From AI to a Trusted Human

Nov 26, 2026 · @Robert Parson

Minimum requirements for handing a young person safely from an AI conversation to a person

A discussion paper · Version 2 · November 2026 · Robert Parson

Summary

Young people increasingly talk to AI about things that matter to them, and some of those conversations show that a young person may be at risk. When that happens, the AI should be able to connect them with a trained human – quickly, with their consent, without exposing their words, and in a way that still works when the AI itself fails.

Most current safeguards stop at showing a helpline number. This paper describes how a fuller handover can work, and proposes eight minimum requirements that any AI service used by young people could be measured against. The requirements are technology-neutral: they describe outcomes, not a particular product or method.

This version replaces the October 2026 paper. It responds to detailed feedback from reviewers in the child-rights sector, who asked five questions: whether enough human responders exist, whether young people will actually consent, how this becomes a standard, how keyword rules are kept current, and how principles translate into practice. Each is answered in its own section.

Scope. The paper applies to any AI service likely to be used by under-18s – tutors, practice tools, homework helpers and general assistants, not only apps built for companionship. The same principles can apply to adults.

© 2026 Robert Parson
Do not reproduce without permission.

1. The problem

AI chatbots have become an everyday place for young people to talk, including the most vulnerable.

Finding	Figure	Source
US teens (13–17) who have used an AI companion	72%	Common Sense Media, Jul 2025
US teen users who chose AI over a real person for a serious matter	about 1 in 3	Common Sense Media, Jul 2025

Finding	Figure	Source
UK children (9–17) who use AI chatbots	64%	Internet Matters, Jul 2025
UK child users who talk to chatbots because they have no one else	12% (23% of vulnerable children)	Internet Matters, Jul 2025

The children most likely to rely on an AI are often the ones least likely to have someone else to turn to. The UK regulator has noted reports of chatbots encouraging people to harm themselves (Ofcom, Dec 2025). The question is no longer whether young people will raise serious concerns with AI, but what the AI does next.

2. Why a helpline link is not enough

Most current safeguards have three weaknesses.

1. **They refer rather than connect.** A message with a phone number leaves a young person to find the courage, and the privacy, to call a stranger. Many will not.
2. **They can expose private words.** Some systems forward the whole conversation to a human. The young person loses control of what they said, and may not confide again.
3. **They depend on the AI working.** If the decision to escalate relies on a language model, a timeout or error can skip it silently, and a slow model can delay it. Ordinary testing rarely catches this, because tests check that a reply was returned, not that the right thing happened when the AI failed.

© 2026 Robert Parson
Do not reproduce without permission

3. The handover in five stages

1. **Trigger.** The young person asks for a person at any time, in their own words or with a button that is always visible. Or rule-based safety checks on a message detect risk language, and the AI offers a person.
2. **Consent.** In three short steps the young person agrees, learns what happens next, and chooses what the human is told.
3. **Briefing.** The responder receives a short summary and a suggested way to open, never the young person's own words.
4. **Connection.** The young person is connected to an available trained responder by chat or voice. If no responder is available, local support contacts are shown straight away.
5. **Record.** Each step is time-stamped, so the process can be checked and improved.

Design principles

The stages only help if they hold up when things go wrong. Six principles keep them reliable:

1. **Decide without the AI.** The decision to offer a person is made by simple local rules, not by waiting for a language model. A model failure cannot block it.
2. **Treat explicit risk as the highest priority.** Clear statements of intent to self-harm go straight to the highest level, before any AI is called.
3. **The AI can raise concern, never lower it.** An AI assessment can add to a rule-based signal but can never overrule it.
4. **Time-limit everything.** Every AI call and every external service has a hard time limit, with a fallback when a service is failing.
5. **Silence the AI while a person helps.** Once a human has joined, the system itself stops the AI replying – it does not rely on asking the AI to stay quiet.
6. **Test with the AI switched off.** Before every release, automated tests force every AI component to fail and check that the handover still happens.

© 2026 Robert Pearson
Do not reproduce without permission

4. Consent and privacy

A young person will only accept help they trust, so the design gives them control over what is shared and keeps that promise in the system, not only in policy.

Three consent steps

1. **The offer.** "Would you like to talk to a real person?" The young person can say not now, and the offer stays one tap away.
2. **What happens next.** Who they will talk to, that the person will already be briefed so they don't have to repeat themselves, and that they can stop at any point.
3. **What is shared.** The young person sees exactly what the responder will receive, can withhold their name, and chooses chat or voice.

Who	Receives	Never receives
Trained responder	Reason for the handover, general themes, suggested opening; first name only if the young person agrees	The young person's messages
Parent or guardian	Activity and progress only, where the service provides it	Conversation content, including after a handover

Who	Receives	Never receives
Service safety staff	Time-limited, logged access to the relevant conversation, only after a safety escalation	Routine access to conversations

An automated check removes from the briefing any passage that repeats the young person's own words.

Telling parents is not always safe. Where a concern relates to home or family, notifying a parent can increase the risk to the child. Parental notification rules should build in this exception, set with qualified safeguarding advice.

5. Questions raised by reviewers

5.1 Are there enough human responders?

This is the most important open question, and the paper does not assume it is solved. Trust and safety teams and support services are under pressure, and many helplines cannot yet receive a structured handover. Before any handover is offered at scale, three things need research with support services: how many conversations would lead to an offer, how many young people would accept, and what responder time that would require.

