

Teachers' Role in Fostering Digital Trust and Mediating Ethical Engagement with Artificial Intelligence - A Literature Review

**Lucianne Zammit^{1*}, Christian Colombo¹, May Agius¹,
Christine Fenech¹, Roger Tirazona¹**

¹University of Malta,

*lucianne.zammit@um.edu.mt

Abstract: The rapid expansion of Artificial Intelligence (AI) in education has intensified questions about how students develop informed and responsible relationships with algorithmic systems. While existing research has largely focused on teachers as adopters or users of AI technologies, far less attention has been given to teachers' role in guiding students' engagement with these systems. This literature review examines how teachers contribute to the development of digital trust in AI-mediated learning environments. Digital trust is conceptualised as the capacity to make informed, critical, and context-sensitive judgements about when and how AI systems should be relied upon. The review synthesises research across three interconnected areas: conceptualisations of digital trust, teachers' mediating practices in AI-supported learning, and the factors shaping teachers' own trust in AI technologies. It further examines systemic constraints on teachers' capacity to perform this mediating role, including gaps in professional preparation and governance frameworks. The Maltese educational context is analysed as a case study illustrating how policy ambition, curriculum reform, and teacher capacity-building interact in practice. The article argues that preparing young people for AI-saturated knowledge environments requires recognising teachers as central mediators of digital trust rather than merely users of educational technologies.

Keywords: digital trust; Artificial Intelligence; teacher mediation; AI Ethics Education

Introduction

The rapid proliferation of Artificial Intelligence (AI) technologies has transformed the educational landscape. Students encounter AI not merely as an abstract concept, but as an embedded presence in their daily lives, from recommendation algorithms shaping their social media feeds to generative AI

tools producing essays, artwork, and code. This ubiquity places teachers at the centre of a complex negotiation between technological capability and human development.

The concept of digital trust, understood as the confidence that individuals place in digital systems to operate reliably, ethically, and securely, has emerged as a critical concern in this landscape. For young people growing up surrounded by AI, developing an appropriate relationship with these technologies requires more than technical literacy. It demands the capacity to critically evaluate AI outputs, recognise biases and limitations, protect personal privacy, and make informed decisions about when and how to rely on automated systems. Teachers, as the primary mediators between students and educational technology, occupy a uniquely important position in fostering this digital trust.

Yet the role of teachers in this domain remains inadequately understood. While a growing body of research examines teachers' own adoption of and attitudes toward AI technologies, far less attention has been paid to teachers' function as guides who help students navigate the ethical complexities of an AI-saturated world. This review addresses this gap by synthesising the literature on AI in education through the lens of teachers' mediating role. It advances the concept of *teacher-mediated digital trust* to describe how educators help students navigate AI systems, the factors that influence teachers' own trust in AI, and the conditions that support or constrain teachers' capacity to foster digital trust among their students.

The article proceeds as follows. It outlines the review methodology, then examines the concept of digital trust and its relevance to AI-mediated learning environments. It then analyses teachers' mediating role, focusing on critical evaluation, ethical reflection, and responsible use of AI-generated knowledge. It considers the factors shaping teachers' own trust in AI, including self-efficacy, professional agency, and pedagogical relevance. It examines systemic constraints on mediation capacity, including gaps in professional preparation and governance. Malta is explored as a case study of how policy ambition, curriculum development, and teacher capacity-building interact. The article concludes with implications for policy, practice, and research.

Methodology

This article adopts a narrative literature review approach to synthesise existing scholarship on teachers' mediating roles in AI-related educational contexts. Narrative reviews suit emerging, conceptually diverse fields where the aim is interpretive synthesis rather than exhaustive aggregation of findings (Baumeister & Leary, 1997; Ferrari, 2015; Grant & Booth, 2009; Snyder, 2019).

Relevant literature was identified through structured searches of major academic databases, including Scopus, Web of Science, and Google Scholar. Searches used keywords including *Artificial Intelligence in education*, *AI literacy*, *digital trust*, *teacher trust in AI*, *AI ethics in education*, and *generative AI in schools*. Additional citation tracking identified frequently referenced studies. The review prioritises peer-reviewed publications produced between 2018 and 2025, a period of accelerated generative AI development and expanding educational research.

To situate the scholarly literature within broader governance and policy debates, the review also includes selected policy documents and conceptual frameworks produced by international organisations such as UNESCO and the European Commission. These sources provide influential guidance on AI governance, digital competence, and the professional responsibilities of educators. Including such sources is consistent with narrative approaches that integrate research and policy literature (Grant & Booth, 2009; Snyder, 2019).

The aim of the review is interpretive rather than exhaustive. Narrative literature reviews do not attempt to systematically catalogue all available studies but instead aim to identify key conceptual patterns, debates, and knowledge gaps within a body of literature (Baumeister & Leary, 1997; Ferrari, 2015). Accordingly, the analysis synthesises literature across three interconnected analytical dimensions that recur across the field:

1. conceptualisations of digital trust in AI-mediated environments,
2. teachers' mediating practices in supporting students' engagement with AI-generated knowledge, and
3. the institutional and professional conditions that shape teachers' capacity to guide students' ethical and critical engagement with AI.

The review seeks to clarify how teacher-mediated digital trust can help explain the evolving professional responsibilities of educators in AI-enabled learning environments.

Defining Digital Trust in AI

Digital trust is commonly defined as “the attitude that an agent will help achieve an individual’s goals in a situation characterised by uncertainty and vulnerability” (Lee & See, 2004, p. 51). Two elements are salient for education: trust is meaningful only under uncertainty, and involves willingness to rely on an agent despite recognised risk. In AI-mediated learning, these conditions are amplified as algorithmic systems shape access to information, feedback, and evaluation in opaque ways.

