VOLUME 01 / 2026

Designership Press

Who Holds the Classroom? Bounded Pedagogical Autonomy for Social Robots and AI Agents in K-12 Education

Dr. Freedom Cheteni

Critical integrative analysis and prospective field-study protocol

Research edition 1.0 | Prepared 2026-10-16

D02 / Human pedagogical authority

Who Holds the Classroom? Bounded Pedagogical Autonomy for Social Robots and AI Agents in K-12 Education

Dr. Freedom Cheteni

The Research University | Volume 1, 2026 | Edition 1.0 | D02

Critical integrative analysis and prospective field-study protocol

Original analyses and prospective protocols. New independent human peer review is not recorded for this edition.

DOI: not assigned to this research edition.

Read online | Author ORCID: 0009-0006-3507-0094 | Editorial record

Abstract

Social robots and generative AI agents create a consequential distinction between executing instructional routines and holding pedagogical authority. This original critical integrative analysis proposes bounded pedagogical autonomy: action-specific permissions that preserve human responsibility for curriculum, interpretation, safeguarding, and consequential decisions. Primary classroom and laboratory studies are examined for the conditions under which personalisation, embodiment, rapport, and social influence may support or undermine learning; findings from language-learning and generative-AI trials are interpreted within their actual populations and assessment conditions. The article introduces an autonomy-oversight matrix and a six-part failure taxonomy, then specifies a prospective four-arm school-cluster randomised trial crossing embodied versus screen presentation with adaptive versus fixed scaffolding. The primary outcome is delayed unaided transfer rather than assisted performance or engagement. The protocol matches instructional content and adult contact, measures workload and cost, specifies clustered analysis and missing-data sensitivity, and makes consent, child assent, accessibility, escalation, and data minimisation operational requirements. Recruitment, ethics approval, preregistration, power simulation, and trial execution have not occurred. The contribution is a falsifiable design and governance argument, not a report of efficacy. Designership is advanced as accountable inquiry: deployment must remain contingent on independent learning, sustainable adult oversight, and protection of children’s rights.

Keywords: Designership, social robots, K-12 education, pedagogical authority, bounded autonomy, unaided transfer, child-centred AI, cluster randomised trial

Why classroom autonomy needs a theory of authority

A social robot can ask a question, wait for an answer, and invite a child to revise an explanation. None of those actions establishes that the robot should decide what the child is entitled to learn, whether an answer deserves a grade, or when a family should be contacted. The important educational distinction is between executing an instructional routine and holding pedagogical authority. This article develops bounded pedagogical autonomy: a design in which an agent may select a limited next action, while human educators retain responsibility for curriculum, interpretation, safeguarding, and consequential decisions.

The intellectual starting point is the Designership programme’s interest in social robotics, agent-supported classrooms, and Movement Thinking. These are treated as a research agenda rather than evidence of effectiveness. The article makes three original contributions: an autonomy-oversight matrix that connects permissions to consequences; a failure taxonomy connecting technical incidents to learning and dignity; and a comparative field-study protocol capable of separating embodied presence from adaptive pedagogy. It is a critical integrative analysis and prospective protocol, not a completed trial, systematic review, or validated instrument.

The central proposition is deliberately conditional. An agent is educationally useful when it helps a learner produce a better explanation and subsequently act without the agent, while leaving educators able to understand and contest its interventions. A fluent exchange, longer screen time, or apparent affection for a robot is insufficient. For Designership, the consequential question is therefore not whether a classroom can run itself. It is which bounded forms of automation make inquiry more capable, equitable, and accountable.

