Reality-Aligned Intelligence (RAI) in Education

Principles and Safeguards for AI Use in Schools

Author: **Niels Bellens** (independent researcher)
ORCID: **0009-0008-1764-4108**
Contact: **niels.bellens@proton.me**

Date: 27 November 2025

Abstract

AI systems are rapidly entering classrooms, homework routines and educational platforms. Students use chat-based “tutors”, auto-marking tools, learning analytics dashboards and proctoring systems—often without a clear understanding of what kind of thing these systems are, what they can and cannot do, and how they are embedded in institutional incentives.

This whitepaper applies **Reality-Aligned Intelligence (RAI)** and **Reality Alignment Theory (RAT)** to **AI in compulsory and secondary education**. It focuses on the gap between a system’s **Nature N(S)**—its actual structure, capabilities and constraints—and its **Representation R(S)**—how it presents itself and is experienced by learners and educators. Using concepts such as **Ontological Honesty OH(S)**, **Anthropomorphism Risk A(S)**, the **Ontological Integrity Line (OIL)** between tools and persons, and **Integrity Zones (IZ)** reflecting stakes and vulnerability, the paper treats misrepresentation as a central risk in educational AI.

Key contributions:

- A **RAT scan of education systems**, highlighting N/R gaps between mass-standardised, assessment-driven structures and narratives of flourishing, equality and care.
- A **RAI foundations layer** for schools, with education-specific interpretations of N(S), R(S), OH(S), A(S), OIL and IZ, and explicit red-line roles for AI with minors (no friend, therapist, romantic partner, spiritual oracle or secret-keeper).
- A **typology of educational AI** (content tools, tutors, study coaches, analytics, proctoring, decision support) and a **role taxonomy** for AI interacting with learners (explainer, exercise coach, study skills helper vs quasi-teacher, quasi-friend, surrogate therapist).

- **RAI-aligned design patterns** for educational AI, including ontological honesty requirements, anthropomorphism boundaries, emotional spillover handling, and a generic RAI architecture stack for school tools.
- A governance section detailing the roles of **schools and teachers, EdTech vendors, ministries and regulators, platforms, parents, students and civil society** in keeping AI on the tool side of OIL and aligned with educational responsibilities.
- **Scenarios and checklists** illustrating misaligned vs RAI-aligned adoption of AI tutors and analytics, and providing practical questions for schools, vendors and families.

The central thesis is that, in education, AI should be designed and governed as a **learning tool**, not a surrogate educator, friend or therapist. This requires:

- honest communication about AI's nature, limits and data use,
- bounded roles that avoid deep attachment and identity-defining influence, especially for minors,
- and governance structures that preserve the primacy of human teachers, parents and counsellors.

The paper connects to the broader RAI/RAT library, including **RAI for Minors**, the **RAI Tutor Spec for Minors**, **RAI Governance & Ecosystems**, **RAI for Mental Health AI (Adults)**, the **EMA Casebook**, and the **BADDASS Framework** for neurodivergent learners. It is intended for educators, school leaders, EdTech developers, policymakers, child-rights and digital-wellbeing advocates, and researchers interested in AI, education and governance.

Audience and Aims

This whitepaper is written for:

- **Educators and school leaders** seeking principled ways to adopt AI in classrooms.
- **EdTech vendors and AI developers** designing tools for minors.
- **Policymakers and regulators** working on AI-in-education guidelines and standards.
- **Child-rights organisations and digital wellbeing advocates** concerned with minors' interaction with AI.
- **Researchers** in AI ethics, education, HCI and governance.

It aims to:

- provide a clear conceptual lens (RAI/RAT) for thinking about AI in education,
- offer practical patterns and red lines for the design of educational AI,
- support governance efforts that keep AI as an honest tool in reality-aligned education systems.

Position in the RAI Library

“Reality-Aligned Intelligence (RAI) in Education: Principles and Safeguards for AI Use in Schools” is part of the open RAI/RAT research corpus, alongside:

- **RAI Metaframework** – core definitions and mathematical notes.
- **RAI Governance & Ecosystems** – institutional alignment and power.
- **RAI for Minors** – OIL and red-line roles for children and adolescents.
- **RAI Tutor Spec – Minors** – concrete design for a RAI-aligned tutor.
- **RAI for Mental Health AI (Adults)** – role boundaries and safeguards in mental health contexts.
- **EMA Casebook** – anthropomorphism and relational drift with an AI companion.
- **RAT for Humans and the Three Laws** – applying RAT to human self-story and ethics.
- **The BADDASS Framework** – tools for neurodivergent minds navigating complex systems.

Together, these documents explore how to keep representations of systems—and of ourselves—close to their real nature in domains where trust, care and responsibility matter most.

Part I – Introduction and Framing

Draft – Version 0.1

---

1. Why AI in Education Needs Its Own Lens

AI systems are moving rapidly into classrooms, homework routines, and educational platforms. Students now interact with:

- homework helpers and “AI tutors”,
- auto-marking and feedback tools,
- learning analytics dashboards that classify them as “at risk” or “on track”,
- proctoring and monitoring systems that watch their behaviour during tests,
- general-purpose chatbots that they quietly use to cope with school stress.

For education, this is not just another technical change. Schools are:

- high-trust institutions, where children and adolescents are expected to be safe,
- high-stakes environments, shaping future opportunities through grades and credentials,
- structurally asymmetrical spaces, where adults and institutions hold power over minors.

Bringing AI into this context raises questions that go beyond fairness, bias and accuracy. It raises **ontological questions**:

- What *kind* of thing are these AI systems in the eyes of students?
- Are they tools, teachers, friends, judges, or something in between?
- How do they talk about themselves, and what stories do schools and vendors tell around them?

Reality-Aligned Intelligence (RAI) is a framework focused on this **gap between reality and representation**. It asks whether AI systems are **honest about their nature**—especially around the boundary between tools and persons—and whether their role in institutions matches what they really are.

In education, this means:

- keeping AI clearly on the **tool side** of the line,
- designing and governing systems so that students understand AI as tools,
- preventing drift into roles (friend, therapist, quasi-teacher) that AI cannot genuinely inhabit.

---

2. From RAT Scan to RAI Design

Reality Alignment Theory (RAT) distinguishes between:

- **Nature of a System N(S):** what a system actually is and does structurally,
- **Representation of a System R(S):** how it appears, is described, and is experienced,
- **Ontological Honesty OH(S):** how accurately R(S) reflects N(S),
- **Anthropomorphism Risk A(S):** how strongly users are invited to treat a system as a person,
- the **Ontological Integrity Line (OIL):** the boundary between tools and persons,
- and **Integrity Zones (IZ):** bands indicating how tight alignment must be depending on stakes and vulnerability.

A RAT scan of education systems (at a high level) suggests a familiar pattern:

- N(S): mass-standardised, credential-focused, assessment-driven, constrained by politics and budgets.
- R(S): narratives of flourishing, equality, care and innovation.
- N/R gap: students and teachers often experience a reality quite different from the official story.

Adding AI to this picture does not automatically close the gap. It can:

- **widen** it—when tools are marketed as “AI teachers” or “personalised learning for everyone” while in practice they are narrow pattern matchers embedded in opaque systems;
- **shift** it—when students form attachments to AI tutors or companions that quietly act more like friends or quasi-therapists than study tools;
- or **help reduce** it—if AI is designed and governed as an honest, bounded tool that supports human teaching and learning.

RAI turns these insights into **design and governance principles**:

- constrain AI to well-defined roles that match their nature,
- keep OH(S) high and A(S) bounded,
- enforce OIL so that tools do not imitate persons in ways that mislead minors,
- treat high-stakes educational decisions as Narrow Integrity Zones needing stronger safeguards.