Within AI research, trust is operationalised across dimensions including performance (accuracy, reliability), integrity (fairness, adherence to ethical norms), benevolence (alignment with users’ interests), and openness (transparency, explainability) (Nazaretsky et al., 2025; Raza et al., 2025; Saveljeva & Volkova, 2025). These dimensions encompass moral and relational judgements about how AI systems operate. In educational contexts, trust cannot be reduced to system accuracy. European policy guidance similarly emphasises that AI integration in education must respect fundamental rights, promote fairness, and preserve learning integrity (European Commission, 2026). Students may place unwarranted trust in fluent but unreliable outputs, or conversely, reject tools that are statistically robust but poorly understood. For this reason, recent scholarship emphasises the need for *calibrated trust*, defined as an alignment between users’ reliance on AI systems and those systems’ actual capabilities and limitations (Cukurova & Miao, 2024; Fan et al., 2025). Both over-trust and under-trust are pedagogically problematic: the former risks epistemic complacency and cognitive offloading, while the latter may foreclose legitimate learning opportunities.

Trust in AI is also shaped by institutional and governance contexts. Weak safeguards around data privacy, surveillance, and accountability can erode trust even when AI systems perform effectively (Collins Umoke et al., 2025; Nazaretsky et al., 2025). In schooling, where students are structurally vulnerable and participation is compulsory, trust is inseparable from perceptions of institutional responsibility. Teachers therefore mediate trust not only in specific tools, but also in the educational system that authorises their use.

This positions AI literacy as a necessary, though insufficient, condition for digital trust. Without the capacity to interrogate AI outputs, recognise bias, understand probabilistic generation, and question provenance, students cannot develop informed judgements about when reliance is appropriate (Cukurova & Miao, 2024; Fan et al., 2025). This aligns with European frameworks defining AI literacy as the capacity to understand AI's opportunities and risks and to use these technologies responsibly (European Commission, 2026). Digital trust is thus not a stable attitude but a learned epistemic disposition requiring active cultivation and continual recalibration.

For the purposes of this review, *digital trust in AI* is therefore understood as students' and teachers' capacity to make informed, critical, and context-sensitive judgements about when and how to rely on AI systems in learning and decision-making. This foregrounds trust as a pedagogical and ethical achievement rather than a default response to technological adoption.

Teachers as Mediators of AI Experiences

This section examines teachers' role in mediating students' AI experiences and fostering digital trust.

As AI becomes embedded in educational practice and everyday life, teachers are increasingly positioned as mediators who shape the epistemic, ethical, and relational conditions under which students engage with AI-generated knowledge (Cukurova & Miao, 2024; Jose et al., 2025). This perspective is also reflected in European policy guidance, which emphasises that teachers and school leaders must evaluate whether AI tools align with pedagogical goals, learning objectives, and ethical principles before integrating them into educational practice (European Commission, 2026).

A key challenge concerns the destabilisation of epistemic authority. Machine-generated outputs often appear fluent and authoritative, obscuring the probabilistic nature of their generation and blurring the distinction between plausibility and reliable knowledge. Students may therefore struggle to evaluate the credibility and provenance of AI-generated information (Jose et al., 2025). Within this context, teachers increasingly assume the role of epistemic mentors, guiding students in interrogating AI outputs, identifying bias, and distinguishing plausibility from justified knowledge (Camilleri, 2023; Cukurova & Miao, 2024). This shifts attention from the technical operation of

AI toward cultivating critical reasoning. Recent work positions generative AI as a catalyst for rethinking pedagogy and assessment, requiring educators to actively shape how students engage with AI-generated knowledge (Montebello, 2025).

The literature also emphasises teachers as human-centred safeguards. AI systems may simulate dialogue and personalise content but lack the empathy and contextual sensitivity required for meaningful educational relationships. Consequently, teachers play a crucial role in ensuring that technological innovation does not undermine the relational and ethical foundations of education (Akman Yeşilel, 2025; Lombardi Calleja et al., 2024). This requires balancing the efficiency afforded by AI with the protection of student agency, wellbeing, and dignity, particularly where algorithmic decisions shape learning outcomes (Adams et al., 2022; Gouseti et al., 2025).

This role is often framed as human-in-the-loop governance, emphasising the need for human oversight in educational AI use (European Commission, 2026; Nazaretsky et al., 2025). In practice, teachers critically evaluate AI systems, determine pedagogical appropriateness, and maintain responsibility for instructional decisions. Teachers thus function not only as facilitators but as curators of the algorithmic environments in which learning takes place (Cukurova & Miao, 2024).

A further dimension concerns data protection. AI-based educational technologies rely on large quantities of student data, raising concerns about privacy, consent, and potential misuse. Teachers often act as the first line of defence, evaluating compliance with requirements such as the GDPR, ensuring data anonymisation, and securing consent where necessary (Akman Yeşilel, 2025; Cukurova & Miao, 2024). In doing so, they translate abstract regulatory principles into everyday classroom practice. Policy guidance reinforces this responsibility, noting that educators must understand how AI systems process student data and ensure that their use complies with regulatory frameworks such as the GDPR and the EU Artificial Intelligence Act (European Commission, 2026).

Teacher mediation also addresses the challenges generative AI poses to academic integrity. Plagiarism detection technologies have been criticised for limited reliability and high false-positive rates, and recent scholarship increasingly advocates for assessment designs that emphasise process,

reflection, and transparency in students' use of AI tools (Cotton et al., 2024; Evangelista, 2025; Leaton Gray et al., 2025). Strategies such as in-class writing, oral examinations, and reflective AI documentation encourage students to account critically for AI's role in their work. These shift the focus from punitive detection to pedagogical accountability.

