

Principles for Relational Safeguards in AI Mental Health Tools



Time-Bound

Interventions must be time-bound. AI tools should not operate as 'always on' therapeutic interventions. Users need space outside of digital input to practise skills and integrate learning in their own time. Sessions should have clear openings and closings.



Supportive

Tools must be supportive – not directive. AI may provide space for psychoeducation, journalling, language-finding, and other non-relational supports. It should not create treatment plans or attempt to replace the reflexive, human processes of therapy.



Safeguarded

The tool must provide safeguarding support and human escalation. AI tools should be explicit about their safeguarding capacity and limits. When risk thresholds are met, they must signpost users to real, human support, and wherever possible, activate engagement with real, live intervention services.



Adjunctive

The interventions must be adjunctive to therapy: not therapy itself. Where used alongside therapy, AI should support, not replace, therapist judgement. Summaries or prompts should be optional, reviewable at the therapist's discretion, and never override clinical reflection or supervision.



Transparent

The tool must be transparent about its limitations. AI must communicate its boundaries clearly, reminding users that it is a supportive tool for reflection and organisation, not a therapeutic relationship. It should also be transparent about the approaches and assumptions embedded in its design.



User-Autonomous

Users must retain total autonomy and choice. They must be able to opt in or out of features, adjust levels of interaction, and decide how their data is stored or deleted. Tools should be designed to empower, not to steer or prescribe. Users should be able to feedback at any, and on any, points.



NATIONAL
COUNSELLING &
PSYCHOTHERAPY
SOCIETY

Contact: Meg Moss
Head of Public Affairs & Advocacy
meg@ncps.com