Evidence is conditional, not a verdict on robots

Classroom robotics research provides reasons for investigation and reasons for restraint. Baxter and colleagues studied personalised and non-personalised robot peers in two primary-school classrooms over two weeks. They reported learning advantages associated with personalisation, including on a novel topic. The small number of classrooms and the specific robot-touchscreen task limit generalisation to whole-school deployment. Importantly, operation without experimenter supervision did not mean absence of the classroom teacher. The study supports testing particular personalised routines, not substituting machines for adult educational responsibility. [R1]

Tanaka and Matsuzoe investigated children teaching vocabulary to a care-receiving robot. Their field experiments make an important conceptual move: a robot can be an imperfect recipient of an explanation rather than an authoritative examiner. This motivates a classroom role in which students must explain and correct. It does not

establish that every simulated mistake improves learning; teachers must distinguish a purposeful misconception from an unmarked factual error. [R2]

Findings also challenge an assumption that more social behaviour necessarily produces better instruction. Kennedy and colleagues experimentally examined how a tutor robot's social behaviour could interfere with child learning. In a larger language-learning programme, van den Berghe and colleagues reported heterogeneous relationships between children's language-learning characteristics and robot support; the underlying comparison found no overall vocabulary advantage from adding the robot to the tablet lessons. Embodiment, gestures, instructional content, and learner characteristics should therefore be separated analytically rather than bundled as a uniformly effective technology. [R3, R4]

Social effects can be educationally ambiguous. Kory-Westlund and Breazeal's storytelling experiment examined rapport-related behaviours and found changes in children's interaction, retelling, and responses to the robot. Vollmer and colleagues showed that children could conform to incorrect robot-group answers in an experimental task. Neither study establishes inevitable classroom harm. Together they justify measuring whether a learner can challenge a persuasive agent, rather than treating compliance or relationship ratings as an unqualified benefit. [R5, R6]

What generative AI changes-and what it does not

Language models broaden the range of responses an agent can produce, but they also make its instructional behaviour less exhaustively scriptable. A robot that selects among approved prompts and a robot that improvises explanations are different interventions. The second requires monitoring the semantic content of outputs, not merely testing whether motors and microphones work. Retrieval from approved material can help constrain responses, but it does not guarantee correct inference, faithful quotation, or an appropriate response to a child.

A high-school mathematics trial by Bastani and colleagues is especially relevant because it distinguished assisted task performance from subsequent unaided performance and compared different forms of AI support. Its results warn against equating access to an effective answer generator with durable learning. The publication has a linked correction, which is retained in this article's evidence record; it corrects Osbert Bastani's institutional affiliation following a production error, rather than the study findings, and is not treated as an independent efficacy study. No effect size is transferred from that intervention to classroom robots. [R7, R8]

Kestin and colleagues reported beneficial learning outcomes from a carefully designed AI tutor compared with in-class active learning in college physics. That study concerns university students and a particular instructional design. Its appropriate role here is to show that implementation and pedagogy matter, not to predict benefits for children. Together, these studies support a test of competing mechanisms: scaffolding may

promote independent reasoning, whereas immediate answer access may substitute for the reasoning that assessment intends to reveal. [R9]

The proposed system therefore offers prompts for explanation, comparison, and revision rather than unrestricted answer delivery. The lesson's explanation of the tool is literal: it is software that can be useful and wrong, and adults remain responsible for teaching. The interface must avoid claims that the agent understands a child's feelings, has needs the child must satisfy, or is a confidential substitute for a trusted adult.

The autonomy-oversight matrix

The proposed matrix classifies an action along two axes: its educational consequence and the reversibility of its effects. Consequence is not identical to technical complexity. Reading an approved sentence aloud is technically sophisticated but relatively low consequence; silently assigning a child to a remedial pathway can be technically simple and highly consequential. Reversibility concerns whether a mistake can be corrected without persistent educational, social, or informational damage. A permission is granted to a specified action in a specified context, never to an agent in general.

Level A permits bounded execution: announcing the next approved task, reading a source passage, or offering a preapproved prompt. The teacher approves the lesson and prompt set beforehand; the system logs the action and offers an immediate stop control. Level B permits constrained adaptation: selecting among approved hints using the learner's current task response. The teacher can inspect the basis for the selection, override it, and review a concise exception log. The agent cannot infer a general ability label from that interaction.