The mediating role thus contributes to preserving the human-relational minimums of education (Jose et al., 2025). As AI systems assume a growing role in generating content, feedback, and recommendations, teachers remain responsible for providing the contextual judgement, ethical reasoning, and interpersonal understanding that automated systems cannot replicate (Kim, 2024; Cukurova & Miao, 2024). Teacher mediation is therefore a central professional responsibility, ensuring that technological innovation enhances rather than diminishes the integrity of educational practice.

Understanding Teacher Trust in AI

Teachers' capacity to foster digital trust among students is shaped by their own relationship with AI technologies. Trust influences not only whether teachers adopt AI tools, but how they integrate them pedagogically and ethically (Viberg et al., 2025). Teacher trust in AI is complex, multidimensional, and cannot be adequately captured by traditional technology adoption models alone.

Early research relied heavily on the Technology Acceptance Model (TAM), which emphasises perceived usefulness and ease of use (Davis, 1989). More recent studies argue this framework is insufficient for AI, as it overlooks uncertainty and ethical risk. Choi et al. (2023) demonstrate that *perceived trust* functions as a distinct mediator of adoption, shaping behavioural intention independently of usability. More contemporary scholarship adopts socio-cognitive models, defining trust as willingness to rely on an agent under vulnerability and uncertainty (Lee & See, 2004; Nazaretsky et al., 2025).

Empirical research identifies multiple trust dimensions relevant to teachers. Nazaretsky et al. (2022) developed a multi-factor instrument for compulsory schooling that distinguishes between performance-related factors (such as perceived benefit and self-efficacy) and ethical or relational concerns (including transparency, loss of human judgement, and lack of emotional understanding). Their findings reveal a consistent tension: teachers may recognise AI's pedagogical potential while simultaneously distrusting its

opacity and moral limitations. Trust, therefore, is not a unitary attitude but a negotiated judgement.

Pedagogical orientation further shapes trust. Teachers with constructivist beliefs are more likely to view AI as compatible with personalised and student-centred learning, whereas those with more transmissive orientations express greater reluctance (Choi et al., 2023). Nevertheless, a persistent pattern across studies is algorithm aversion: teachers consistently report higher trust in human expertise than in AI systems, even when AI outputs are demonstrably accurate (Akman Yeşilel, 2025; Nazaretsky et al., 2022). This could reflect concerns about de-professionalisation and the erosion of pedagogical judgement rather than fear of technological failure.

Importantly, research distinguishes between general digital competence and AI-specific literacy. Lucas et al. (2024) identify a strong positive relationship between teachers' knowledge of AI and their trust in AI systems, indicating that generic digital skills are insufficient for confident and ethical adoption. Teachers who lack AI-specific understanding are more likely to adopt restrictive or avoidance-based mediation strategies, potentially discouraging responsible student engagement (Lucas et al., 2024; Marín et al., 2025).

Contrary to common assumptions, demographic variables such as age, gender, and teaching experience are not strong predictors of trust in AI (Akman Yeşilel, 2025; Viberg et al., 2025). However, some evidence suggests that highly experienced teachers may exhibit lower trust in AI's diagnostic or evaluative functions, possibly due to heightened sensitivity to pedagogical nuance and error (Akman Yeşilel, 2025).

Teacher anxiety surrounding AI further complicates the trust-adoption relationship. While teachers express concerns about privacy, bias, surveillance, and ethics, anxiety alone does not reliably predict resistance to AI use (Ayanwale et al., 2022). Instead, trust appears to be shaped by perceived relevance, institutional expectations, and professional agency. Teachers may engage with AI despite unresolved concerns, particularly when adoption is implicitly required. This underscores a critical limitation in the literature: trust does not guarantee ethical or pedagogically sound practice. Without sufficient agency, transparency, and support, teachers risk deferring to algorithmic authority rather than mediating it.

Thus, research on teacher trust indicates that confidence in AI is constructed through knowledge, perceived competence, and supportive institutional conditions. Trust is best understood as a prerequisite for active mediation, enabling teachers to engage critically with AI while retaining responsibility for judgement and student learning. Importantly, this is a distinct objective from equipping teachers to use AI confidently themselves: a teacher may achieve sufficient trust in AI tools for personal adoption while still lacking the pedagogical preparation needed to guide students' critical engagement with them. This distinction – between teacher trust in AI and teacher capacity to mediate student trust in AI – is examined as a structural gap in the following section.

Gaps in Teacher Capacity and Systemic Barriers

The literature reveals a persistent mismatch between the expectation that teachers mediate AI engagement and their actual capacity to do so. While policy positions teachers as ethical guides, empirical research shows many educators feel underprepared to navigate AI-related risks. This concern is echoed in European policy discussions, which recognise that teachers often feel uncertain about rapidly evolving AI technologies and therefore require structured professional learning opportunities to build confidence in their use (European Commission, 2026).

Teachers report anxiety related to privacy, surveillance, bias, and academic integrity, alongside concerns about the erosion of human connection and professional judgement (Ayanwale et al., 2022; Akman Yeşilel, 2025). These concerns are not merely attitudinal but structural. Marín et al. (2025) and Collins Umoke et al. (2025) argue that in the absence of clear governance frameworks, teachers are left to manage high-stakes ethical decisions individually, amplifying uncertainty and risk aversion.

Time pressure emerges as a dominant systemic barrier, with educators struggling to acquire AI-specific knowledge alongside existing demands (Chounta et al., 2022). Lucas et al. (2024) observe that educators are placed under a dual burden: they are expected to gain mastery over rapidly evolving AI tools while simultaneously teaching students about them. This results in what the International Task Force on Teachers for Education 2030 (2025) describes as "incidental learning," where teachers acquire AI knowledge in a fragmented, ad-hoc manner rather than through structured professional development. Consequently, many educators rely on self-exploration without

clear institutional guidance, leading to misconceptions and uneven implementation (Yuan et al., 2025).