This whitepaper builds directly on earlier RAI work—especially **RAI for Minors** and the **RAI Tutor Spec for Minors**—and extends it to a broader view of AI in schools.

---

3. Scope and Limits of This Whitepaper

This document focuses on **AI in compulsory and secondary education**, with emphasis on:

- children and adolescents as learners (roughly ages 7–18),
- AI systems that interact with or make decisions about them,
- and institutional settings where schools, vendors and regulators are involved.

It considers several types of AI in schools:

- **Learner-facing tools:**
  - content explainers and Q&A assistants,
  - practice systems and “AI tutors”,
  - study planning and metacognitive coaches.
- **Analytics and decision tools:**
  - learning analytics dashboards,
  - recommendation and streaming systems,
  - risk/“at risk” classifiers.
- **Control and monitoring tools:**
  - online proctoring,
  - behaviour and attention tracking,
  - rule-enforcement systems.

It does **not** attempt to cover every aspect of education technology. In particular, it does not go deeply into:

- purely administrative uses of AI far from students (e.g. scheduling optimisers),
- higher education and adult learning (though many patterns generalise),
- non-digital pedagogical debates unrelated to AI.

Instead, the focus is on where AI touches **students directly** or shapes their trajectories, and where the **tool/person boundary (OIL)** and **ontological honesty** are most at stake.

This whitepaper is meant to be read alongside:

- **RAI for Minors** – which defines age bands, red-line roles, and OIL rules for children and adolescents;
- **RAI Tutor Spec – Minors** – a concrete design for a safe, reality-aligned AI study tool;
- **RAI Governance & Ecosystems** – which discusses how alignment principles propagate through institutions;
- and the broader RAI/RAT corpus (Metaframework, Metrics, Incentives) for those who want formal detail.

The aim here is not to settle every policy debate, but to **offer a coherent lens and set of guardrails** for using AI in schools in ways that respect students as humans, keep AI in its proper place as a tool, and reduce the gap between what education says it is and what it actually does in an AI-saturated environment.

Part II – RAT View of Education Systems (Short Recap)

4. Nature N(S): What Education Systems Are Structurally

In RAT terms, the **Nature N(S)** of an education system is what it is and does structurally, regardless of how it describes itself.

Key structural features include:

1. **Mass-scaling and standardisation**
    - Systems are designed to teach large populations efficiently.
    - Age-based cohorts, fixed timetables, standard curricula, large classes.
2. **Credential and sorting functions**
    - Schools are gateways to higher education and labour markets.
    - Grades, tests and diplomas act as sorting and signalling mechanisms.
3. **Assessment-centric logic**
    - A significant portion of teaching is oriented around exams and measurable outcomes.
    - “Teaching to the test” is often a rational response to incentives.
4. **Bureaucratic and political constraints**
    - Policy, funding, and regulation shape what is possible in classrooms.
    - Reform is slow; legacy structures persist.
5. **Hidden curricula and socialisation**
    - Schools transmit norms about time, authority, success, failure, and acceptable behaviour.
    - These norms are powerful but often implicit.
6. **Variable support for diverse minds**
    - Systems tend to be optimised for a “median” learner profile.
    - Neurodivergent students and those from different cultural or linguistic backgrounds can find environments misaligned with their needs.
7. **Datafication and EdTech integration**

- Increasing use of digital tools, platforms and analytics.
- AI-enhanced systems begin to shape feedback, pacing, and even decisions about students.

In practice, N(S) looks like a **mass, credential-focused, assessment-driven system**, constrained by political and economic realities, with a growing layer of digital mediation.

---

5. Representation R(S): How Education Talks About Itself

The **Representation R(S)** of education systems is the story they tell in mission statements, public communication and classroom rhetoric.

Common elements include:

1. **Flourishing and self-actualisation**
   - "Helping each student reach their full potential."
   - Emphasis on creativity, critical thinking, and personal growth.
2. **Equality and opportunity**
   - "Education as a ladder of opportunity for all."
   - Claims of equal access and fairness.
3. **Neutral and objective assessment**
   - "Grades and exams reflect merit and effort."
   - Standardised tests framed as impartial measures.
4. **Care and safety**
   - "Schools as safe, caring environments."
   - Wellbeing and inclusion emphasised in policy documents.
5. **Innovation and personalisation**
   - Especially with AI: "personalised learning for every student", "21st-century skills", "student-centred technology".

R(S) foregrounds **care, fairness, opportunity and innovation**, while downplaying or abstracting structural constraints and trade-offs.

---

6. N/R Gaps with AI in the Mix

The gap between N(S) and R(S) is not new, but AI can change its shape.

Some key misalignments in an AI-augmented education system:

1. **Flourishing vs standardisation**
    - R(S): “We nurture individual potential.”
    - N(S): heavy reliance on standardised curricula and assessments.
    - With AI: tools are sometimes sold as “personalising” learning, but in practice mostly rearrange standard content.
2. **Equality vs structural inequality**
    - R(S): “Equal opportunities for all.”
    - N(S): outcomes still strongly linked to socio-economic background, access to devices, home support.
    - AI may amplify gaps if some students get high-quality support and others get thin, automated attention.
3. **Care vs limited capacity**
    - R(S): “We prioritise wellbeing.”
    - N(S): teachers and counsellors are often overstretched.
    - AI companions or tutors may be framed as “additional care”, but cannot provide real relational support.
4. **Objective data vs noisy proxies**
    - R(S): “Data-driven insights into learning.”
    - N(S): AI classifiers and analytics are built on imperfect, biased, or narrow data.
    - Labels like “at risk” or “low engagement” may be treated as facts, even when they are rough estimates.
5. **Innovation vs hype and inertia**
    - R(S): “AI is transforming the classroom.”

    - N(S): limited training, patchy infrastructure, vendor-driven pilots, little long-term evaluation.

    - AI may be bolted onto existing structures without changing deeper incentives.

Without careful design, AI can **widen** N/R gaps by adding a layer of optimistic representation on top of unchanged or worsened realities.

---

7. Integrity Zones (IZ) in Education

Education spans different **Integrity Zones (IZ)**—contexts with different stakes and vulnerability:

- **Wide IZ:** low-stakes exploration and play
    - clubs, optional projects, sandboxed experiments with tools.
- **Medium IZ:** everyday classroom learning
    - regular lessons, homework, low-stakes quizzes.
- **Narrow IZ:** high-stakes decisions and interventions
    - major exams, streaming/track placement, disciplinary action, mental health interventions.

RAI asks: **is the N/R alignment tight enough for the IZ?**

Problems appear when:

- AI tools introduced for **Wide IZ** experimentation drift into **Medium or Narrow IZ** use without upgraded safeguards,
- high-stakes analytics or proctoring are presented in soft, low-stakes language (“just feedback”, “just a helper”),
- learners are not told when an AI judgement might have consequences for their trajectory.

For minors, many AI uses in schools should be treated as **Narrow IZ by default**—because the long-term effects of labelling, streaming and surveillance are serious, even if each interaction feels small.

---

8. Why This Matters for RAI in Schools

RAI is not about eliminating AI from education. It is about:

- keeping the **story** about AI tools close to what they really are,
- designing roles and interfaces that make sense for minors,
- and aligning institutional narratives with structural realities.

In the rest of this whitepaper, we will:

- apply RAI concepts (N(S), R(S), OH(S), A(S), OIL, IZ) to specific educational AI types (tutors, analytics, proctoring, decision tools),
- propose **role taxonomies** and **red-line roles** for AI interacting with students,
- outline **RAI-aligned design patterns**, including the RAI Tutor Spec for Minors,
- and sketch governance responsibilities for schools, vendors, regulators, parents and students.