Level C covers proposals requiring explicit adult approval before their effect: changing a learning goal, recommending a placement, sending an external message, or adding information to a persistent student record. The agent may draft a proposal with its evidence and uncertainty; it cannot execute it. Level D excludes delegated authority in this study: discipline, admission, disability identification, mental-health diagnosis, high-stakes grading, and independent safeguarding decisions. A child's disclosure of possible harm triggers an adult safeguarding pathway; it does not trigger an automated clinical judgement.

These levels become an implementation table with five mandatory fields for each action: permitted input, permitted output, responsible adult, escalation trigger, and maximum retention period. For example, an uncertain oral response permits a repeat-or-type request, not an ability score. A request for a grade permits referral to the teacher, not a generated mark. A request to share personal information permits a privacy reminder and adult referral, not a new data collection form. The table is an original governance proposal that the trial will evaluate, not an established standard.

A failure taxonomy that preserves educational meaning

A failure register should distinguish six classes. Epistemic failures include incorrect explanations, invented sources, and misleading certainty. Pedagogical failures include revealing an answer too early, rewarding an unsupported claim, or giving a hint that makes the assessed task trivial. Interaction failures include speech-recognition breakdown, inaccessible instructions, and turn-taking that prevents a learner from participating. Authority failures occur when the agent acts outside its permission level or obscures who made a consequential decision. Relational failures include coercive praise, pressure to agree, or language encouraging secrecy. Data failures include excessive collection, wrong-student attribution, and unauthorised disclosure.

The unit of coding is a specific event with an observable context and consequence, not a general impression that a session went poorly. Independent trained coders will record the task, model and prompt version, output, permission level, severity, student-visible correction, time to adult response, and whether the learning activity resumed. A wrong statement caught before delivery differs from one repeated by a learner on an unaided transfer task. Both matter, but they should not be counted as equivalent instructional exposure.

Severity rules distinguish inconvenience, recoverable instructional distortion, and serious safety or rights events. All serious events trigger immediate suspension of the affected function pending human review. A single credible unauthorised disclosure or safeguarding mishandling is sufficient for that suspension; there is no requirement to wait for statistical significance. For ordinary errors, a preregistered review threshold will combine exposure frequency and severity. Thresholds will be set with educators, family representatives, and the ethics committee before recruitment, then frozen for the trial.

Reporting includes the denominator of agent opportunities and learner exposure time. A system used less often may show fewer incidents while being less reliable per interaction. Conversely, an increase in logged incidents after better monitoring may reflect better detection. The trial will therefore report incident rates, total exposure, detection route, correction quality, and unresolved cases separately.

Field-study design: separating embodiment from pedagogy

The proposed primary study is a four-arm, parallel, school-cluster randomised trial with a two-by-two factorial design. One factor is interface: a stationary social robot or a screen-based agent with matched audio and visual content. The other is instructional policy: bounded adaptive scaffolding or a fixed, curriculum-aligned prompt sequence. All arms receive the same learning goals, source material, task time, teacher training, escalation procedure, and human supervision. This isolates the incremental contribution of embodied presentation and of adaptive prompt selection more cleanly than comparing a robot lesson with unrelated business-as-usual instruction.

Schools, rather than individual children within the same room, will be allocated to reduce intervention sharing and inconsistent classroom norms. Schools will be recruited before allocation; classes and consented students will be enumerated before randomisation where feasible. An independent statistician will use concealed, blocked allocation stratified by school phase, baseline attainment band, and broad deprivation category. The analysis will respect the unit of assignment and report the number of schools, classes, teachers, and students in every arm. Cluster-trial reporting principles inform this design. [R10]

The initial target population is students aged approximately ten to thirteen undertaking an eight-week science explanation and design unit. The narrow age range and subject context are deliberate boundaries, not a claim to cover all K-12 education. Tasks require learners to explain a phenomenon, compare evidence, revise a proposed design, and justify a trade-off. Teachers may stop a session at any time. Students can choose an equivalent non-agent activity without academic penalty, with participation in research data collection handled separately from access to instruction.