The governance landscape compounds these difficulties. Policy frameworks addressing AI ethics, data protection, and accountability frequently lag behind classroom implementation (Gouseti et al., 2025; Marín et al., 2025). Without institutional clarity regarding acceptable use, liability, and safeguards, teachers are required to interpret abstract ethical principles in situ, often without legal or professional protection. This contributes to cautious or defensive practices, particularly in relation to assessment and student data use.

Critically, existing research remains skewed towards teachers as users or adopters of AI tools, rather than as mediators who cultivate students' capacity for critical evaluation and ethical judgement. Although AI literacy is increasingly promoted in policy rhetoric, AI ethics education remains underprioritised in classroom practice, and rarely assessed in ways that capture students' reasoning or decision-making (Ma et al., 2025; Wiese et al., 2025). Consequently, there is limited empirical insight into teachers' confidence, strategies, or perceived competence in guiding students through epistemic uncertainty, bias recognition, and privacy protection.

When mediation capacity is underdeveloped, teachers may default to restrictive approaches or uncritical adoption, both undermining calibrated digital trust. Addressing this requires shifting attention from technology adoption to professional agency and the institutional conditions enabling informed judgement. Recent research on AI literacy education highlights teacher preparation and access to accessible pedagogical tools as critical factors shaping the successful implementation of AI-related learning in schools (Busuttil, 2025a).

The challenge facing education systems is not lack of willingness to engage with AI, but lack of structured support for the professional judgement that mediation requires. Where policy ambition outpaces capacity-building, the burden of managing trust and risk falls on individual teachers. Recent European policy frameworks, therefore, stress the need to equip educators with the knowledge, skills, and institutional support required to critically evaluate AI systems and guide students' ethical engagement with them (European Commission, 2026). The following section examines Malta as a case study of this dynamic, illustrating how a dense policy environment and strong

societal endorsement of AI intersect with uneven mediation capacity at the classroom level.

The Maltese Context

This section examines Malta as a case study, foregrounding the dynamics of small-state implementation, where global AI discourses intersect with a highly centralised education system. The section is structured across three interconnected levels: policy and societal expectations, subject-specific classroom enactments, and teacher capacity-building through initial and ongoing professional education. Together, these levels illuminate how teacher mediation becomes both structurally demanded and unevenly supported in practice.

Policy ambition and societal expectations

Malta has positioned itself as an enthusiastic adopter of AI-enabled innovation at national and educational policy levels. Malta's Ethical AI Framework articulates an ambition to establish the country as a global leader in the field (Malta.AI Taskforce, 2019), while the Digital Education Strategy 2025–2030 (Directorate for Digital Literacy and Transversal Skills, 2025) prioritises AI literacy and transversal skills. This is reinforced by the Eurobarometer survey, indicating strong public agreement in Malta that digital skills are essential for societal participation and should be embedded across schooling and higher education. Attitudes are characterised by openness, with only 11% of respondents indicating AI does not belong in education (European Commission, 2025).

However, favourable attitudes coexist with high expectations of schools and teachers. The same Eurobarometer data show widespread agreement that teachers should play a central role in supporting students' safe, responsible, and critical engagement with digital technologies, including AI, as well as in mitigating risks related to wellbeing, disinformation, and misuse. Analytically, this combination of openness and expectation is significant. Acceptance of AI is not accompanied by a relaxation of professional responsibility. Instead, teachers are implicitly positioned as the primary mediators of trust, risk, and ethical use, even as policy rhetoric emphasises innovation and experimentation (European Commission, 2025). This dynamic reinforces a distinction developed earlier in the review between acceptance, trust, and readiness. Low resistance to AI may enable rapid uptake, but without sustained pedagogical

investment it also heightens the risk of uncritical adoption. In this sense, Malta exemplifies how policy ambition and societal endorsement can intensify, rather than resolve, the demand for teacher mediation. This pattern reflects what the international literature identifies as structural demand displacement: professional responsibilities multiply faster than the support structures available to sustain them.

Within the Catholic Church school sector, which serves a substantial proportion of the Maltese student population, AI integration is explicitly framed through the lens of human-centred education (Lombardi Calleja et al., 2024). Rather than imposing bans, leadership in this sector has increasingly adopted policies that embrace AI while foregrounding human dignity and ethical responsibility.

Subject-specific classroom enactments

At the classroom level, subject-specific research illustrates how these policy and societal dynamics translate into pedagogical strain. Evidence from secondary education, particularly within informatics, highlights both opportunity and disruption. Teachers participating in workshops on large language models reported enhanced student engagement and support for personalised learning, alongside persistent concerns regarding assessment validity, student overreliance, and the erosion of critical evaluation skills (Busuttil, 2025b).

Other studies identify a clear implementation gap between curricular ambition and classroom reality. While the ICT C3 curriculum is content-innovative, Busuttil and Vassallo (2025) report that only 43% of ICT teachers feel confident teaching advanced topics such as AI and cybersecurity. This suggests that curriculum reform has outpaced system-wide capacity-building, a challenge likely to be exacerbated in non-state schools where resource constraints may amplify inequities (Vassallo et al., 2025).

An expanding body of Maltese scholarship documents AI-related educational experimentation across sectors and subject areas. Research has examined AI in early years classrooms (Azzopardi, 2025; Calleja & Camilleri, 2025; Mifsud et al., 2025), generative AI in Maltese language education (Żammit, 2025; Oliveri, 2025; Bezzina & Wright, 2025), AI-supported science teaching (Schembri Frendo & Vassallo, 2025; Spiteri & Dingli, 2024; Zammit et al., 2022), and

system-level analyses of AI adoption and narratives (Aquilina, 2025; Pulis & Mallia, 2025).

