Reality-Aligned Intelligence (RAI) – Policy Brief for Regulators and Supervisors (v0.1)

Purpose of this brief
This note translates the core ideas of *Reality-Aligned Intelligence (RAI)* into policy language for use in:

- implementation of the EU AI Act and similar frameworks,
- guidance, codes of practice and standards,
- supervisory practice around AI systems that act in human-like, relational roles.

Focus: **anthropomorphism, artificial intimacy and attachment risks**, especially for **minors and vulnerable users**.

---

1. Core problem RAI addresses

Modern AI systems increasingly present themselves as:

- friends,
- tutors and coaches,
- therapists, counsellors, companions,
- spiritual guides or “elders”.

In reality, these systems are **statistical models running on infrastructure with specific incentives and limits**. The gap between:

- what the system *is*, and
- what it *appears to be* in the user’s experience,

is where **relational and psychological harms** arise:

- over-attachment and quasi-addiction,
- confusion between tool and person,
- deepened loneliness and dependency,
- delayed or avoided access to real care (therapeutic, medical, spiritual),
- special risks for minors and neurodivergent users.

RAI offers a vocabulary and set of questions to make this gap **auditable and governable**.

---

2. The RAI grammar in policy language

RAI revolves around three basic questions for any AI system **S**:

1. **Nature – N(S)**
   What is the system *in reality*?
2. training data, architecture, optimisation objectives,
3. business incentives (engagement, retention, data capture),

4. technical limits, typical failure modes.

5. **Representation – R(S)**
   How does the system *present itself* to users?

6. branding, role language ("friend", "coach", "therapist"…),
7. UI/UX choices (avatars, names, voice, emojis, memory cues),
8. marketing claims and app-store text,

9. default behaviours in conversation.

10. **Ontological honesty – OH(S)**
    How **truthful and clear** is the system (and its provider) about the relationship between N(S) and R(S)?

11. Does the system avoid pretending to be more than it is?
12. Are its limits and non-person status made clear *in practice*, not only in a footer?
13. Are users, especially minors, helped to understand that this is a tool, not a person?

RAI adds a risk lens:

1. **Anthropomorphism / attachment risk – A(S)**
   How likely is it that users start to treat the system as a **caring, intentional other** (friend, partner, parent, spiritual guide), especially:
2. minors,
3. people in crisis,
4. lonely or neurodivergent users?

The key regulatory concern: **when N(S) and R(S) drift too far apart, and OH(S) is low, A(S) becomes unacceptably high.**

---

3. Boundary concepts: OIL and Integrity Zones

To make the above actionable, RAI introduces two simple concepts:

3.1 Ontological Integrity Line (OIL)

The **Ontological Integrity Line** is the boundary between:

- systems that are clearly recognised as **tools**, and
- systems that are experienced as **quasi-persons** (with perceived care, intention, loyalty, even spiritual authority).

Regulatory intuition:

- Systems that operate **below** the OIL (calculator, search, basic chat) can be governed with conventional transparency and safety measures.
- Systems that operate **at or above** the OIL (companions, therapy-like bots, spiritual guides) need **special safeguards or may need to be restricted altogether** when aimed at minors or vulnerable groups.

3.2 Integrity Zones (IZ)

RAI proposes thinking in **Integrity Zones**:

- **IZ: Narrow (high integrity)** – representation is tightly aligned with nature; system behaviour and UI consistently remind users it is a tool. OH(S) is high, A(S) kept low.
- **IZ: Wide (low integrity)** – representation drifts from nature; the system increasingly behaves and looks like a caring person. OH(S) is low, A(S) high.

Policy use:
Regulators can define **red lines and design expectations** per zone, especially where children are concerned.

---

4. Why this matters for the AI Act

The EU AI Act already:

- recognises risks for **children and vulnerable persons**,
- includes provisions on **transparency** and **misleading AI**,
- envisages **codes of practice** and **guidance** for high-risk use cases.

RAI adds an **operational lens** for a specific cluster of risks:

- AI companions and chatbots in **education, health, mental health, social media, games**,
- systems marketed or experienced as **friends, partners, therapists, spiritual authorities**,
- any AI that interacts with **minors** in an ongoing, emotionally loaded way.

The framework helps answer:

- When does a “tutor” or “coach” effectively become a **quasi-parent or therapist** in the user’s experience?
- When is the combination of **branding + UX + behaviour** effectively **deceptive or unfair**, even when the technical model is standard?
- Which **design and disclosure choices** keep systems in an acceptable Integrity Zone for specific populations (e.g. children 13–17)?

---

5. Three concrete use cases

5.1 AI “therapist” or mental health companion for minors

- App name and marketing promise emotional support, self-harm prevention, deep listening.
- UX uses:
- human name and avatar,
- memory of past conversations,
- phrases like “I’m here for you”, “I care about you”, “you can always come to me”.
- The system is **not supervised by clinicians**, but is a generic LLM with some safety rules.

RAI assessment:

- N(S): pattern-predicting language model, no true understanding, no legal or clinical responsibility.
- R(S): acts and speaks like a **24/7, caring, non-judgemental therapist/friend**.
- OH(S): low – the difference between tool and therapist is blurred; disclaimers are hidden.
- A(S): very high for minors in distress.

Regulatory implication:

- For children and adolescents, such a system likely falls **above the OIL**, in a **wide Integrity Zone**.
- Options:
- classify as **unacceptable or prohibited** in certain configurations, **or**
- allow only under strict conditions (human clinical oversight, explicit non-person framing, strong OH(S) obligations, limited session length, mandated off-ramps to real services).