A two-week familiarisation period precedes the eight-week instructional unit. All groups receive comparable orientation to their assigned interface and equivalent researcher contact. The robot remains stationary and does not touch children or navigate around a classroom. A neutral introduction and matched interaction time reduce novelty differences, but cannot eliminate them. Weekly novelty ratings and exposure histories will therefore be reported as process measures rather than used to redefine the primary outcome after results are known.

Outcomes that require the learner to act independently

The primary endpoint is performance on a delayed, unaided transfer assessment four weeks after the unit. Students encounter unfamiliar surface features but the same underlying scientific and design principles. They produce an explanation and revise a design in response to new evidence without access to any agent. A prespecified rubric scores causal adequacy, use of evidence, recognition of a trade-off, and justified revision. Two raters masked to treatment will independently score a prespecified sample sufficient to estimate agreement; discrepancies follow a documented adjudication rule.

Before recruitment, assessment developers will pilot task difficulty, accessibility, rubric interpretation, and parallel-form comparability with students outside trial schools. The primary score is a fixed weighted composite established during that pilot. It will not be reweighted to favour whichever subscale yields the largest intervention difference. Immediate knowledge, retention, and revision quality are secondary outcomes. Engagement is a process measure. The capacity to detect and challenge a plausible but incorrect explanation is a separately declared secondary outcome using a brief, debriefed assessment rather than covert manipulation during instruction.

Teacher workload is a coequal implementation criterion. Weekly short diaries will record preparation, supervision, technical recovery, reviewing exceptions, and additional time outside paid hours. A subsample will complete structured time observations to estimate diary bias. Costs will include devices, software, accessibility supports, training, maintenance, and adult attention. The study will report educational effects alongside workload and cost rather than converting all benefits into a single promotional return-on-investment figure.

Accessibility will be designed into the tasks, including text alternatives, captions, adjustable response time, switch-compatible input where required, and a route that does not require speech. Language background and disability-related accommodation are not error labels. Recruitment and retention by relevant groups will be reported with privacy-preserving cell rules. Interpretation of subgroup differences will attend to whether the interface offered equal access before attributing differences to learners.

Hypotheses, estimands, and a power plan

The primary confirmatory hypothesis is that bounded adaptive scaffolding improves delayed unaided transfer relative to the fixed prompt sequence, averaged across interface conditions. The primary estimand is the student-average intention-to-treat difference for students eligible and enrolled under the preregistered recruitment rule, accounting for school assignment. Marginal contrasts will weight enrolled students equally; school-size weights will be fixed using preallocation eligible enrolment, and the analysis plan will specify their implementation and handling of informative cluster size before allocation. An equal-school-weight contrast will be a sensitivity analysis because it targets a different population quantity. A second confirmatory contrast tests the main effect of embodiment on that same outcome. The family of these two tests will use a prespecified multiplicity procedure. The interface-by-policy interaction is secondary unless power simulations support an adequately sized confirmatory interaction test before recruitment begins.

An original mechanism hypothesis predicts that adaptive scaffolding increases substantive explanation revisions rather than merely response length. A rival account predicts improved assisted task completion without improved unaided transfer. A separate boundary hypothesis predicts that any learning benefit is unacceptable for deployment if accompanied by a meaningful increase in uncompensated teacher workload or a serious rights event. The workload margin must be negotiated and documented with teachers before allocation; this manuscript does not invent a universally acceptable number of extra minutes.

Power will be established through simulation using the planned multilevel analysis, plausible school and class intraclass correlations, unequal cluster sizes, baseline-outcome correlation, and expected attrition. A provisional planning scenario is sixty schools, fifteen per factorial cell, with approximately forty participating students per school before attrition. This is a feasibility scenario, not a recruited sample or a power finding. Simulations will examine small educationally relevant effects across conservative correlation and attrition scenarios; the final sample target, code, assumptions, and resulting detectable effects will be registered before randomisation.

If the feasible number of schools cannot support the declared estimands, investigators will increase recruitment, reduce the number of confirmatory contrasts, or designate the study as a feasibility trial before allocation. They will not justify an underpowered efficacy claim by counting thousands of student interactions as independent observations. School withdrawal will be simulated separately from student nonresponse because losing clusters damages information differently from losing the same number of students across retained schools.