Research focusing on teacher beliefs further highlights variation linked to pedagogical orientation. Busuttil and Calleja (2025) found that mathematics teachers with discovery-oriented or connectionist philosophies were more likely to view ChatGPT as a tool for fostering active learning, while expressing concern about its limitations in handling visual representations and the risk of bypassing conceptual understanding.

Taken together, this body of work points to rapid innovation accompanied by conceptual and ethical fragmentation. Concerns related to data privacy, bias, learner autonomy, assessment integrity, and pedagogical purpose recur across disciplines (Aquilina, 2025; Pulis & Mallia, 2025). Crucially, these concerns are not subject-specific. Rather, they underscore the centrality of teacher mediation in translating policy intent, ethical principles, and technological affordances into coherent classroom practice. This variation reinforces the finding that trust in AI is shaped by pedagogical orientation rather than technical familiarity, suggesting mediation capacity must be cultivated within disciplinary contexts.

Teacher capacity-building through initial and ongoing professional education

Empirical research reveals a nuanced picture of cautious optimism and uneven confidence among teachers (Busuttil & Calleja, 2025; Vassallo et al., 2025). Although low resistance to AI may appear encouraging, it risks obscuring a deeper pedagogical issue: openness does not equate to readiness to teach with, about, or through AI in ethically robust ways. In the absence of sustained professional preparation, acceptance may accelerate uncritical adoption, intensifying epistemic, safeguarding, and integrity risks.

Across studies, continuous professional development emerges as the decisive factor in strengthening teacher agency. Busuttil and Calleja (2025) demonstrate that hands-on workshops can shift perceptions, reframing AI as a pedagogical resource rather than a disruptive threat. However, they emphasise that isolated interventions are insufficient. Sustained, pedagogy-driven training is required to address misconceptions and embed ethical reasoning in practice. Lombardi Calleja et al. (2024) describe a multiplier approach within Church schools, whereby school leaders and heads of department are trained as mentors who

support professional learning communities, providing ongoing, contextualised guidance.

Evidence from initial teacher education further illustrates how mediation capacity can be cultivated. Research on pre-service teachers' engagement with ChatGPT through assessment-as-learning frameworks shows that AI-generated materials function primarily as provisional scaffolds rather than pedagogically complete artefacts (Busuttil, 2025c). Participants reported limitations relating to linguistic accuracy, differentiation, and classroom realism, necessitating substantial human judgement and revision.

These constraints did not result in rejection but prompted reflective evaluation and deeper engagement with pedagogical principles. Assessment emerged as a critical site where the inadequacy of unmediated AI use became visible, requiring pre-service teachers to calibrate reliance and align instructional decisions with contextual realities. This provides empirical support for the claim that generative AI reconfigures, rather than diminishes, teacher agency.

Strategic alignment and implementation challenges

Recent policy developments indicate a growing effort to consolidate AI education initiatives within a coherent strategic framework. The Digital Education Strategy 2025–2030 aims to standardise the development of AI literacy and transversal digital competencies across all sectors of the education system (Directorate for Digital Literacy and Transversal Skills, 2025). The strategy emphasises collaboration between the Ministry for Education, MDIA, and tertiary institutions to ensure innovation is accompanied by pedagogical and ethical guidance. The strategy thus recognises that teacher preparation must extend beyond technical skills to include critical and ethical engagement (Lombardi Calleja et al., 2024).

Within the compulsory education system, this strategic alignment is reflected in a number of curricular developments. At early years level, the Digital Education Strategy promotes the development of responsible digital literacy, encouraging the use of age-appropriate digital storytelling and exploratory tools while maintaining strong safeguarding practices. In the later years of primary education, Action 1.1.3 of the strategy establishes that by Year 6, students should acquire a basic understanding of AI technologies and their everyday applications, including image generation, text-to-speech systems, and algorithmically mediated search (Directorate for Digital Literacy and

Transversal Skills, 2025). These developments signal a shift from viewing digital competence primarily as an operational skill toward a broader conception of technological understanding that includes ethical awareness and critical evaluation. At secondary level, this policy direction is reflected in the ICT C3 curriculum, which explicitly introduces emerging technologies such as Artificial Intelligence, blockchain, and robotics alongside cybersecurity and digital ethics (Catania, 2023).

Within the Maltese curriculum, the Ethics subject provides a rare institutional space where the moral implications of emerging technologies are explicitly examined, positioning ethical reflection as a core component of AI literacy. The Ethics programme emphasises concepts such as responsibility, integrity, consequences, and critical decision-making, which are increasingly applied to students' digital lives and interactions with emerging technologies. For example, modules such as *Responsible Use of Innovative Technology* invite students to examine ethical dilemmas involving algorithmic bias, misinformation, and the societal implications of automated systems (Tirazona et al., 2024). Ethics lessons thus provide structured opportunities for students to develop the reflective judgement necessary for responsible digital participation.

This alignment between technology-focused and values-oriented curricula illustrates a shift towards a whole-system approach to AI literacy, in which ethical reflection and critical thinking complement technical competence across subjects.

Despite these promising developments, tension between strategic ambition and implementation capacity persists (Vassallo et al., 2025). Teacher preparedness to translate AI literacy and digital ethics into practice remains uneven. Teachers require not only technological familiarity but the pedagogical competence to guide students in evaluating algorithmic outputs, recognising bias, and navigating privacy and accountability (Nazaretsky et al., 2022; Jose et al., 2025).

In this small-state context, where policy networks are dense and curricular innovation rapid, these gaps are particularly visible. Without sustained professional learning, institutional guidance, and opportunities for educators to engage critically with AI within their own disciplinary contexts, the responsibility for cultivating digital trust risks being unevenly distributed

across schools and individual teachers, potentially widening disparities in students' opportunities to develop informed relationships with AI technologies.

