

NEWSLETTER #3



SHIELD

European AI-Powered Gender
Based Violence Reduction Initiative



If an AI chatbot is meant to support young people facing online gender-based violence, what should it never get wrong?

This question is becoming central to the SHIELD project. As partners continue developing an AI-powered chatbot and support pathway, the focus is not only on what artificial intelligence can do, but on what young people, parents, educators, youth workers, and gender-based violence experts would actually trust. This matters because gender-based violence is not only a technical problem; it is rooted in unequal power relations and can move across offline and online spaces [1,2].

Building Trust Into AI

Over the past months, SHIELD partners have been gathering evidence through needs assessment activities, focus groups, and expert discussions across partner countries. These activities are helping the project move from broad ambition to concrete design priorities: how should a chatbot speak to young people, what kind of guidance should it provide, when should it hand over to human support, and what safeguards must be in place from the beginning?

Young people need language that feels clear, non-judgemental, and easy to recognise. Parents and guardians are concerned that AI tools may lack empathy or emotional sensitivity. Youth workers, educators, and experts are especially attentive to risks such as bias, errors, unsafe escalation, and data misuse. Across these perspectives, privacy stands out as a non-negotiable condition for trust.

Online gender-based violence is shaped by shame, fear, uncertainty, social pressure, and distrust of reporting systems. A chatbot that simply detects harmful language is therefore not enough. To be useful, AI must be embedded in a wider support pathway that is privacy-first, trauma-informed, human-in-the-loop, and accountable to the people it is meant to serve. This aligns with wider guidance on technology-facilitated gender-based violence and ethical AI, which emphasises safety, transparency, context, and safeguards for affected communities [2,3,4].

WHAT THIS MEANS FOR SHIELD

The goal is not to build “more AI” for its own sake. The goal is to build safer, more trustworthy digital support for young people navigating online harm. Evidence on conversational agents suggests that such tools are most appropriate when designed as complements to human care, not substitutes for professional judgement or safeguarding systems [5,6].

In the coming months, SHIELD will continue refining chatbot requirements, strengthening safeguarding procedures, and working with partners to ensure that ethical design remains at the centre of development. Trust will not come from technology alone. It will come from careful listening, transparent choices, and continued collaboration with the communities SHIELD aims to support.



Co-funded by
the European Union