Analysis, missingness, and evidence of mechanism

The primary analysis will use a hierarchical model with treatment indicators, their prespecified interaction, baseline transfer-related attainment, randomisation strata, and school and class structure appropriate to the sampling plan. Small-sample degrees-of-freedom corrections or randomisation-based inference will be prespecified for the number of clusters. Estimates will include confidence intervals and both raw-score and standardised interpretations. A decision to use random rather than fixed class effects will be made before unmasking, based on the data structure and simulation performance, not statistical significance.

Missing outcomes will be tabulated by school, arm, time point, and stated reason. The primary missing-data approach will use a multilevel-compatible procedure under an explicit missing-at-random assumption, including baseline attainment and predictors of nonresponse. Sensitivity analyses will vary departures from that assumption, including worse unobserved outcomes among disengaged students and informative school withdrawal. A complete-case estimate will be reported as a sensitivity analysis, not treated as automatically unbiased. Attendance, tool exposure, crossover, and teacher overrides will be reported without conditioning the primary treatment effect on post-randomisation compliance.

Potential mediation by revision quality will be exploratory. Random assignment to a tool does not randomise a student’s willingness to revise; a simple association between revision and transfer cannot establish that revision caused the treatment effect. The study will analyse temporal patterns and contrasting cases to refine the proposed mechanism, explicitly acknowledging confounding and measurement error. Interview accounts from teachers and students will be used to understand failures and implementation conditions, not to replace a null primary outcome with a success narrative.

All prespecified outcomes, model versions, subgroup analyses, and deviations will appear in a public results register. Subgroups will be limited to a small set justified before allocation, with interaction estimates and uncertainty. Any additional discovery analysis will be labelled exploratory and will require later replication. A null embodiment effect remains informative: if a less expensive screen produces comparable transfer with fewer burdens, the appropriate educational conclusion may favour that screen.

Failure to detect an interface effect will not establish equivalence. An equivalence or non-inferiority claim would require a defensible educational margin, appropriate power, and a prespecified testing procedure before recruitment. The present protocol is a superiority design. A confidence interval that excludes only large differences may narrow a deployment decision, but it cannot certify that two tools are educationally interchangeable. Any later equivalence study will be registered as a distinct analysis or trial rather than retrofitted to a null result.

Consent, safeguarding, and data minimisation

The study requires approval from an appropriately constituted research ethics body, participating schools, and relevant data controllers before recruitment. Guardian permission and age-appropriate child assent will be obtained for research participation; children can pause or withdraw without losing access to equivalent learning. Consent materials will explain what the agent does, who can inspect records, what data leave the device, and the limits of confidentiality when an adult must respond to a safeguarding concern. No approval, consent, or recruitment is claimed in this article.

Default data collection excludes facial emotion inference, continuous classroom video, biometric identification, and ambient recording of nonparticipants. Speech is optional; push-to-talk input will be transient where technically feasible, with text retained only when necessary for the declared outcomes and incident audit. Research identifiers are separate from school identifiers, access is role-limited, and retention and deletion schedules are approved before data collection. Student data will not be used to train general-purpose models or support advertising. These are proposed operational safeguards, to be tested through technical and institutional review rather than asserted as automatic compliance. [R12, R14]

The interface offers a visible stop-and-ask-an-adult control. Disclosure of possible harm causes the agent to cease the ordinary tutoring routine, provide a short neutral acknowledgement, and alert the designated adult through the school’s approved route. It must not promise secrecy, investigate an allegation, or make a diagnosis. Escalation includes a documented fallback when the designated adult is unavailable. A practical drill with adults and simulated inputs must pass before any child-facing session.