Conclusion: Implications for Policy, Practice, and Research

Synthesising international and Maltese scholarship, this review demonstrates that the educational significance of AI lies not primarily in its technical capabilities, but in the ways human judgement, ethical reasoning, and professional responsibility are exercised around its use.

Digital trust – the capacity for informed, context-sensitive judgements about when and how to rely on AI – constitutes a pedagogical and ethical task that must be actively supported, as both uncritical reliance and categorical rejection undermine meaningful learning. Teachers occupy a critical mediating role in this process: as epistemic guides supporting students in interrogating AI-generated outputs, recognising bias, and maintaining intellectual agency, and as ethical gatekeepers translating principles of privacy, accountability, and fairness into classroom practice.

Educators' trust in AI is shaped less by demographic factors than by AI-specific knowledge, professional agency, and institutional conditions. Importantly, trust alone does not guarantee ethically robust practice. Where governance frameworks are unclear and professional preparation is fragmented, teachers may either defer to algorithmic authority or adopt restrictive approaches that limit meaningful engagement with AI. In both cases, opportunities to cultivate calibrated digital trust among students are weakened.

The Maltese case confirms and sharpens this finding. Despite sophisticated policy frameworks, high public acceptance, active curricular reform, and emerging whole-system alignment between technology-focused and values-oriented subjects, the mediation deficit identified across the international literature reproduces itself at the classroom level. Malta thus demonstrates that no combination of policy ambition and societal endorsement substitutes for the sustained, contextualised professional formation of teachers as mediators of digital trust.

A significant gap remains in the literature: while considerable research examines teachers as adopters of AI, far less attention addresses teachers as mediators who cultivate students' capacity to critically evaluate algorithmic

systems, recognise bias, protect their privacy, and navigate ethical dilemmas in AI-mediated environments. Closing this gap calls for a reorientation of research toward teacher mediation as a professional practice in its own right – examining how the epistemic, ethical, and relational dimensions of mediation develop across subject areas and career stages, what forms of professional education most effectively cultivate them, and how governance frameworks can create the institutional conditions they require. Further research is needed to examine how students in Malta understand, trust, and adopt AI tools in their learning. Evidence on students’ trust judgements and patterns of use would help inform teachers’ professional development, particularly in relation to AI literacy and the mediation of students’ engagement with AI-generated knowledge.

The key question is not whether teachers should use AI, but whether systems support them in guiding students’ engagement with it. The evidence consistently suggests they are not: professional preparation is fragmented, governance frameworks lag behind classroom realities, and the mediating role of teachers remains undertheorised and underresourced. This is not a technical adjustment but a structural priority. Where education systems do not deliberately cultivate teachers as mediators of digital trust, the formation of students’ relationships with AI will be shaped by platform design, algorithmic defaults, and peer influence – rather than by pedagogical intention.

This review advances teacher-mediated digital trust as a framework for research, policy, and professional practice. Its core claim is that calibrated, critical engagement with AI is not a disposition students develop independently; it is cultivated through sustained pedagogical relationships with teachers who are themselves epistemically confident, ethically grounded, and institutionally supported. Recognising teachers as mediators reframes AI integration from a question of tool provision to one of professional formation, institutional design, and the values education systems choose to enact. How those values are articulated, resourced, and protected will determine whether AI in education deepens or diminishes the quality of human judgement it is supposed to support.

References

Adams, C., Pente, P., Lemermeyer, G., Turville, J., & Rockwell, G. (2022). Artificial intelligence and teachers’ new ethical obligations. *International Review of Information Ethics*, 31(1). <https://doi.org/10.29173/irrie483>

- Akman Yeşilel, D. B. (2025). English language teachers' perspectives on digital trust in AI-supported education. *British Educational Research Journal*. <https://doi.org/10.1002/berj.4209>
- Aquilina, K. (2025). AI adoption among Europe's school evaluators: Awareness and challenges. *Malta Journal of Education*, 06(01). <https://doi.org/10.62695/tqqm2623>
- Ayanwale, M. A., Sanusi, I. T., Adelana, O. P., Aruleba, K. D., & Oyelere, S. S. (2022). Teachers' readiness and intention to teach Artificial Intelligence in schools. *Computers and Education: Artificial Intelligence*, 3. <https://doi.org/10.1016/j.caeai.2022.100099>
- Azzopardi, S. (2025). Whimsical wonders or digital dangers: Exploring the influence of Artificial intelligence during story reading in an early years classroom. *Malta Journal of Education*, 06(01). <https://doi.org/10.62695/wsri5131>
- Baumeister, R. F., & Leary, M. R. (1997). Writing narrative literature reviews. *Review of General Psychology*, 1(3), 311–320. <https://doi.org/10.1037/1089-2680.1.3.311>
- Bezzina, A., & Wright, E. (2025). Using Artificial intelligence for personal, social and career development: Possibilities and challenges. *Malta Journal of Education*, 06(01). <https://doi.org/10.62695/khrp4909>
- Busuttil, L. (2025a). Constructing AI literacy: A hands-on approach for school children. *Informatics in Education*, 24(4), 761–779. <https://doi.org/10.15388/infedu.2025.26>
- Busuttil, L. (2025b). Large language models in secondary informatics education: Teachers' perspectives. 18th International Conference on Informatics in Schools: Situation, Evolution, and Perspectives, ISSEP 2025, Germany. 29-40.
- Busuttil, L. (2025c). Leveraging generative AI in pre-service teacher training: Insights into assessment as learning. In *International Conference on Higher Education Learning Methodologies and Technologies Online* (pp. 46-60). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-94002-6_4
- Busuttil, L., & Calleja, J. (2025). Teachers' beliefs and practices about the potential of ChatGPT in teaching mathematics in secondary schools. *Digital Experiences in Mathematics Education*, 11(1), 140–166. <https://doi.org/10.1007/s40751-024-00168-3>
- Busuttil, L., & Vassallo, D. (2025). Evaluating the ICT C3 curriculum in Maltese schools: teacher perspectives. *Malta Review of Educational Research*, 19(1), 37-65.
- Calleja, J., & Camilleri, P. (2025). Primary school teachers' perceptions towards the use of generative AI in teaching using lesson study. *International Journal for Lesson and Learning Studies*, 14(3), 237–252. <https://doi.org/10.1108/IJLLS-11-2024-0268>
- Camilleri, P. (2023). A future past? Rationalising the formalisation of an ethical objective in teacher preparation today for the machine autonomy of tomorrow. In J. Casas-Roma, J. Conesa, & S. Caballe (Eds.), *Technology, users and uses. Ethics and human interaction through technology and AI* (pp. 336–363). Ethics International Press Ltd.