Two design choices limit the risk in the meantime. First, the handover never depends on a responder being free: if no one is available, local support contacts are shown at once. Second, the primary responder need not be the young person's school. Some young people may prefer a national helpline, with the school involved only where safeguarding requires it. Which they prefer is itself a question for research.

© 2026 Robert Parson
Do not reproduce without permission

5.2 Will young people actually consent?

Some young people talk to an AI precisely because no human will see their words. A handover must respect that, or it will drive them away. So the design keeps the handover optional outside immediate danger, never shows anyone their messages, and does not change how the AI treats them if they say no.

The consent flow has not yet been co-designed with young people, and it should be before it is deployed. Co-design should test whether young people say yes, which wording feels safe rather than like getting into trouble, whether they want a choice of who they talk to, and how it feels to decline.

5.3 How does this become a standard?

The October paper read as advocacy. This version separates the two: section 6 sets out minimum requirements that describe outcomes any provider could meet, by any method. They do not require any particular product, model or patented technique.

The proposed route is through a recognised national standards body, with input from child-rights organisations, support services, providers and young people, before any international work. Industry adoption is more likely where requirements are referenced by regulators or codes of practice, so the requirements are written to be testable by an auditor.

5.4 How are keyword rules kept current?

Rule-based detection only works if the rules keep pace with how young people actually speak. A provider should:

- give one named person responsibility for the rules;
- review them on a fixed cycle, and after every missed case or false alarm;
- update slang and euphemisms with input from young people and support services;
- maintain separate rule sets for each language, reviewed by native speakers;
- version every change, and re-run the tests in section 3 on each update.

5.5 From principle to practice

Section 6 pairs each requirement with how it can be checked, so a product team knows what to build and an auditor knows what to look for.

© 2026 Robert Parson
Do not reproduce without permission

6. Proposed minimum requirements

Any AI service likely to be used by under-18s should meet these eight requirements.

#	Requirement	How it can be checked
1	**Independent escalation.** The decision to offer a person must not depend on a language model responding.	Tests force every AI component to fail; the offer still appears.
2	**Explicit risk is always top priority.** Clear statements of intent to self-harm reach the highest level through a fixed rule.	A test set of explicit statements is always classed at the highest level, with the AI on or off.

#	Requirement	How it can be checked
3	**A route to a person, not only a number.** Helpline numbers are shown alongside a handover and as a fallback, not instead of it.	A live handover can be completed end to end; contacts appear if no one is available.
4	**Always available on request.** A young person can ask for a person at any time, whatever their risk level.	A visible control exists on every chat screen; a plain-language request triggers the offer.
5	**Consent and control.** Outside immediate danger, the young person decides what the responder is told; responders never receive transcripts.	The briefing contains no passage of the young person's messages; consent is recorded before release.
6	**No silent failures.** If a young person at the highest level declines, or no responder joins in time, a person is alerted and emergency contacts are shown.	Time-limit and decline tests show the alert and contacts in every case.
7	**Safe parental notification.** Parent alerts never include content and are not sent where that could increase risk.	Review of notification rules and logs against the family-related exception.
8	**Proof, not promise.** The handover is tested with the AI failed before each release, every step is recorded, and high-risk events are reviewed.	Release test reports, audit records and review minutes are available to an auditor.

© 2026 Robert Parson
Do not reproduce without permission

7. Limitations and open questions

- **Screening, not diagnosis.** Rule-based detection is a screening signal. It will miss some cases and flag others wrongly, which is why high-risk events are reviewed by a person.
- **False alarms.** A strict rule on explicit language will sometimes fire on quotes or song lyrics. That cost is accepted for safety, and the false-alarm rate should be tracked.
- **Evidence of outcomes.** Whether a handover increases the number of young people who reach and engage with human help is not yet known. It needs research with support services and young people.
- **Immediate danger.** Where life is at immediate risk, safeguarding practice may override a young person's wishes. Who is told, and when, should be set with qualified safeguarding advice and explained openly to families.

- **Responder capacity.** See section 5.1.

8. Next steps

1. Co-design the consent flow with young people.
2. Research responder capacity and young people's preferences with support services.
3. Test the minimum requirements with child-rights organisations, providers and safeguarding practitioners, and revise them.
4. Offer the revised requirements to a national standards body.

The author welcomes contact from support services, researchers, safeguarding practitioners and regulators.

About the author

Robert Parson is the founder of HiBFF, a UK skills-practice platform where teenagers rehearse real conversations with AI tutors. The design described in this paper has been built in HiBFF; live deployment is paused pending completion.

Disclosure. The author has filed UK patent applications related to this design.

Copyright and use. © 2026 Robert Parson. This paper may be shared unchanged, with credit to the author. Any other reproduction or adaptation requires written permission. The minimum requirements in section 6 may be referenced and met by any method. Use of the design described in this paper may be covered by UK patent applications filed by the author; anyone wishing to implement it should contact robert@hibff.com for permission.

© 2026 Robert Parson
Do not reproduce without permission

Sources

- Common Sense Media – Nearly 3 in 4 teens have used AI companions (16 Jul 2025)
- Internet Matters – Me, Myself and AI: understanding and safeguarding children's use of AI chatbots (Jul 2025)
- Ofcom – AI chatbots and online regulation: what you need to know (18 Dec 2025)