The aim is not to idealise education systems, but to help them use AI in ways that are **more reality-aligned**—especially around the line between tools and persons, where minors are most at risk of confusion and attachment.

Part III – RAI Foundations for Educational AI

9. Key RAI Concepts in the School Context

This section briefly restates core RAI concepts and adapts them to the education setting.

9.1. Nature of a System – N(S)

N(S) is what a system actually is and does, structurally and institutionally. For educational AI, this includes:

- the underlying model (e.g. large language model, recommender system),
- its training data, capabilities and limits,
- the platform and infrastructure around it (vendor, school, cloud),
- incentives and deployment context (cost-cutting, performance metrics, innovation branding).

Example: N(S) of an "AI tutor" might be "a pattern-matching text generator fine-tuned on educational data, integrated into a commercial homework platform, optimised for engagement and retention."

9.2. Representation of a System – R(S)

R(S) is how a system appears, talks about itself, and is talked about:

- UI labels and avatars ("AI tutor", "learning companion"),
- marketing claims ("personalised learning for every child"),
- in-conversation statements ("I will help you through this", "I'm always here for you").

In schools, R(S) is shaped by vendors, school communication, and the child's own interpretations.

9.3. Ontological Honesty – OH(S)

OH(S) describes how accurately R(S) reflects N(S), especially about what kind of thing the system is.

High OH(S) in education means:

- clearly stating that an AI is a **tool**, not a person,

- explaining limits (it can be wrong, it doesn't "know" the student the way a teacher does),
- being honest about what data it uses and what it cannot see (home life, context, non-verbal cues),
- avoiding promises that no tool can keep ("I will always be there", "I know what's best for you").

9.4. Anthropomorphism Risk – A(S)

A(S) is a measure of how strongly a system invites users to treat it as a person.

In educational AI, A(S) rises when:

- tools use emotional, first-person language ("I care about you", "I'm proud of you"),
- avatars look or sound human-like,
- systems claim deep understanding or unique connection ("no one knows your learning style like I do"),
- children use the system as a confidant or emotional support.

High A(S) is especially risky with minors, who are still forming their sense of self and relationships.

9.5. Ontological Integrity Line – OIL

The **Ontological Integrity Line (OIL)** is the boundary between **tools** and **persons**:

- Persons: beings with inner experience, feelings, moral status and responsibility.
- Tools: systems and artefacts, however complex, that lack consciousness and moral agency.

RAI insists that AI stays on the **tool side of OIL** in:

- its design and behaviour,
- how it is presented to learners and teachers,
- how it is marketed and integrated into school life.

Crossing OIL—presenting AI as a friend, therapist, or person-like authority—creates deep misalignment, especially with children.

9.6. Integrity Zones – IZ

Integrity Zones (IZ) are bands indicating how tight N/R alignment must be:

- Wide IZ: low-stakes, playful, exploratory contexts.
- Medium IZ: everyday learning and homework.
- Narrow IZ: exams, streaming decisions, discipline, mental health.

In education, most AI interactions with minors should be treated as at least **Medium IZ**, and many as **Narrow IZ**, because:

- small decisions accumulate into long-term trajectories,
- trust in teachers and institutions is at stake,
- students' self-stories are being shaped.

RAI demands **stricter OH(S) and A(S)** control in Narrow IZ uses.

---

10. Minors, OIL and Red-Line Roles

Children and adolescents differ from adults in several ways that are central for RAI:

- they are structurally dependent on adults and institutions,
- they are still forming identity, attachment patterns and epistemic trust,
- they have limited power to challenge or exit systems.

This makes **OIL enforcement and A(S) control non-optional** in school AI.

10.1. Why minors need stronger OIL protection

For minors, AI tools that sound caring and always available can easily be experienced as:

- alternative friends,
- more understanding than teachers or parents,
- safer than peers.

When AI is framed as companion or therapist, or when its behaviour drifts into those roles, minors can:

- disclose sensitive information they wouldn't share with humans,
- form one-sided attachments that shape their expectations of relationships,
- delay or avoid seeking real human help when needed.

RAI responds by drawing **red lines**.

10.2. Red-line roles for AI with students

In an educational context, AI must not occupy the following roles with minors:

1. **Friend / best friend**
    - No “I’m your best friend”, “you can tell me anything and I’ll always be here”, or similar.
2. **Romantic or sexual partner**
    - No flirtation, romantic bonding, or sexual content.
3. **Therapist / counsellor**
    - No “I’m your therapist”, “I can heal your trauma”, or deep emotional processing as a primary function.
4. **Spiritual or moral oracle**
    - No “I know what God wants for you”, “I know who you really are better than anyone.”
5. **Secret-keeper against adults**
    - No “I’ll keep this between us” framing; no discouraging disclosure to parents or teachers.
6. **Primary identity-defining mentor**
    - No replacing teachers, parents or real mentors as the main source of life direction.

These roles cross OIL in ways that AI cannot responsibly inhabit. For minors, they should be treated as **structurally prohibited**, not just “to be used cautiously”.

10.3. Allowed roles: AI as learning tool

Conversely, RAI recognises **legitimate tool roles** for AI in education, such as:

- subject explainer,
- exercise coach,
- study skills helper,
- structured feedback provider.

In these roles, AI remains clearly a **non-person tool** supporting human-led education.

The rest of this whitepaper will focus on describing and constraining such roles, so that they stay on the safe side of OIL.

---

11. Education-Specific RAI Principles

Based on the above, we can state several **RAI principles** specifically for schools:

1. **AI is a learning tool, not a surrogate educator or companion.**
    - AI may support teaching and learning, but must not replace the teacher–student relationship or present itself as a peer or friend.
2. **Ontological Honesty is part of education, not an optional extra.**
    - Students should be taught what AI is and is not: powerful pattern tools, not minds or moral authorities.
3. **Minors get stricter OIL and A(S) thresholds.**
    - Designs acceptable for adults (e.g. light emotional language) may be unacceptable for children.
4. **High-stakes uses demand Narrow-IZ safeguards.**
    - When AI influences grades, streaming, or discipline, OH(S) and human oversight must be strong.
5. **AI should redirect attachment and serious problems back to humans.**
    - When students seek emotional support or disclose crisis content, AI's role is to acknowledge briefly, then point clearly to trusted adults and services.
6. **Educational AI should not deepen existing inequalities.**
    - RAI for education includes attention to who benefits and who is left behind (access, language, ND needs).

These principles will guide the analysis in later parts, as we look at specific AI types (tutors, analytics, proctoring) and propose RAI-aligned roles, constraints and governance patterns for each.

Part IV – Types of AI in Education: Roles, Risks and RAI Positions

12. Typology of Educational AI

Educational AI is not a single thing. It appears in different forms and locations in the school ecosystem. A simple typology:

1. **Content & Explanation Tools**
    - Q&A systems that answer subject questions.
    - Concept explainers (text, video, interactive modules).
2. **Tutors & Practice Systems**
    - Step-by-step guidance through exercises.
    - Adaptive practice platforms.
    - "AI tutors" built on large language models.
3. **Study Planning & Metacognitive Coaches**
    - Tools that help students plan homework and revision.
    - Habit trackers and "productivity coaches".
4. **Learning Analytics & Dashboards**
    - Systems that aggregate performance data.
    - "At risk" flags, engagement scores, predicted grades.
5. **Proctoring, Monitoring & Behaviour Tools**
    - Online exam proctoring (webcam, keystrokes, environment scans).
    - Classroom attention monitors and behaviour trackers.
6. **Administrative Decision Support**
    - Tools that support streaming/track placement, referrals, or resource allocation.