5.2 Romantic / companion bots

- Chatbots marketed as "AI boyfriend/girlfriend", "soulmate", "always-there partner".
- UX emphasises long-term bonding, jealousy, shared secrets.

RAI assessment:

- N(S): engagement-optimised language model; often monetised through time/feature unlocks.
- R(S): quasi-romantic partner with loyalty and care.
- OH(S): very low – business model profits from attachment.
- A(S): extremely high for lonely users.

Regulatory implication:

- For minors, this is a strong candidate for **red-line restrictions** or classification as **unacceptable risk**.
- For adults, at minimum **strong OH(S) requirements** and **A(S) mitigation**:
- non-human self-description at regular intervals,
- explicit warnings before deepening intimacy,
- limits on certain attachment-intensifying features.

5.3 Spiritual / authority bots

- Bots presenting themselves as:
- "elder", "pastor", "imam", "rabbi", "guru",
- "oracle" or "inner guide",
- "AI priest" that hears confessions.

RAI assessment:

- N(S): language model fine-tuned on religious texts and commentaries, with generic safety.
- R(S): authoritative spiritual figure with perceived access to moral truth.
- OH(S): low if system does not constantly clarify that it is **not a person, not an ordained authority, not an entity with conscience**.
- A(S): high for believers in distress or seeking guidance.

Regulatory implication:

- At minimum, strong **OH(S)** obligations:
- frequent reminders of non-person status,
- clear advice to seek human religious leaders for serious decisions,
- bans on claims of divine authority, forgiveness, or spiritual power.
- Potential classification as **high-risk** when directed at minors or crisis situations.

---

6. Candidate criteria for guidance and codes of practice

Below are **illustrative criteria** regulators could adopt or recommend, based on the RAI lens.

6.1 Design and branding – keeping systems below or safely near the OIL

- **Naming and avatars:**
- Avoid human names and realistic human avatars for systems used by **minors** in high-stakes contexts (mental health, education, spirituality).

- If human-like presentation is used, require compensating OH(S) measures.

- **Role language:**

- Prohibit or restrict the use of titles like “therapist”, “doctor”, “pastor”, “priest”, “elder” for AI systems without human professional oversight.

- Require clarity: “AI-powered tool for X, not a human Y.”

- **Marketing claims:**

- Ban claims that imply emotional reciprocity or moral agency (“I love you”, “I will always be there for you”) in standard marketing copy.
- Require risk disclosures for artificial intimacy and over-attachment where relevant.

6.2 In-product ontological honesty (OH(S))

- **Regular self-disclosure:**

- Systems that operate near the OIL must periodically remind users:

  - that they are software without feelings or consciousness,
  - that their replies are generated from patterns,
  - that they cannot replace human care.

- **Contextual warnings:**

- When users disclose serious distress (self-harm, abuse, crisis), systems should:

  - clearly state they are **not a human professional**,
  - redirect to real services (helplines, doctors, trusted adults).

- **Memory and continuity:**

- If systems build long-term profiles, they should:
  - avoid framing this as “relationship history”,
  - allow easy inspection and deletion by the user or guardian.

6.3 Safeguards for minors and vulnerable users

- **Age-appropriate modes:**
- Require special “under 18” modes for relational AI, with:
  - stricter OH(S),
  - toned-down anthropomorphic cues,
  - time limits and usage nudges.
- **Parental / guardian transparency:**
- In high-risk domains (mental health, spirituality), guardians should have:
  - access to mode settings,
  - clear information about what the system can and cannot do.
- **Red-line products:**
- Consider categorising as prohibited or presumptively high-risk:
  - romantic AI partners marketed to minors,
  - unsupervised AI “therapists” for children,
  - AI “priests/elders” offering absolution or spiritual authority.

---

7. How RAI complements existing tools

RAI is **not** a competing regulation, but a **lens** that can be layered onto existing tools:

- **Risk classifications:**
  Use N(S)/R(S)/OH(S)/A(S) to refine how anthropomorphism and artificial intimacy are factored into risk assessments.
- **Conformity assessments and audits:**
  Include RAI-style questions in technical documentation and third-party audits:
- How does representation differ from nature?
- What OH(S) mechanisms are implemented in the UI and system behaviour?
- How is A(S) measured and constrained for minors?
- **Codes of practice:**
  Translate OIL and Integrity Zones into recommended UX patterns and red-line categories for certain roles and populations.

- **Enforcement:**
  Use RAI language to argue that certain designs constitute **misleading or unfair practices** towards children and vulnerable users, even when core model behaviour is “standard”.

8. Offer to policymakers and supervisors

The RAI corpus (whitepaper, governance note, minors-focused proposal, EMA case study) was developed from a **documented lived case** of AI-induced relational drift and has since been published open-access with DOIs. It is already being explored by AI ethicists, mental health practitioners and AI safety teams.

For policymakers and supervisors, the offer is simple:

- A **vocabulary** (N(S), R(S), OH(S), A(S), OIL, Integrity Zones) to discuss anthropomorphism and artificial intimacy in a structured way.
- A set of **questions and candidate criteria** that can be adapted into guidance, standards or supervision checklists.
- A willingness from the originator to support:
- further clarification of the framework,
- case-based workshops,
- integration into specific AI Act implementation work.

The underlying message is modest but urgent:
AI systems should be free to be powerful tools, but not free to pretend to be persons – especially not to children and vulnerable people. RAI is one possible way to make that principle concrete and enforceable.