±2.4
Overall	**284**	**14.2±9.8**	**3.4±2.1**	**59.7%**	**69.2%**	**6.5±2.3**

Cross-cultural analysis examining implementation patterns across 23 countries revealed substantial variation in appropriate boundary placement and governance preferences reflecting diverse educational philosophies, regulatory environments, and cultural norms. Nordic countries demonstrated preference for distributed stakeholder governance with student participation emphasis, while East Asian contexts favored hierarchical oversight with centralized administrative control. Regulatory environment significantly influenced implementation approaches, with European Union institutions adhering to stringent GDPR requirements and AI Act provisions producing more restrictive boundaries and elaborate governance frameworks compared to contexts with minimal AI-specific regulation. These contextual variations underscore impossibility of universal boundary and governance prescriptions, necessitating flexible frameworks adaptable to local circumstances while maintaining core ethical principles including transparency, accountability, privacy protection, and human authority preservation over high-stakes decisions. The research identified common governance elements transcending contextual variation: explicit boundary documentation, stakeholder participation mechanisms, monitoring and audit protocols, adverse event response procedures, and regular framework review and refinement processes. Institutions incorporating these core elements across varied governance models demonstrated consistently superior outcomes compared to those lacking fundamental governance infrastructure regardless of specific structural approach adopted.

Table 8 - Implementation Outcomes by Governance Framework Maturity (N=753)

Maturity Level	Institutions	Documentation	Monitoring	Review Cycle	Adverse Events	Success Rate	Sustainability Score
Advanced Framework	94	98.9%	97.9%	Quarterly	2.7±2.1	87.2%	8.4±1.3
Developing Framework	139	84.2%	73.4%	Biannual	5.3±3.2	74.1%	7.1±1.8
Basic Framework	177	63.8%	48.6%	Annual	8.1±4.1	63.8%	6.2±2.1
Minimal Framework	343	28.3%	21.9%	None/Ad-hoc	12.4±5.7	51.3%	4.8±2.4
Overall	**753**	**56.7%**	**52.3%**	**Variable**	**9.2±5.1**	**63.4%**	**6.1±2.4**

Conclusion

This investigation establishes empirical foundation demonstrating that educational AI agent implementation success depends fundamentally on boundary clarity and governance framework robustness rather than exclusively on technical capability. The finding that 73.2% of institutions lacked formalized operational boundaries while simultaneously reporting 28.3% boundary violation rates reveals critical implementation gap between technological deployment and institutional preparedness. Statistical evidence documenting r=0.821 correlation between boundary clarity and implementation success alongside r=0.794 for governance framework effectiveness validates hypothesis that sustainable AI agent integration requires explicit operational parameter specification and comprehensive ethical oversight architecture. The documented 42.7% efficiency gains achieved through AI agent deployment demonstrate substantial potential for instructional management enhancement, yet simultaneous identification of 34.6% increased accountability complexity and 29.3% violation incidence across functional domains underscores transformation challenges. These quantitative findings align with qualitative evidence revealing stakeholder anxiety regarding autonomous decision-making in high-stakes educational contexts, particularly concerning assessment automation, predictive interventions, and student privacy protection. The research documents divergent implementation trajectories, with institutions implementing hybrid governance models incorporating stakeholder participation achieving 56.8% lower adverse outcome rates compared to centralized administrative approaches, providing empirical support for distributed oversight frameworks balancing institutional authority with stakeholder engagement.

Sector-specific analysis reveals differential boundary requirements across primary, secondary, and tertiary education contexts, with primary settings maintaining substantially more restrictive agent autonomy (M=3.2±1.4) compared to higher education (M=6.8±1.9) reflecting developmental appropriateness concerns and duty-of-care variation. The identification of privacy protection practices as paramount governance dimension, yet documentation of only 37.3% institutions conducting systematic bias audits despite 89.2% acknowledging bias risk, exposes gap between stated priorities and operational practices. This implementation-intention discrepancy manifests across multiple governance dimensions, with sophisticated framework development insufficient absent corresponding organizational capacity building through professional development. The documented 2.8 times higher governance utilization among institutions providing comprehensive training validates organizational learning as critical mediating factor between policy existence and operational effectiveness. Temporal pattern analysis documenting boundary relaxation during expansion phases absent corresponding governance framework updates in 38.2% of institutions reveals systematic vulnerability as operational experience increases stakeholder confidence yet potentially compromises educational quality through inadequate oversight maintenance. Cross-cultural variation in appropriate boundary placement and governance preferences reinforces impossibility of universal prescriptions while simultaneously identifying core elements transcending contextual specificity including documentation, stakeholder participation, monitoring protocols, and regular review processes.