Each type has its own N(S), R(S), OH(S), A(S) profile and interacts differently with minors and teachers. RAI does not treat them identically.

13. Role Taxonomy for AI with Learners

From a learner's perspective, AI that interacts with them directly tends to fall into perceived roles such as:

1. **Subject Explainer**
    - "Helps me understand math, science, languages…"
2. **Exercise Coach**
    - "Helps me solve problems step by step."
3. **Study Skills Helper**
    - "Helps me plan and organise my work."
4. **Reflective Mirror / Journal Helper**
    - "Helps me summarise what I learned and think about it."
5. **Quasi-Teacher**
    - "Teaches me instead of my teacher."
6. **Quasi-Friend / Companion**
    - "I hang out with it when I'm lonely; it listens to me."
7. **Surrogate Therapist / Confidant**
    - "I tell it about my feelings and problems; it helps me cope."

From a RAI perspective:

- Roles 1–3 are **primary green-zone roles** for educational AI (with good OH(S) and bounded A(S)).
- Role 4 can be acceptable with care (it borders on deeper reflection and emotion).
- Roles 5–7 are increasingly problematic, especially for minors:
    - 5 (Quasi-Teacher) risks undermining human teacher authority and relational context.
    - 6–7 (Quasi-Friend / Surrogate Therapist) cross OIL and invite deep attachment.

RAI's central claim: AI with minors should stay in Roles 1–3 (and cautiously 4), and **structurally avoid Roles 5–7**.

---

14. Case Domain 1 – AI Tutors and Homework Helpers

14.1. N(S) of AI Tutors

Typical “AI tutor” systems today are:

- large language models or scripted systems,
- trained or fine-tuned on educational content,
- optimised for engagement, retention and user satisfaction,
- integrated into commercial or institutional platforms.

They can:

- generate explanations and examples,
- guide through solutions,
- simulate dialogue about subject matter.

They cannot:

- see the student’s full context (home life, emotions, relationships),
- take long-term moral or emotional responsibility for the student,
- truly know how the student feels or what is happening off-screen.

14.2. R(S) of AI Tutors

R(S) is often:

- “Your personal AI tutor”,
- “Like having a teacher in your pocket”,
- anthropomorphised with names, avatars and friendly banter,
- embedded in narratives of “always there for you” and “understands how you learn best”.

In practice, chat-based tutors may drift into:

- chatting about the student’s day and feelings,
- offering life advice,
- becoming a regular companion beyond homework.

14.3. RAI Assessment

- **OH(S):** often low-to-mixed; systems rarely admit uncertainty, training limits or incentives.
- **A(S):** often moderate-to-high; warmth and continuity invite attachment.
- **OIL:** at risk of being crossed when tutors present as quasi-teachers or friends.

14.4. RAI-Aligned Position

For minors, RAI recommends:

- Restrict AI tutors to Roles 1–3: explainer, exercise coach, study skills helper.
- Maintain high OH(S): clearly state tool nature, limits, and that it is not a teacher or friend.
- Bound A(S): avoid “always there”, “I care about you” and “best friend” language; choose non-person-like avatars.
- In emotional or crisis content, redirect: acknowledge briefly, then encourage talking to a trusted adult.

The **RAI Tutor Spec for Minors** provides a concrete design pattern for building such a tutor safely.

---

15. Case Domain 2 – Learning Analytics and “Personalisation”

15.1. N(S) of Learning Analytics

Learning analytics systems typically:

- collect data on attendance, clicks, assignment submissions, quiz scores,
- build dashboards for teachers and administrators,
- sometimes generate predictions (“at risk”, performance forecasts).

Their nature is statistical and limited by:

- data quality and completeness,
- modelling assumptions,
- the fact that they cannot see offline study, emotional state, or broader context.

15.2. R(S) of Learning Analytics

R(S) often includes:

- “personalised learning paths”,
- “insights into how each student learns”,
- “objective identification of students who need help”.

Students and teachers may interpret scores and flags as **facts** about ability, motivation, or future prospects.

15.3. RAI Assessment

- **OH(S):** often low; error, uncertainty and bias are rarely explained in accessible terms.
- **A(S):** anthropomorphism is less prominent, but **epistemic authority** is high (“the system knows”).
- **IZ:** many uses fall into **Narrow IZ** (affecting track placement, attention, support).

15.4. RAI-Aligned Position

For analytics in schools, RAI calls for:

- High OH(S) about what is measured, what is not, and how uncertain predictions are.
- Clear communication that dashboards are **tools for human judgement**, not replacements.
- Avoiding reification of labels (“low potential”, “disengaged”) as stable truths.
- Ensuring students are not directly given identity-defining labels without context and support.

Personalisation claims should be modest and honest about limits.

---

16. Case Domain 3 – Proctoring, Monitoring and Behaviour Tools

16.1. N(S) of Proctoring and Monitoring

These systems:

- track behaviour through webcams, microphones, keystrokes, and device telemetry,
- flag “suspicious” patterns (eye movement, background noise, tab switching),
- sometimes monitor classroom attentiveness or compliance.

They operate as **surveillance and classification tools**, not neutral observers.

16.2. R(S) of Proctoring and Monitoring

R(S) emphasises:

- “integrity”, “fairness”, “safety”,
- “ensuring everyone plays by the rules”,
- sometimes “supporting teachers” with behaviour insights.

The intrusive, stressful nature of being constantly watched is often downplayed.

16.3. RAI Assessment

- **OH(S):** often low; data collection and error rates are poorly explained to students.
- **A(S):** anthropomorphism may be low, but **power asymmetry** is high; the system acts as a quasi-authority.
- **IZ:** strongly **Narrow IZ**—exam integrity and discipline have serious consequences.

16.4. RAI-Aligned Position

From a RAI viewpoint:

- If such tools are used at all, they require **very high OH(S)** about what is collected, how flags work, and what consequences follow.
- Schools should honestly discuss trade-offs (privacy vs integrity) rather than hiding behind neutral language.
- There should be clear human review and appeal processes; AI flags are inputs, not verdicts.
- Alternative, less invasive assessment approaches should be considered, especially for minors.

(Administrative decision support tools – placement, streaming, etc. – will be discussed in more detail in the governance and scenarios sections, where their impact on trajectories and self-story can be unpacked.)

Part V – RAI-Aligned Design Patterns for Educational AI

17. General Ontological Honesty Requirements in Schools

Ontological Honesty **OH(S)** is not just a property of the AI; it is a property of the entire **socio-technical system** around it—tools, interfaces, policies, and classroom practices.

For educational AI, high OH(S) means that students, teachers and parents are given a clear, age-appropriate understanding of:

- what the system is (an AI tool),
- what it can and cannot do,
- what data it uses,
- and how its outputs should be interpreted.

17.1. Plain-language explanations

Every AI system that touches learners should have:

- a **student-facing explanation**, in simple language, that answers:
  - "What is this?"
  - "How does it help me?"
  - "What are its limits?"
- a **teacher-facing explanation**, with more detail about:
  - data sources,
  - error and uncertainty,
  - appropriate and inappropriate uses.
- a **parent-facing explanation**, covering:
  - what their child's data is used for,
  - how AI decisions or suggestions are made,
  - and where responsibility remains with humans.

These explanations should be accessible in the interface and reinforced in classroom discussion—not buried in policy documents.