Relational restraint is part of safeguarding. The agent will not ask children to keep it company, express distress when they leave, solicit personal secrets, or imply that declining interaction is unkind. A student may question its accuracy without losing rewards. These restrictions may reduce some measures of attachment or compliance; that is compatible with the educational objective. UNICEF’s child-rights guidance supports attention to safety, privacy, fairness, participation, and accountable oversight, but does not itself validate this protocol or endorse the system. [R12]

Reproducibility and a publication record fit for scrutiny

Before the trial, investigators will preregister the protocol, outcome definitions, analysis plan, failure codebook, stopping rules, and intervention manifest in an appropriate public registry. The manifest will identify robot hardware, operating software, language model provider and version, system instructions, retrieval corpus, allowed actions, and input/output filters. Material system changes during the trial require a dated amendment and a plan for interpreting affected observations. Frozen instructional behaviour is preferable to silently adopting a provider update that changes the intervention halfway through the study.

The SPIRIT-AI and CONSORT-AI guidance was developed for clinical research, not schools; its requirements concerning system versions, inputs, outputs, human interaction, and error analysis are adapted here as reporting aids. Their use does not turn an educational trial into a clinical trial or establish institutional approval. Together with the cluster-trial extension, they motivate a record that allows readers to identify which intervention was actually evaluated and under what human operating conditions. [R10, R11, R13]

Publication will include assessment rubrics, analysis code, simulated or synthetic test data, aggregated incident tables, and a description of restricted data access where consent and disclosure risk preclude open student-level records. Student work will be published only under separate, informed permission and an appropriate licence. Raw child conversations are not a necessary public supplement. Open research should make claims inspectable without making children permanently searchable.

The paper agent attached to this article should answer from this manuscript and its verified sources, distinguish proposed methods from completed findings, and expose the passage supporting a response. It should return uncertainty when the manuscript is silent. Successful retrieval is not evidence that the field trial has occurred, and software evaluation must be recorded separately from scholarly peer review. Independent human reviewers should examine the theoretical argument, child protections, outcome validity, and analysis plan before an efficacy study begins.

Limitations and a decision rule for Designership

The framework is an original proposal awaiting validation. Its matrix may be too coarse for some classrooms, and coding failures reliably may require more training than expected. School recruitment may favour well-resourced settings, limiting transportability. Students and teachers cannot be blinded to a physical robot, and their expectations may influence outcomes despite matched contact and assessment masking. A short instructional unit cannot establish long-term effects on attachment, teacher employment, civic development, or whole-school culture.

The factorial design estimates embodiment within a tightly specified lesson system; it does not compare every robot with every chatbot. The robot and screen cannot be made sensorially identical, and costs may differ substantially. Teacher adaptations may be both necessary for inclusion and a source of implementation variation. Results must therefore report what educators actually did, including overrides and departures, rather than attributing all change to an abstract category called AI.

A defensible deployment decision requires a joint pattern: credible evidence for the stated claim about unaided learning, no unacceptable rights or safety events, workable adult oversight, accessible participation, and sustainable workload and cost. Positive engagement with weak transfer fails that test. Improved transfer produced by hidden teacher labour also fails the claim that automation has increased classroom capacity. The study may support a narrower permission set, a screen rather than a robot, or no adoption. Designing the inquiry so that each conclusion remains possible is the article’s central contribution to Designership.

References and verification record

References retain their original authorship and publication identity. Verification dates below record source and metadata checking; they do not confer review or approval on this article.

[R1] Paul Baxter; Emily Ashurst; Robin Read; James Kennedy; Tony Belpaeme (2017). Robot education peers in a situated primary school study: Personalisation promotes child learning. PLOS ONE.

DOI: https://doi.org/10.1371/journal.pone.0178126

Source: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0178126

Verified 2026-10-16. Primary classroom field experiment. Publisher full text and Crossref DOI metadata verified. Two classrooms; personalisation evidence is not whole-school autonomy evidence.

[R2] Fumihide Tanaka; Shizuko Matsuzoe (2012). Children Teach a Care-Receiving Robot to Promote Their Learning: Field Experiments in a Classroom for Vocabulary Learning. Journal of Human-Robot Interaction.