- Catania, J. (2023). *ICT C3 Syllabus*. Government of Malta. <https://c3.skola.edu.mt/syllabus/>
- Choi, S., Jang, Y., & Kim, H. (2023). Influence of pedagogical beliefs and perceived trust on teachers' acceptance of educational Artificial Intelligence tools. *International Journal of Human-Computer Interaction*, 39(4), 910–922. <https://doi.org/10.1080/10447318.2022.2049145>
- Chounta, I. A., Bardone, E., Raudsep, A., & Pedaste, M. (2022). Exploring teachers' perceptions of Artificial Intelligence as a tool to support their practice in Estonian K-12 Education. *International Journal of Artificial Intelligence in Education*, 32(3), 725–755. <https://doi.org/10.1007/s40593-021-00243-5>
- Collins Umoke, C., Nwangbo, S. O., & Onwe, O. A. (2025). Stakeholder trust and AI in education: A policy perspective on data privacy, bias, and decision-making. *IIARD Journals*, 8(3), 54–68. <https://doi.org/10.56201/rjpst.vol.8.no3.2025.pg54.68>
- Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. *Innovations in Education and Teaching International*, 61(2), 228–239. <https://doi.org/10.1080/14703297.2023.2190148>
- Cukurova, M., & Miao, F. (2024). *AI competency framework for teachers*. UNESCO Publishing.
- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319–340.
- Directorate for Digital Literacy and Transversal Skills. (2025). *Digital education strategy 2025-2030*. Government of Malta.
- European Commission. (2025). *Future needs in digital education*. Publications Office of the European Union. <https://doi.org/10.2766/6684184>
- European Commission. (2026). *Guidelines on the ethical use of artificial intelligence and data in teaching and learning for educators* (Revised edition). Publications Office of the European Union. <https://doi.org/10.2766/7967834>
- Evangelista, E. D. L. (2025). Ensuring academic integrity in the age of ChatGPT: Rethinking exam design, assessment strategies, and ethical AI policies in higher education. *Contemporary Educational Technology*, 17(1). <https://doi.org/10.30935/cedtech/15775>
- Fan, Y., Tang, L., Le, H., Shen, K., Tan, S., Zhao, Y., Shen, Y., Li, X., & Gašević, D. (2025). Beware of metacognitive laziness: Effects of generative artificial intelligence on learning motivation, processes, and performance. *British Journal of Educational Technology*, 56(2), 489–530. <https://doi.org/10.1111/bjet.13544>
- Ferrari, R. (2015). Writing narrative style literature reviews. *Medical Writing*, 24(4), 230–235. <https://doi.org/10.1179/2047480615Z.000000000329>
- Gouseti, A., James, F., Fallin, L., & Burden, K. (2025). The ethics of using AI in K-12 education: A systematic literature review. *Technology, Pedagogy and Education*, 34(2), 161–182. <https://doi.org/10.1080/1475939X.2024.2428601>

- Grant, M. J., & Booth, A. (2009). A typology of reviews: An analysis of 14 review types and associated methodologies. *Health Information and Libraries Journal*, 26(2), 91–108. <https://doi.org/10.1111/j.1471-1842.2009.00848.x>
- International Task Force on Teachers for Education 2030. (2025). *Promoting and protecting teacher agency in the age of artificial intelligence*. UNESCO Publishing.
- Jose, B., Cleetus, A., Joseph, B., Joseph, L., Jose, B., & John, A. K. (2025). Epistemic authority and generative AI in learning spaces: Rethinking knowledge in the algorithmic age. *Frontiers in Education (Lausanne)*, 10. <https://doi.org/10.3389/educ.2025.1647687>
- Kim, J. (2024). Types of teacher-AI collaboration in K-12 classroom instruction: Chinese teachers' perspective. *Education and Information Technologies*, 29(13), 17433–17465. <https://doi.org/10.1007/s10639-024-12523-3>
- Leaton Gray, S., Edsall, D., & Parapadakis, D. (2025). AI-based digital cheating at university, and the case for new ethical pedagogies. *Journal of Academic Ethics*, 23(4), 2069–2086. <https://doi.org/10.1007/s10805-025-09642-y>
- Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80.
- Lombardi Calleja, N., Trapani Maggi, R., & Gové, J. (2024). Navigating AI in Catholic church schools: Leadership transformation and ethical Considerations. In M. Cutajar, & C. Bonello (Eds.), *Educational Leadership in Transition: Innovative Perspectives and Practices for the 21st Century, in Malta and Beyond* (pp. 239–264). Malta University Distributors.
- Lucas, M., Zhang, Y., Bem-haja, P., & Vicente, P. N. (2024). The interplay between teachers' trust in artificial intelligence and digital competence. *Education and Information Technologies*, 29(17), 22991–23010. <https://doi.org/10.1007/s10639-024-12772-2>
- Ma, M., Ng, D. T. K., Liu, Z., & Wong, G. K. W. (2025). Fostering responsible AI literacy: A systematic review of K-12 AI ethics education. *Computers and Education: Artificial Intelligence*, 8, 100422. <https://doi.org/10.1016/j.caeai.2025.100422>
- Malta.AI Taskforce (2019). *Malta: Towards trustworthy AI*. Parliamentary Secretariat for Financial Services, Digital Economy and Innovation, Office of the Prime Minister. <https://mdia.gov.mt/national-strategies/malta-ai-strategy-and-vision/>
- Marín, Y. R., Caro, O. C., Rituay, A. M. C., Llanos, K. A. G., Perez, D. T., Bardales, E. S., Tuesta, J. N. A., & Santos, R. C. (2025). Ethical challenges associated with the use of Artificial Intelligence in university education. *Journal of Academic Ethics*, 23(4), 2443–2467. <https://doi.org/10.1007/s10805-025-09660-w>
- Mifsud, C. L., Bonello, C., & Kucirkova, N. I. (2025). Exploring the multifaceted roles of social robots in early childhood literacy lessons: Insights from a Maltese classroom. *International Journal of Social Robotics*, 17(7), 1235–1249. <https://doi.org/10.1007/s12369-025-01290-x>
- Montebello, M. (2025). Rethinking pedagogy and assessment in the age of generative AI. In *Artificial intelligence in education: Creating an equitable, creative, and effective learning environment*. IntechOpen. <https://doi.org/10.5772/intechopen.1012982>