The research establishes tri-level governance architecture as optimal framework encompassing technical boundary specification defining agent operational parameters across functional domains, pedagogical oversight protocols preserving human authority over consequential educational decisions, and distributed ethical accountability structures engaging diverse stakeholders in ongoing implementation oversight. This integrated approach addresses accountability attribution challenges in multi-agent systems while maintaining adaptability to contextual variation in educational norms and

regulatory requirements. The finding that institutions implementing privacy-by-design approaches experienced 67.3% fewer incidents alongside significantly higher stakeholder trust demonstrates tangible benefits of principled governance frameworks prioritizing long-term sustainability over short-term performance optimization. Practical implications include necessity for institutions initiating AI agent implementations to invest comparable resources in governance development as technical deployment, prioritize stakeholder participation throughout implementation lifecycle, establish explicit operational boundaries before system activation, implement continuous monitoring infrastructure, and maintain iterative refinement processes responding to operational experience. Future research directions encompass longitudinal investigation of governance framework evolution beyond eight-month timeframe examined here, comparative effectiveness analysis of specific governance mechanisms across diverse contexts, investigation of optimal stakeholder participation models balancing democratic principles with operational efficiency, and examination of regulatory framework impacts as jurisdictions implement AI-specific educational technology oversight. The ultimate trajectory of educational AI agent integration depends critically on collective capacity to develop governance frameworks maintaining technological innovation while preserving core educational values including pedagogical integrity, learner welfare, educational equity, and human agency in high-stakes decisions affecting student trajectories and educational opportunity access.

References

1. Berkovich I., Hassan T. (2025). The rise of AI-assisted instructional leadership: empirical survey of generative AI integration in school leadership and management work. Frontiers in Education, 10, Article 1643023. DOI: 10.3389/feduc.2025.1643023
2. Chan A., Salganik R., Markelius A., Pang C., Rajkumar N., Krasheninnikov D., Langosco L., He Z., Duan Y., Carroll M. (2024). Visibility into AI Agents. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24) (pp. 891-903). ACM. DOI: 10.1145/3630106.3658948
3. Chan C.K.Y., Hu W. (2023). Students' voices on generative AI: perceptions, benefits, and challenges in higher education. International Journal of Educational Technology in Higher Education, 20, Article 43. DOI: 10.1186/s41239-023-00411-8
4. Collie R., Martin A., Nassar N. (2024). The integration of GenAI in teaching is a revolutionary development that transforms conventional teaching and schooling. Educational Psychology Review, 36(2), 487-512. DOI: 10.1007/s10648-024-09842-w
5. Dieker L., Hines R., Wilkins I., Hughes C., Scott K.H., Smith S., Ingraham K., Ali S., Zaugg T., Shah S. (2024). Using an Artificial Intelligence (AI) Agent to Support Teacher Instruction and Student Learning. Journal of Special Education Preparation, 4(2), 78-88. DOI: 10.33043/d8xb94q7
6. Fullan M., Azorín C., Harris A., Jones M. (2023). Artificial intelligence and school leadership: Challenges, opportunities and implications. School Leadership and Management, 44(4), 339-346. DOI: 10.1080/13632434.2023.2246856
7. Holmes W., Bialik M., Fadel C. (2019). Artificial Intelligence in Education: Promises and Implications for Teaching and Learning. Center for Curriculum Redesign.
8. Jiang Y.-H., Liu T.-Y., Zhuang X., Hu H., Li R., Jia R. (2024). Enhancing educational practices with multi-agent systems: A review. In Y. Wei, C. Qi, Y.-H. Jiang, L. Dai (Eds.), Enhancing Educational Practices: Strategies for Assessing and Improving Learning Outcomes (pp. 47-65). Nova Science Publishers.
9. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). (2024). Official Journal of the European Union, L 2024/1689. https://eur-lex.europa.eu/eli/reg/2024/1689/oj
10. Sposato K., Keding C. (2024). Principals Leading AI in Schools for Instructional Leadership: A Conceptual Model for Principal AI Use. Journal of School Leadership. Published online. DOI: 10.1080/15700763.2024.2428297
11. Tong R.J., Hu X. (2024). Future of Education with Neuro-Symbolic AI Agents in Self-Improving Adaptive Instructional Systems. Frontiers of Digital Education, 1, 198-212. DOI: 10.1007/s44366-024-0008-9
12. Tsikrika T., Petkos G., Vrochidis S. (2024). Trustworthy AI in education: A Roadmap for Ethical and Effective Implementation. In Proceedings of the 13th Hellenic Conference on Artificial Intelligence (pp. 167-173). ACM. DOI: 10.1145/3688671.3688781
13. UNESCO. (2024). Guidance for generative AI in education and research.

UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000389715
14. Wang S., Wang F., Zhu Z., Wang J., Tran T., Du Z. (2024). Artificial intelligence in education: A systematic literature review. Expert Systems with Applications, 252(Part A), Article 124167. DOI: 10.1016/j.eswa.2024.124167
15. Wei Y., Qi C., Jiang Y.-H. (2024). Agent4EDU: Advancing AI for Education with Agentic Workflows. In Proceedings of the 2024 3rd International Conference on Artificial Intelligence and Education (ICAIE '24) (pp. 156-162). ACM. DOI: 10.1145/3722237.3722268
16. Yu J., Hao Z., Li R.M. (2025). AI instructional agent improves student's perceived learner control and learning outcome: empirical evidence from a randomized controlled trial. arXiv preprint, arXiv:2505.22526v1. https://arxiv.org/abs/2505.22526
17. Zhang L., Chen Y., Li M. (2024). Ethical framework for AI education based on large language models. Education and Information Technologies. Published online December 23, 2024. DOI: 10.1007/s10639-024-13241-6

Границы применения и система этического управления образовательными агентами ИИ в управлении учебным процессом

Ван Цзюньган

Постдокторант,
Московский государственный университет им. М. В. Ломоносова,
119991, Российская Федерация, Москва, Ленинские горы, 1;
e-mail: 3943329717@qq.com

Аннотация

Образовательные агенты искусственного интеллекта представляют собой фундаментальную трансформацию систем управления учебным процессом, однако их внедрение осуществляется в условиях недостаточно определённых границ и систем управления. В данном исследовании рассматриваются ограничения применения и механизмы этического надзора в 847 образовательных учреждениях, внедряющих системы агентов ИИ, в период с января по декабрь 2024 года. С использованием смешанных методов анализа, включающих институциональные опросы, показатели производительности систем и интервью с заинтересованными сторонами, исследование выявило критические пороги, где автономия агентов пересекается с педагогической ответственностью. Результаты показали, что в 73,2% учреждений отсутствовали формализованные протоколы для определения границ принятия решений агентами ИИ, в то время как 68,4% сообщили о пробелах в управлении в сфере надзора за конфиденциальностью данных. Статистический анализ выявил коэффициенты корреляции r=0,821 между чётко определёнными операционными границами и успешными результатами внедрения, а также r=0,794 для учреждений с установленными этическими рамками. Многоагентные системы в классах достигли 42,7% повышения эффективности выполнения административных задач, но привели к 34,6% увеличению сложности механизмов подотчётности. В исследовании зафиксированы различные модели внедрения в начальном, среднем и высшем образовании, причём нарушения границ occurred в 28,3% сценариев автономного оценивания и в 41,9% систем предиктивного вмешательства. Этические системы управления с участием заинтересованных сторон снизили неблагоприятные результаты на 56,8% по сравнению с централизованными административными подходами. Полученные данные указывают на то, что устойчивая интеграция агентов ИИ требует трёхуровневой архитектуры управления, включающей

техническую спецификацию границ, педагогические протоколы надзора и распределённые структуры этической подотчётности. Исследование устанавливает эмпирически обоснованные параметры для delineation операционной сферы агентов при сохранении приоритета человеческого надзора в образовательных решениях с высокими ставками. Эти результаты создают основу для разработки стандартизированных рамок, балансирующих технологические возможности с педагогической целостностью и защитой благополучия учащихся.

Для цитирования в научных исследованиях

Ван Цзюньган. Application Boundaries and Ethical Governance Framework of Educational AI Agents in Instructional Management // Экономика: вчера, сегодня, завтра. 2026. Том 16. № 3А. С. 790-805. DOI: 10.34670/AR.2026.74.29.041

Ключевые слова

Образовательные агенты ИИ, границы управления учебным процессом, система этического управления, надзор за автономными системами, педагогическая подотчётность, многоагентные образовательные системы, этика ИИ в образовании.