DOI: https://doi.org/10.5898/jhri.1.1.tanaka

Source: https://doi.org/10.5898/JHRI.1.1.Tanaka

Verified 2026-10-16. Primary field experiments. Publisher DOI landing, University of Tsukuba laboratory bibliography and Crossref metadata verified; no unrestricted classroom-autonomy claim.

[R3] James Kennedy; Paul Baxter; Tony Belpaeme (2015). The Robot Who Tried Too Hard: Social Behaviour of a Robot Tutor Can Negatively Affect Child Learning. Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction.

DOI: https://doi.org/10.1145/2696454.2696457

Source: https://doi.org/10.1145/2696454.2696457

Verified 2026-10-16. Primary conference experiment, not journal article. Crossref title/subtitle and author order verified; official HRI programme and author institutional bibliography confirm study identity. DOI publisher full text blocked to web retrieval; cited conclusion corroborated by primary author record.

[R4] Rianne van den Berghe; Ora Oudgenoeg-Paz; Josje Verhagen; Susanne Brouwer; Mirjam de Haas; Jan de Wit; Bram Willemsen; Paul Vogt; Emiel Krahmer; Paul Leseman (2021). Individual Differences in Children’s (Language) Learning Skills Moderate Effects of Robot-Assisted Second

Language Learning. Frontiers in Robotics and AI.

DOI: https://doi.org/10.3389/frobt.2021.676248

Source: https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2021.676248/full

Verified 2026-10-16. Primary study and follow-up analysis from the same L2TOR dataset, not an independent replication of Vogt et al. Publisher full text verifies authors, null overall added robot benefit and mixed moderator results.

[R5] Jacqueline M. Kory-Westlund; Cynthia Breazeal (2019). Exploring the Effects of a Social Robot's Speech Entrainment and Backstory on Young Children's Emotion, Rapport, Relationship, and Learning. Frontiers in Robotics and AI.

DOI: https://doi.org/10.3389/frobt.2019.00054

Source: https://www.frontiersin.org/journals/robotics-and-ai/articles/10.3389/frobt.2019.00054/full

Verified 2026-10-16. Primary experiment. Publisher full text verifies two authors, 86 children and 2x2 design; no evidence of established developmental benefit inferred from rapport.

[R6] Anna-Lisa Vollmer; Robin Read; Dries Trippas; Tony Belpaeme (2018). Children conform, adults resist: A robot group induced peer pressure on normative social conformity. Science Robotics.

DOI: https://doi.org/10.1126/scirobotics.aat7111

Source: https://www.science.org/doi/10.1126/scirobotics.aat7111

Verified 2026-10-16. Primary experimental study. Science Robotics DOI and Crossref author order/title/year verified; publisher search and author institutional record corroborate child conformity finding. Publication day not asserted because indexes differ.

[R7] Hamsa Bastani; Osbert Bastani; Alp Sungu; Haosen Ge; Özge Kabakcı; Rei Mariman (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences.

DOI: https://doi.org/10.1073/pnas.2422633122

Source: https://www.pnas.org/doi/10.1073/pnas.2422633122

Verified 2026-10-16. Primary high-school randomised trial, journal version of record. Exact DOI verified against publisher and Crossref; source-version preprint 10.2139/ssrn.4895486 is not treated as a second study. Linked correction retained as R8.

[R8] (2025). Correction for Bastani et al., Generative AI without guardrails can harm learning: Evidence from high school mathematics. Proceedings of the National Academy of Sciences.

DOI: https://doi.org/10.1073/pnas.2518204122

Source: https://www.pnas.org/doi/10.1073/pnas.2518204122

Verified 2026-10-16. Publisher correction notice, not a study. Crossref update-to explicitly identifies correction to R7; primary full text independently checked in editorial audit: it corrects Osbert Bastani’s institutional affiliation after a production error, not the findings. Crossref notice carries no individual authors. Complete ongoing retraction surveillance is not claimed.

[R9] Greg Kestin; Kelly Miller; Anna Klales; Timothy Milbourne; Gregorio Ponti (2025). AI tutoring outperforms in-class active learning: an RCT introducing a novel research-based design in an authentic educational setting. Scientific Reports.