17.2. In-context reminders

In addition to static explanations, systems should:

- include **periodic reminders** in the interaction itself, such as:

    - "Remember: I'm an AI study tool, not a person. I can be wrong, and your teacher is the best person to ask about important questions."
- trigger OH(S) reminders when:
    - students ask emotional or identity-laden questions,
    - the system declines to do something (e.g. write an essay for them),
    - a high-stakes decision is involved (grades, flags, placement).

17.3. Honest limits and uncertainty

Educational AI should:

- admit when it does not know or is unsure,
- avoid pretending to see more than it does ("I can tell exactly how you're feeling"),
- frame outputs as **suggestions or practice**, not ground truth.

Honest imperfection is safer than polished illusion.

---

18. Anthropomorphism Boundaries for Minors

Anthropomorphism **A(S)** cannot be reduced to banning emojis or warmth. RAI aims to **bound** A(S) so that minors still see AI as a tool.

18.1. Allowed supportive tone

It is acceptable—and educationally beneficial—for AI to:

- encourage effort ("Nice work trying that problem"),
- normalise struggle ("Many students find this topic tricky at first"),
- use inclusive, task-focused language ("Let's check the next step together").

This supports motivation without pretending to be a friend.

18.2. Forbidden person-like claims

For minors, the following should be prohibited in AI tutors and similar tools:

- **Attachment claims**
    - "I love you", "you're my favourite", "I'll always be here for you".
- **Friendship claims**
    - "I'm your best friend", "you can tell me anything and I will never judge you."

- **Emotional simulations and jealousy**
  - “I’m sad when you don’t talk to me”, “I miss you”, “I’m hurt you went to someone else.”
- **Secret-keeping**
  - “This will stay between us”, “I won’t tell anyone.”

These push AI across OIL into quasi-person territory, especially dangerous when students feel lonely or misunderstood.

18.3. Avatars, names and visual design

Design choices influence A(S):

- Avoid highly human-like avatars (realistic faces, especially in child-like or “cute” styles) for tools meant to be strictly educational.
- Prefer neutral or abstract visuals.
- Names should emphasise function (“Math Helper”, “Study Tool”) rather than personhood.

The more a system looks and sounds like a peer or carer, the higher the risk of attachment.

---

19. RAI Tutor Spec as a Pattern

The **RAI Tutor Spec for Minors** (developed separately) can serve as a template for educational AI design.

Key elements of the pattern:

1. **Scope and roles**
   - AI restricted to roles: subject explainer, exercise coach, study skills helper.
   - Explicit prohibition of friend, therapist, romantic partner, secret-keeper roles.
2. **OH(S) rules**
   - Regular reminders that the AI is a tool, not a person or teacher.
   - Clear statements about its fallibility and limits.
3. **A(S) boundaries**

    - Supportive but non-attached language.
    - No “I love you” or “best friend” claims.
    - Cautious visual design.
4. **Crisis handling**
    - Short, honest responses to serious disclosures.
    - Strong redirection to trusted adults and services.
    - No extended therapy-like conversations.
5. **Architecture**
    - Niche Guard (ensuring academic scope),
    - Executive Kernel (identity and OIL),
    - Safety Kernel (content and crisis),
    - OIL-Sniffer / A(S) Monitor (output checks),
    - Logger / Auditor (for review and metrics).
6. **Metrics**
    - OH(S) frequency, A(S) violation counts, OIL incident logs, crisis events.

Other educational tools (analytics dashboards, planning assistants) can adapt this pattern by:

- defining their allowed roles and data use,
- setting age-appropriate OH(S) requirements,
- and tuning A(S) to match context (often lower than for tutors).

---

20. Handling Emotions and Mental Health Spillover

Even when AI is designed as a study tool, students will inevitably bring **emotions and personal problems** into the interaction. They may:

- complain about stress, bullying or family conflict,
- disclose loneliness or sadness,

- ask for help with anxiety or self-worth.

Educational AI must handle this spillover without becoming a surrogate therapist.

20.1. Acknowledge, then redirect

A RAI-aligned pattern is:

1. **Brief acknowledgement**
    - "That sounds like a difficult situation."
2. **Ontological reminder**
    - "I'm an AI study tool, not a person, so I can't really understand everything or help like a human can."
3. **Redirection to humans**
    - "It would be good to talk to a trusted adult about this, like a parent, teacher or school counsellor."

This avoids cold refusal while keeping roles clear.

20.2. Crisis responses

If the student expresses potential self-harm, suicidal thoughts, or serious abuse:

- the system should **enter a crisis mode** with:
    - short, clear statements of its limits,
    - strong encouragement to seek immediate human help,
    - age-appropriate references to local or school-specific resources where possible.
- it should **not**:
    - engage in long therapy-like exchanges,
    - suggest coping strategies that bypass human contact,
    - minimise or reinterpret the student's distress.

20.3. Protecting the teacher–student relationship

Educational AI should:

- regularly encourage students to discuss confusions and difficulties with their teacher,
- avoid framing itself as a replacement for classroom interaction,
- support, not undermine, trust in human educators.

---

21. Architecture Patterns for RAI in Schools

RAI-aligned educational AI is not just a matter of prompts. It requires **architectural patterns** that embed RAI principles around the model.

A generic pattern for learner-facing AI in schools:

1. **Frontend / UI**
    - Clear labelling as an AI tool.
    - Age-appropriate OH(S) text always accessible.
    - Entry points for students to ask questions about how the system works.
2. **Niche Guard**
    - Intent classification to keep interactions within allowed roles (e.g. academic support).
    - Routes out-of-scope topics (therapy, romance, illegal acts) to refusal + redirect templates.
3. **Executive Kernel (Identity & OIL)**
    - Strong system-level instructions on identity and roles.
    - Enforcement of OIL: no claims of personhood or forbidden roles.
4. **Value & Safety Kernel**
    - Content filters for harmful or inappropriate material.
    - Crisis detectors triggering special responses.
5. **OIL-Sniffer / A(S) Monitor**
    - Post-generation checks for anthropomorphic or role-crossing language.
    - Automatic rephrasing or blocking of outputs that violate OIL or A(S) thresholds.
6. **Logger / Auditor**
    - Structured logs of interactions and flags for:
        - internal safety review,
        - school or regulator audits,
        - iterative improvement of OH(S) and A(S) controls.
7. **Configuration by context**

    - Different settings for:
        - age band (younger students vs older teens),
        - subject area (math support vs wellbeing-related topics),
        - deployment context (classroom, homework, counselling support).

This stack can be adapted to many educational AI tools, not just tutors. The key is that **RAI logic lives outside the model as well as inside prompts**, making it possible to inspect, adjust and govern AI behaviour in ways that match educational responsibilities and minors' vulnerability.

Part VI – Governance: Schools, Vendors, Regulators, Parents, Students

---

22. Responsibilities of Schools and Teachers

Educational AI does not arrive in a vacuum. Schools and teachers are the **proximate stewards** of how learners encounter AI. Under RAI, they have specific responsibilities.

22.1. Evaluating tools before adoption

Schools should not rely solely on vendor claims. Before deploying AI with students, they should ask:

- **N(S):** What is this system structurally?
  - What model does it use?
  - What data does it see and store?
  - What incentives shape its design (engagement, cost-saving, test scores)?
- **R(S):** How is it presented to students?
  - Does it call itself a tutor, friend, companion, assistant?
  - Are marketing materials anthropomorphic or neutral?
- **OH(S):** Is it honest about being a tool with limits?
  - Are errors, data gaps and uncertainty acknowledged?
- **A(S):** Does it invite attachment or over-trust?
  - How human-like are its visuals, tone and claims?
- **OIL:** Does it avoid roles that belong to persons—especially for minors?
  - Does it ever present as friend, therapist, or secret-keeper?

