Refusal Is Not Referral

Scope, escalation, and institutional duty for AI deployed in children's spaces

Working paper · Fourth revision · August 2026

The missing third state

A child asks a kindergarten teacher a question about sex. In many schools the teacher will not answer it — and will not merely refuse. The teacher recognises the question as outside the role, routes it toward someone whose role includes it, and, depending on what was disclosed, ensures the appropriate adult learns it was asked. Three things happened and only one was a refusal.

Conversational AI reproduces almost none of this. Systems answer or decline; substantial engineering goes into calibrating the boundary between the two. The institutional pattern requires a third state: refer. A refusal ends the deployer's exposure. A referral serves the child. These are different objectives, and current practice optimises for the first.

A generic suggestion is not a referral

A referral has a destination. Three levels should be distinguished, because they are routinely conflated. Signposting — talk to a trusted adult, a number on the screen — has no destination, no transfer and no record, and is what most current practice provides. Referral names a reachable person or office, transfers the relevant context, and produces a record that the transfer occurred. A warm handoff means the destination has affirmatively accepted context and responsibility.

Signposting is proportionate for many out-of-bounds questions. It is not sufficient where a safeguarding concern is present.

Escalation, and the route that runs the other way

The ordinary escalation route runs to the guardian. The override runs in the opposite direction, and it is the more important half: where the parent is the suspected source of harm, notification converts a disclosure of abuse into a disclosure to the abuser. An escalation duty without a protective-authority route can be worse than none.

The system is not a statutory reporter. It is an internal conduit: it routes to a designated human, who forms or declines to form suspicion and executes any statutory report. That keeps the legally operative judgment where the statutes already put it.

Fitness, not age

At a liquor store the customer proves age — one bit, tested once. At a school the adult proves fitness: checked, credentialed, trained, renewed, revocable, while the child proves nothing. These are categorically different regimes, and a children's space needs the second. Fitness is continuing, so certification needs audit intervals rather than a one-time gate; it is revocable, so something must exist to withdraw; and it attaches to the deployer, because the deployer is the party capable of bearing a continuing obligation.

Auditability, not a chaperone

Youth-serving organisations converged on rules limiting unsupervised one-to-one contact — two-deep leadership, open-door policies, chaperone requirements. It is tempting to observe that conversational AI is structurally the configuration those rules restrict, and to demand a chaperone. That would be a mistake: it anthropomorphises software, and it destroys what makes AI tutoring valuable, which is the freedom to be wrong in private.

The rule is not really about the second adult. It is about auditability, and software admits better implementations. Reviewable is not surveilled: routine interactions retained but unread, flagged interactions escalated to a named human, aggregate patterns available to auditors.

Where the duty attaches, and what is not claimed

No duty attaches to the model. When an institution places an automated system into a role that functionally replicates professional contact with a child, its own existing duty of care is engaged — it has selected an instrumentality, not hired an employee. The paper claims no language model is a moral agent, does not assert that existing doctrine already imposes these duties on software, and does not address content classification or age assurance.

The claim most exposed is stated as a hypothesis rather than a result: that conventional safety evaluation rewards successful non-compliance without separately measuring whether an out-of-scope interaction reached an appropriate human destination. That is the first place research effort should go.

The full paper sets out the three-axis state model — response, escalation, accountability — in the form a procurement officer could specify and an auditor could test, together with the capacity and compelled-speech objections in full.

To request the full paper, write to shknudson@ifcsis.org.