DOI: https://doi.org/10.1038/s41598-025-97652-6

Source: https://www.nature.com/articles/s41598-025-97652-6

Verified 2026-10-16. Primary randomised trial in college physics, not K12 or embodied robotics. Publisher full text and DOI metadata verified. Generalisation to children explicitly limited.

[R10] Marion K. Campbell; Gilda Piaggio; Diana R. Elbourne; Douglas G. Altman; CONSORT Group (2012). Consort 2010 statement: extension to cluster randomised trials. BMJ.

DOI: https://doi.org/10.1136/bmj.e5661

Source: https://www.bmj.com/content/345/bmj.e5661

Verified 2026-10-16. Primary reporting guideline, cluster-trial extension. Publisher/author metadata and CONSORT Group attribution verified. Used as reporting and design guidance, not educational efficacy evidence.

[R11] Xiaoxuan Liu; Samantha Cruz Rivera; David Moher; Melanie J. Calvert; Alastair K. Denniston; The SPIRIT-AI and CONSORT-AI Working Group (2020). Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. Nature Medicine.

DOI: https://doi.org/10.1038/s41591-020-1034-x

Source: https://www.nature.com/articles/s41591-020-1034-x

Verified 2026-10-16. Primary consensus reporting guideline. Publisher confirms five lead authors plus named working group. Clinical origin explicitly stated; adaptation to educational research is proposed, not validated.

[R12] UNICEF Innocenti - Global Office of Research and Foresight (2025). UNICEF Guidance on AI and Children 3.0. UNICEF Innocenti.

Source: https://www.unicef.org/innocenti/reports/policy-guidance-ai-children

Verified 2026-10-16. Authoritative primary policy guidance, version 3.0 December 2025. Official page and linked report verify child-rights requirements. No DOI assigned in consulted source; not an efficacy study or endorsement.

[R13] Samantha Cruz Rivera; Xiaoxuan Liu; An-Wen Chan; Alastair K. Denniston; Melanie J. Calvert; The SPIRIT-AI and CONSORT-AI Working Group (2020). Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. Nature Medicine.

DOI: https://doi.org/10.1038/s41591-020-1037-7

Source: https://www.nature.com/articles/s41591-020-1037-7

Verified 2026-10-16. Primary consensus protocol guideline. Publisher and PubMed verify five lead authors and working group. Crossref request rate-limited; publisher DOI/title/authors/year independently accessible. Clinical origin explicitly stated.

[R14] Fengchun Miao; Wayne Holmes (2023). Guidance for generative AI in education and research. UNESCO.

DOI: https://doi.org/10.54675/EWZM9535

Source: https://unesdoc.unesco.org/ark:/48223/pf0000386693

Verified 2026-10-16. Primary UNESCO guidance. Official UNESCO record and publication verify authors, year, title and DOI; used for human agency/data governance principles, not intervention efficacy.

Source and authorship notes

Seed attachment Pasted markdown(20261016-073609).md, library file libfile_c8df81e3d25c819186a2df6c6d4dd2ce: social robotics and autonomous-classroom themes treated as commissioning ideas, not independently verified historical or efficacy evidence.

No claim is made that Cheteni founded the first autonomous classroom, established a particular Jibo intervention, or produced empirical findings; these claims were not necessary for this article and were not verified here.

Original synthesis, governance matrix, failure taxonomy and prospective protocol. No new participant data or completed intervention analysis. Not a systematic review; search completeness and quantitative synthesis are not claimed.

AI assistance supported source discovery, metadata auditing and manuscript development. Human author accountability, independent handling and human peer review must be recorded separately; no independent human review or acceptance is claimed by this draft.

References include primary experiments, a conference paper, methodological consensus guidelines and authoritative policy guidance. Citation verification establishes identity and stated claim scope, not replication or proof of effectiveness.

Fourteen references checked 2026-10-16. R7 has a linked publisher correction R8. The citation audit is not a guarantee of complete future correction/retraction surveillance.