- Nazaretsky, T., Ariely, M., Cukurova, M., & Alexandron, G. (2022). Teachers' trust in AI-powered educational technology and a professional development program to improve it. *British Journal of Educational Technology*, 53(4), 914–931. <https://doi.org/10.1111/bjet.13232>
- Nazaretsky, T., Mejia-Domenzain, P., Swamy, V., Frej, J., & Käser, T. (2025). The critical role of trust in adopting AI-powered educational technology for learning: An instrument for measuring student perceptions. *Computers and Education: Artificial Intelligence*, 8. <https://doi.org/10.1016/j.caeai.2025.100368>
- Oliveri, F. (2025). *Maltese spelling and error correction*. [Unpublished Bachelor's Dissertation]. University of Malta.
- Pulis, A., & Mallia, M. (2025). The narratives of Artificial Intelligence: A critical view of an emerging tool. *Malta Journal of Education*, 6(02), 123–136. <https://doi.org/10.62695/rrui1032>
- Raza, S., Sapkota, R., Karkee, M., & Emmanouilidis, C. (2025). *TRiSM for Agentic AI: A review of trust, risk, and security management in LLM-based agentic Multi-Agent Systems*. <http://arxiv.org/abs/2506.04133>
- Saveljeva, J., & Volkova, T. (2025). A survey on digital trust: Towards a validated definition. *Digital*, 5(2), 14.
- Schembri Frendo, C., & Vassallo, D. (2025). Educational neuroscience meets AI: A framework for secondary science teaching. *Malta Journal of Education*, 06(01). <https://doi.org/10.62695/rrui1028>
- Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104, 333–339. <https://doi.org/10.1016/j.jbusres.2019.07.039>
- Spiteri, B. J., & Dingli, A. (2024). Enhancing creativity through personalised AI Tutoring: The DAVE system. 17th Annual International Conference of Education, Research and Innovation, 383–390. <https://doi.org/10.21125/iceri.2024.0198>
- Tirazona, R., Caruana, K., & Zammit, L. (2024). *Ethics Year 8 Syllabus: Consequences, Integrity, and Responsibility*. Ministry for Education, Sport, Youth, Research, and Innovation. https://curriculum.gov.mt/wp-content/uploads/2024/09/syllabus_ethics_yr8_2024.pdf
- Vassallo, D., Govus, G., & Busuttil, L. (2025). Between ambition and implementation: a comparative study of Malta's ICT C3 curriculum in the context of small-state digital education. *Journal of Information Technology Education*. <https://doi.org/10.28945/5674>
- Viberg, O., Cukurova, M., Feldman-Maggor, Y., Alexandron, G., Shirai, S., Kanemune, S., Wasson, B., Tømte, C., Spikol, D., Milrad, M., Coelho, R., & Kizilcec, R. F. (2025). What explains teachers' trust in AI in education across six countries? *International Journal of Artificial Intelligence in Education*, 35(3), 1288–1316. <https://doi.org/10.1007/s40593-024-00433-x>
- Wiese, L. J., Patil, I., Schiff, D. S., & Magana, A. J. (2025). AI ethics education: A systematic literature review. *Computers and Education: Artificial Intelligence*, 8, 100405. <https://doi.org/10.1016/j.caeai.2025.100405>

- Yuan, S., Mohamad Nasri, N., & Jamaludin, K. A. (2025). Exploring EFL lecturers' needs for AI-assisted teaching strategies: A qualitative basis for module development in Henan, China. *Forum for Linguistic Studies*, 7(12). <https://doi.org/10.30564/fls.v7i12.12396>
- Zammit, M., Voulgari, I., Liapis, A., & Yannakakis, G. (2022). *The road to AI literacy education: From pedagogical needs to tangible game design*. European Conference on Games Based Learning. <https://www.um.edu.mt/library/oar/handle/123456789/80765>
- Żammit, J. (2025). Secondary school teachers' experiences with generative AI in Maltese language teaching. *Technology, Knowledge and Learning*, 30(4), 2103–2121. <https://doi.org/10.1007/s10758-025-09861-7>

Acknowledgements

This work was funded by the Malta Digital Innovation Authority (MDIA) Applied Research Grant (MARG). The funded project is entitled “Education for Technology with Human Responsibility, Integrity, Values, and Ethics” with the acronym wE-THRIVE. The support of the project collaborators who contributed to the broader research agenda on ethics, digital technologies, and education is gratefully acknowledged.