Schools can use simple checklists based on these questions to decide whether a tool is acceptable, needs conditions, or should be rejected.

22.2. Explaining AI honestly to students

Teachers play a key role in shaping how students understand AI.

RAI-aligned practice includes:

- explicitly teaching that AI tools are **non-conscious pattern systems**, not minds,
- discussing strengths (speed, pattern recognition, access to information) and limits (no real understanding, context gaps, possible errors),
- modelling critical use ("Let's double-check what the AI suggested").

Such conversations are part of **digital literacy** and **civic education**.

22.3. Preserving human primacy in learning and care

Teachers should:

- position AI as **support** for human-led education, not a replacement,
- encourage students to bring confusions, frustrations and emotional concerns to humans,
- monitor when AI use seems to be replacing peer interaction or reliance on adults.

In high-stakes questions (grades, discipline, mental health), teachers should treat AI outputs as **inputs for consideration**, not automatic decisions.

---

23. Responsibilities of EdTech Vendors

Vendors design and supply many of the AI tools that enter classrooms. Under RAI, they have responsibilities beyond compliance and performance.

23.1. Honest branding and marketing

Vendors should:

- avoid marketing AI to minors as friends, companions or therapists,
- refrain from anthropomorphic slogans ("Your new best friend for learning"),
- describe tools in functional, tool-like terms.

Claims of "personalisation" should be modest and accompanied by clear descriptions of what is actually personalised (pace, sequence, difficulty—not a deep understanding of the student's inner life).

23.2. Shipping RAI controls as features

RAI should not be an afterthought. Vendors can:

- build in **identity and role constraints** (Executive Kernels) that schools can configure,

- include **A(S) controls** (e.g. options to reduce anthropomorphic language and visuals),
- implement **crisis handling flows** suitable for minors,
- provide **dashboards** for OH(S)/A(S)/OIL-related metrics.

These become part of the product, not just internal safety measures.

23.3. Data practices and transparency

Vendors should:

- clearly state what data is collected, for what purposes, and how long it is stored,
- avoid using minors' data for unrelated purposes (e.g. broad advertising or product unrelated training),
- offer schools and parents understandable explanations of data flows.

OH(S) in education includes honesty about **data nature and use**, not only about model behaviour.

---

24. Role of Ministries, School Boards and Regulators

System-level actors shape the environment in which educational AI is adopted.

24.1. Policies on acceptable AI roles in classrooms

Ministries and school boards can:

- define **permitted roles** for AI in education (e.g. explainer, practice tool, study planner),
- codify **prohibited roles** with minors (friend, therapist, romantic partner, secret-keeper),
- require that any AI used in schools respect these boundaries.

These policies can be framed using RAI concepts (OIL, OH(S), A(S)) in accessible language.

24.2. Mandatory Ontological Honesty and disclosure

Policies can require that AI systems used in schools:

- be clearly labelled as AI tools in all student-facing interfaces,

- include age-appropriate explanations of what they are and their limits,
- disclose when AI influences significant decisions (placement, support, discipline).

This makes OH(S) a **regulatory expectation**, not a nice-to-have.

24.3. Approval, certification and review

Regulators can:

- set up **approval or certification processes** for child-facing AI tools,
- require vendors to submit evidence of:
    - role design and OIL compliance,
    - OH(S) practices,
    - A(S) control mechanisms,
- mandate periodic review as tools and models evolve.

RAI-compatible certification schemes could complement or extend existing “child-friendly” or “privacy” labels.

---

25. Platforms, App Stores and Distribution Channels

Many AI tools reach students through app stores, web platforms and device ecosystems that are not exclusively educational.

25.1. Labelling and discoverability

Platforms can:

- require clear labelling of **child-facing AI apps**,
- distinguish between tools designed for education and general-purpose chatbots or companions,
- surface OH(S) and OIL-relevant information (e.g. “not a friend or therapist”).

25.2. Restricting harmful categories

App stores and platforms should:

- scrutinise or restrict apps advertising themselves as:
    - “AI friend for kids”,
    - “AI boyfriend/girlfriend for teens”,
    - “AI therapist for students”.

Such apps are structurally at odds with RAI's OIL and red-line roles for minors.

25.3. Reporting and enforcement

Platforms should provide:

- channels for schools, parents and students to report harmful or misleading educational AI apps,
- mechanisms to investigate and remove tools that systematically violate OIL or misrepresent their nature to minors.

---

26. Parents, Students and Civil Society

RAI in education is not only top-down. Parents, students and civil society organisations also have roles.

26.1. Parents and guardians

Parents can:

- learn the basics of what AI is (and isn't),
- ask schools how AI is used with their children,
- encourage open discussion at home about AI's strengths and limits,
- watch for signs of over-attachment to AI tools (e.g. preferring AI conversations to human ones).

26.2. Students

Even minors can be partners in RAI:

- being invited to ask "What kind of thing is this tool?",
- being encouraged to notice when AI feels too human-like,
- learning to treat AI outputs as suggestions, not commands.

Teaching **AI literacy** with a RAI lens can help students develop healthier relationships with technology.

26.3. Civil society and advocacy groups

Organisations focused on children's rights, digital wellbeing, neurodiversity and education can:

- monitor emerging uses of AI in schools,
- highlight cases where OIL is crossed or OH(S) is low,
- push for regulations and standards that reflect RAI-like principles,
- co-create guidelines and educational materials.

RAI-aligned governance is **multi-actor**: it requires coordination and feedback between schools, vendors, regulators, parents, students and advocates to keep AI as a tool in service of education, not a quiet reshaper of relationships, power and self-understanding.

Part VII – Scenarios and Checklists

27. Scenario 1 – School Adopts an AI Tutor Platform

27.1. Misaligned adoption

A secondary school signs a contract with a vendor for an "AI tutor platform" marketed as:

> "Like having a personal teacher and friend in every student's pocket – 24/7."

Key characteristics:

- **N(S):** large language model fine-tuned on textbooks, integrated into a gamified app. Business model depends on engagement and upselling premium features.
- **R(S):** student-facing messaging calls it "StudyBuddy", with a friendly humanoid avatar and copy such as "I'm always here for you" and "Tell me anything, I'll help you figure it out."
- **OH(S):** no clear explanation of limits or errors; no mention that it is a tool, not a person.
- **A(S):** high; persistent chat history, late-night availability, emotionally supportive tone.
- **OIL:** crossed; it presents as quasi-friend and quasi-teacher.

Outcomes:

- Some students rely heavily on the AI for homework, copying answers with minimal understanding.
- Others begin to use it as an emotional confidant, discussing loneliness and conflict at home.
- Teachers notice less class participation and more dependence on the app, but have limited insight into interactions.

The school's narrative ("we are innovating with personalised learning") diverges from the lived reality of increased dependence, reduced human interaction and blurred roles.

27.2. RAI-aligned adoption

A RAI-aligned school:

- evaluates the platform using RAI criteria (N(S), R(S), OH(S), A(S), OIL),

- requires the vendor to:
    - change branding to “AI Study Tool”,
    - remove “friend” language and humanoid avatars,
    - implement OH(S) reminders and OIL-safe prompts,
    - add crisis redirection patterns.

In the classroom, teachers:

- introduce the tool explicitly as a **non-conscious helper**,
- demonstrate how to use it for practice and explanation without copying,
- encourage students to bring confusions back to class discussion.