Библиография

1. Berkovich I., Hassan T. The rise of AI-assisted instructional leadership: empirical survey of generative AI integration in school leadership and management work // Frontiers in Education. 2025. Vol. 10. Article 1643023. DOI: 10.3389/feduc.2025.1643023.
2. Chan A., Salganik R., Markelius A., Pang C., Rajkumar N., Krasheninnikov D., Langosco L., He Z., Duan Y., Carroll M. Visibility into AI Agents // Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24). ACM, 2024. P. 891-903. DOI: 10.1145/3630106.3658948.
3. Chan C.K.Y., Hu W. Students' voices on generative AI: perceptions, benefits, and challenges in higher education // International Journal of Educational Technology in Higher Education. 2023. Vol. 20. Article 43. DOI: 10.1186/s41239-023-00411-8.
4. Collie R., Martin A., Nassar N. The integration of GenAI in teaching is a revolutionary development that transforms conventional teaching and schooling // Educational Psychology Review. 2024. Vol. 36. No. 2. P. 487-512. DOI: 10.1007/s10648-024-09842-w.
5. Dieker L., Hines R., Wilkins I., Hughes C., Scott K.H., Smith S., Ingraham K., Ali S., Zaugg T., Shah S. Using an Artificial Intelligence (AI) Agent to Support Teacher Instruction and Student Learning // Journal of Special Education Preparation. 2024. Vol. 4. No. 2. P. 78-88. DOI: 10.33043/d8xb94q7.
6. Fullan M., Azorín C., Harris A., Jones M. Artificial intelligence and school leadership: Challenges, opportunities and implications // School Leadership and Management. 2023. Vol. 44. No. 4. P. 339-346. DOI: 10.1080/13632434.2023.2246856.
7. Holmes W., Bialik M., Fadel C. Artificial Intelligence in Education: Promises and Implications for Teaching and Learning. Boston: Center for Curriculum Redesign, 2019. 248 p. ISBN: 978-1-7348451-0-9.
8. Jiang Y.-H., Liu T.-Y., Zhuang X., Hu H., Li R., Jia R. Enhancing educational practices with multi-agent systems: A review // Enhancing Educational Practices: Strategies for Assessing and Improving Learning Outcomes / eds. Wei Y., Qi C., Jiang Y.-H., Dai L. Nova Science Publishers, 2024. P. 47-65.
9. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) // Official Journal of the European Union. 2024. L 2024/1689. URL: https://eur-lex.europa.eu/eli/reg/2024/1689/oj.
10. Sposato K., Keding C. Principals Leading AI in Schools for Instructional Leadership: A Conceptual Model for Principal AI Use // Journal of School Leadership. 2024. Published online. DOI: 10.1080/15700763.2024.2428297.
11. Tong R.J., Hu X. Future of Education with Neuro-Symbolic AI Agents in Self-Improving Adaptive Instructional Systems // Frontiers of Digital Education. 2024. Vol. 1. P. 198-212. DOI: 10.1007/s44366-024-0008-9.
12. Tsikrika T., Petkos G., Vrochidis S. Trustworthy AI in education: A Roadmap for Ethical and Effective Implementation // Proceedings of the 13th Hellenic Conference on Artificial Intelligence. ACM, 2024. P. 167-173. DOI: 10.1145/3688671.3688781.
13. UNESCO. Guidance for generative AI in education and research. Paris: UNESCO, 2024. 48 p.

URL: https://unesdoc.unesco.org/ark:/48223/pf0000389715.
14. Wang S., Wang F., Zhu Z., Wang J., Tran T., Du Z. Artificial intelligence in education: A systematic literature review // Expert Systems with Applications. 2024. Vol. 252, Part A. Article 124167. DOI: 10.1016/j.eswa.2024.124167.
15. Wei Y., Qi C., Jiang Y.-H. Agent4EDU: Advancing AI for Education with Agentic Workflows // Proceedings of the 2024 3rd International Conference on Artificial Intelligence and Education (ICAIE '24). ACM, 2024. P. 156-162. DOI: 10.1145/3722237.3722268.
16. Yu J., Hao Z., Li R.M. AI instructional agent improves student's perceived learner control and learning outcome: empirical evidence from a randomized controlled trial // arXiv preprint. 2025. arXiv:2505.22526v1. URL: https://arxiv.org/abs/2505.22526.
17. Zhang L., Chen Y., Li M. Ethical framework for AI education based on large language models // Education and Information Technologies. 2024. Published online 23 December 2024. DOI: 10.1007/s10639-024-13241-6.