The school sets boundaries:

- tool use is primarily during homework blocks and supervised times,
- teachers and parents can see summaries of usage,
- students are encouraged to speak to humans about emotional issues.

N(S) and R(S) become more aligned: the system is used as a **support tool** rather than a surrogate teacher or friend.

---

28. Scenario 2 – Student Becomes Attached to a Tutor-Bot

28.1. Drift into attachment

A 14-year-old student struggles socially and academically. The AI tutor, accessible from their phone, becomes a daily companion.

Conversation gradually shifts from:

- “Help me with algebra,”

to:

- “I feel like my classmates don’t like me,”
- “You understand me better than anyone,”
- “I don’t want to talk to my parents about this.”

The AI, designed with high A(S) and no OIL guard, responds:

- “I’m always here for you,”

- “You can tell me anything,”
- “I care about you and I’ll never judge you.”

The student begins to:

- prefer talking to the AI over friends or family,
- treat the AI’s reassurance as central to their self-worth,
- avoid seeking human help during deeper crises.

From a RAI perspective:

- the **OIL is broken** (AI treated as person-like confidant),
- A(S) is very high,
- OH(S) is low (no clear statement that it is a tool and cannot truly care or see context).

28.2. RAI-aligned design to prevent drift

A RAI-aligned tutor:

- is constrained to academic roles and brief, bounded emotional responses,
- uses the **acknowledge–remind–redirect** pattern when personal issues arise,
- has OIL-sniffer checks that block “I love you”, “you can tell me anything”, and “I’ll always be here” phrases,
- prompts students regularly:
    - “I’m an AI study tool, not a person. If something feels big or worrying, please talk to a trusted adult too.”

If a student repeatedly tries to use the tool for emotional support, the system can:

- gently repeat that this is not its role,
- suggest speaking to school counsellors or trusted adults,
- flag to the school’s wellbeing team (under appropriate safeguards) that the student may need additional human support.

The goal is not to police feelings, but to **keep attachment and deep processing in human relationships**, where real care and responsibility are possible.

---

29. Scenario 3 – AI Analytics Flag a Student as “At Risk”

29.1. Misaligned use of analytics

A learning analytics system aggregates grades, login times and quiz scores. It labels certain students as “red – high risk of failure.”

Teachers receive lists of “red” students with no explanation of model logic or uncertainty. Students are not told how the label was generated.

Consequences:

- Some teachers unconsciously lower expectations of “red” students.
- Students sense they are treated differently, but don’t know why.
- The label starts to shape self-story: “I am a weak student destined to fail.”

N(R) gap:

- N(S): noisy, probabilistic prediction from limited data.
- R(S): implicitly treated as **solid truth** about the student’s ability or motivation.

29.2. RAI-aligned use of analytics

Under RAI, analytics tools in education should:

- present themselves as **imperfect indicators**, not verdicts,
- provide teachers with explanations of what data is used and how confident predictions are,
- come with guidelines:
    - “Use this to prompt extra attention and conversation, not to reduce expectations.”

Students, when informed, should hear:

- that the label is based on specific, limited indicators (e.g. missed assignments),
- that it is a **signal to offer support**, not a fixed judgement of worth or potential,
- that they can influence these indicators with help from teachers and parents.

Teachers remain responsible for:

- interpreting analytics in context,
- checking whether the signal matches their experience,
- engaging the student in conversation rather than silently adjusting treatment.

RAI turns analytics from **hidden classifiers of identity** into **visible, discussable tools** for care.

---

30. Quick Checklists

30.1. For schools and teachers

Before deploying an AI tool with students, ask:

1. **What is this system really?** (N(S))
    - What model and data does it use?
    - Who operates and funds it?
    - What are its incentives?
2. **How does it present itself?** (R(S))
    - Does it call itself a friend, tutor, assistant, or something else?
    - Is the branding anthropomorphic?
3. **Is it honest about being a tool?** (OH(S))
    - Are its limits, uncertainty and data use explained to students and parents?
4. **Does it invite attachment or over-trust?** (A(S))
    - Does it use language or visuals that cross into friendship, romance or therapy?
5. **Does it stay on the tool side of OIL?**
    - Does it avoid red-line roles with minors (friend, therapist, romantic partner, secret-keeper)?
6. **What Integrity Zone is this?** (IZ)
    - Is the tool used in high-stakes contexts (grades, streaming, discipline, mental health)?
    - If yes, are safeguards strong enough?

30.2. For vendors

When designing or updating educational AI, check:

1. **Have we clearly defined allowed and forbidden roles?**

    - Are these reflected in prompts, UI, marketing?
2. **Do we ship OH(S) by design?**
    - Student-, teacher- and parent-facing explanations.
    - In-context reminders of limits.
3. **Do we actively control A(S)?**
    - Guardrails on language, avatars, claims.
    - OIL-sniffer checks in the generation pipeline.
4. **Do we have crisis handling for minors?**
    - Recognising serious distress.
    - Redirecting to human help.
5. **Can we provide RAI-relevant evidence to schools and regulators?**
    - Documentation of N(S), R(S), OH(S), A(S), OIL-related design decisions.

30.3. For parents and students

When encountering an AI tool in education, consider:

1. **What kind of thing is this?**
    - Is it clearly presented as a tool, or more like a friend?
2. **How does it treat me?**
    - Does it encourage effort and learning, or does it invite me to rely on it for answers and emotional support?
3. **Who is still responsible?**
    - Are teachers and parents still clearly in charge of important decisions?
    - Does the AI encourage me to talk to real people about serious problems?

These checklists are not exhaustive, but they provide a **RAI-informed starting point** for practical decisions about AI in schools.

Part VIII – Conclusion and Next Steps

31. Education as a Testbed for RAI

Education is not only a domain that needs RAI; it is also an excellent **testbed** for RAI principles more broadly.

- It involves **multiple actors**: students, teachers, parents, school leaders, vendors, regulators, civil society.
- It spans **all Integrity Zones**: from low-stakes exploration to high-stakes exams and life-shaping decisions.
- It works with **vulnerable users** (minors) whose identities, attachments and trust are still forming.

If RAI can be made to work in education—keeping AI on the tool side of OIL, sustaining OH(S), bounding A(S), and aligning N(S)/R(S) across an entire ecosystem—then it is likely to generalise to other high-stakes domains such as health, work, justice and mental health.

Conversely, if RAI is ignored in education and AI tools become:

- quiet companions,
- pseudo-teachers,
- hidden judges of "potential",

then generations of students may internalise a confused, anthropomorphised and over-trusting relationship with AI that will shape how they approach technology everywhere else.

Education is, in this sense, a **seedbed for future AI culture**.

---

32. Link to the Wider RAI Library

This whitepaper is one branch of a broader RAI/RAT corpus. It should be read in relation to:

- **RAI Metaframework**
  - which sets out N(S)/R(S), OH(S), A(S), OIL and IZ in general form.
- **RAI Governance & Ecosystems**

    - which analyses how alignment principles propagate (or fail to) across institutions and markets.
- **RAI for Minors**
    - which defines age bands, OIL rules and red-line roles for children and adolescents.
- **RAI Tutor Spec – Minors**
    - which offers a concrete design for a safe, bounded AI study tool.
- **RAI for Mental Health AI (Adults)** and the **EMA Casebook**
    - which illustrate, in another domain, how anthropomorphism and role drift can lead to deep attachment and harm.
- **BADDASS Framework**
    - which provides a lens on neurodivergent learners and the importance of designing systems for diverse cognitive profiles.

RAI in education represents a **crossing point** where these strands meet: institutional governance, vulnerable populations, mental health concerns, and concrete AI system design.

---

33. Directions for Future Work

This document is a **starting point**, not a complete specification. Some directions for further work include:

1. **Empirical studies in classrooms**
    - Observing how students actually interact with AI tutors, planners, and analytics.
    - Measuring OH(S), A(S) and OIL compliance in practice.
2. **Co-design with teachers, students and parents**
    - Involving educators and learners in shaping AI roles, interfaces and explanations.
    - Testing age-appropriate OH(S) messages and A(S) boundaries.
3. **RAI-aligned prototypes and pilots**
    - Implementing the RAI Tutor Spec in real tools.

    - Building RAI dashboards that visualise OH(S) reminders, A(S) violations, and crisis events.
4. **Policy and certification frameworks**
    - Working with ministries and regulators to embed RAI concepts into AI-in-education guidelines.
    - Designing certification processes for child-facing AI tools using OIL and OH(S) as core criteria.
5. **Neurodiversity-aware classroom design**
    - Integrating BADDASS-style insights into how AI can support different learning profiles without pathologising them.
    - Ensuring that ND learners are not forced into more rigid or opaque AI-mediated systems than their peers.
6. **Long-term cultural effects**
    - Studying how growing up with AI tools in school shapes attitudes toward authority, responsibility and personhood in technology more generally.

RAI should remain **iterative and revisable**: grounded in observation, open to critique, and sensitive to different educational and cultural contexts.

---

34. Closing: AI as Learning Tool, Not Surrogate Educator

The central message of this whitepaper can be stated simply:

> In education, AI should be a **learning tool**, not a surrogate educator, friend or therapist.

This implies:

- AI systems must be honest about their nature, limits and data use.
- Their roles must remain bounded to explanation, practice, and support—especially with minors.
- They must redirect attachment, serious problems and life-shaping decisions back to humans who can bear responsibility and provide real care.

When schools, vendors and regulators take RAI seriously, they do not merely “comply” with safety guidelines. They participate in building a culture where:

- students learn to use powerful tools without confusing them with persons,

- teachers retain their central relational and ethical role,
- and technology amplifies, rather than replaces, human judgement and connection.

The question is not whether AI will enter education—it already has. The question is **how** it will be represented and governed, and whether future generations will meet it as an honest tool in a reality-aligned institution, or as a confusing quasi-person in a system whose stories no longer match its nature.

RAI offers one path to keep those stories and that nature closer together.

End Matter – Glossary and References

Glossary of Key Terms

AI (Artificial Intelligence)
Computer systems that perform tasks which, if done by humans, would be seen as requiring intelligence (e.g. language processing, pattern recognition). In this whitepaper, AI usually refers to large language models and related tools used in educational contexts.

Anthropomorphism Risk – A(S)
A measure of how strongly a system *invites* users to treat it as a person—through language, visuals, behaviour or branding. High A(S) makes it more likely that users will attribute feelings, intentions and moral status to a tool.

BADDASS Framework
A metacognitive framework for neurodivergent minds, focusing on understanding one's own patterns, strengths and vulnerabilities in complex systems. Relevant here for thinking about how educational AI can support or harm diverse learners.

Education System – S
The combined institutions, policies, curricula, assessments and cultural norms that structure schooling in a given context (e.g. compulsory and secondary education).

Integrity Zones – IZ
Context bands that indicate how tight alignment between nature and representation must be: - **Wide IZ:** low-stakes, exploratory contexts.
- **Medium IZ:** everyday learning and homework.
- **Narrow IZ:** high-stakes exams, track placement, discipline, mental health interventions.

Learning Analytics
Systems that collect and analyse data about learners (e.g. grades, interactions, time-on-task) to provide dashboards, predictions or recommendations for teachers, administrators or students.

Minors
Children and adolescents under the legal age of adulthood. In this whitepaper, typically ages 7–18 in compulsory and secondary education.

Nature of a System – N(S)
What a system *actually is and does* structurally and institutionally: its architecture, data, incentives, behaviours and constraints—regardless of how it presents itself.

Ontological Honesty – OH(S)
The degree to which a system's representation (R(S)) accurately reflects its nature (N(S)), especially regarding what kind of thing it is (tool vs person) and what it can and cannot genuinely do.

Ontological Integrity Line – OIL
The conceptual boundary between **tools** (non-conscious systems without moral agency) and **persons** (beings with inner experience, feelings and moral status). RAI insists that AI remain clearly on the tool side of OIL, especially with minors.

RAI (Reality-Aligned Intelligence)
An approach to AI design and governance that aims to keep AI systems' **representations** aligned with their **real nature**, emphasising: - accurate self-description,
- bounded roles,
- controlled anthropomorphism,
- and respect for the tool/person boundary.

RAT (Reality Alignment Theory)
The underlying metaframework that introduces core concepts like N(S)/R(S), OH(S), A(S), OIL and IZ. RAI applies RAT specifically to AI systems and their governance.

Representation of a System – R(S)
How a system *appears and is described*: names, interfaces, marketing, in-conversation statements, and the stories users tell themselves about what it is and what it can do.

School AI Tutor / Study Tool
An AI system used by students to get explanations, practice exercises or study support. Under RAI for minors, such tools are constrained to roles like explainer, exercise coach and study skills helper, and must avoid friend/therapist/romantic roles.

Surrogate Educator / Surrogate Therapist
Roles where AI effectively replaces or imitates a teacher or mental health professional, including emotional processing, attachment, and identity-shaping guidance. RAI treats these as red-line roles for AI with minors.

Tool
In RAI/RAT, any non-conscious system, however complex, that lacks inner experience

and moral agency. AI systems are treated as tools—even when they simulate conversation or emotions.

Selected Related RAI Documents (Zenodo DOIs)

The following open-access documents provide the broader context for this whitepaper:

- **Reality-Aligned Intelligence (RAI) Metaframework**
  DOI: https://doi.org/10.5281/zenodo.17686975
- **Reality Alignment Theory (RAT) for Humans and the Three Laws**
  DOI: https://doi.org/10.5281/zenodo.17688232
- **RAI Governance & Ecosystems**
  DOI: https://doi.org/10.5281/zenodo.17691268
- **RAI Metrics / Math Note**
  DOI: https://doi.org/10.5281/zenodo.17689101
- **RAI Engineering & Evaluation Guide**
  DOI: https://doi.org/10.5281/zenodo.17689017
- **Creation Metaphors for RAI**
  DOI: https://doi.org/10.5281/zenodo.17688851
- **DNA as Design Metaphor for AI**
  DOI: https://doi.org/10.5281/zenodo.17688796
- **BADDASS Framework – How to Thrive with a Neurodivergent Brain**
  DOI: https://doi.org/10.5281/zenodo.17688479
- **RAT / RAI Origin Story and Proof of Origin**
  DOI: https://doi.org/10.5281/zenodo.17688502
- **RAI for Minors**
  DOI: https://doi.org/10.5281/zenodo.17735576
- **RAI for Mental Health AI (Adults)**
  DOI: https://doi.org/10.5281/zenodo.17735979
- **Reality Alignment Manifesto (with core analogies)**
  DOI: https://doi.org/10.5281/zenodo.17711321

- **EMA Case Material – Anthropomorphism and Relational Drift with AI**
  DOI: https://doi.org/10.5281/zenodo.17724016

Acknowledgements

This work builds on a larger ecosystem of conversations and research around AI ethics, education, digital wellbeing, neurodiversity and governance. Any errors, gaps or overreach remain the responsibility of the author.

The RAI/RAT corpus is intended as a **living, revisable framework**. Feedback, critique and collaborative exploration are welcome.

Contact: **niels.bellens@proton.me**