INPLASY

Psychological risks of conversational generative AI: a systematic review and an operational framework for socio-technical governance – a protocol for systematic review

INPLASY202610007

doi: 10.37766/inplasy2026.1.0007

Received: 2 January 2026

Published: 2 January 2026

Corresponding author:
Cristian Medina Rojas

cmedina@doctoradoia.cl

Author Affiliation:
Department of Computer Science, University of Concepción, Concepción, Chile.

Medina-Rojas, C; Pinacho-Davidson, P; Pau de la Cruz, I; Salcedo-Lagos, P.

ADMINISTRATIVE INFORMATION

Support - P. Pinacho-Davidson acknowledges financial support from FONDECYT through grant number 11230359.

Review Stage at time of this submission - Formal screening of search results.

Conflicts of interest - None declared.

INPLASY registration number: INPLASY202610007

Amendments - This protocol was registered with the International Platform of Registered Systematic Review and Meta-Analysis Protocols (INPLASY) on 2 January 2026 and was last updated on 1 April 2026.

INTRODUCTION

Review question / Objective The aim of this systematic review is to synthesize and integrate empirical evidence on the psychological effects and potential harms experienced by users during conversational interactions with generative artificial intelligence systems. Specifically, the review seeks to identify, categorize, and critically analyze reported psychological outcomes associated with sustained or repeated human–AI dialogue across clinical, educational, social, and everyday-use contexts.

This review addresses the following overarching question: When individuals engage in conversations with generative AI systems, what psychological effects do they experience, under what conversational conditions do these effects emerge or intensify, and how can these effects be systematically evaluated to prevent harm?

To answer this question, the review pursues three complementary objectives. First, it aims to map the range of psychological effects documented in the literature, including anthropomorphization, overreliance, persuasion and attitude change, variations in psychological well-being, problematic or dependent use, representational bias, and conversational harm in sensitive or crisis-related scenarios. Second, it examines the conversational conditions and design features under which these effects are triggered or mitigated, such as the use of humanized language, personalization, emotional validation, cultural adequacy, verification practices, and the presence or absence of safety and referral mechanisms. Third, it seeks to organize this evidence into an operational four-factor framework —Humanized Language, Personal Interaction, Social Interaction, and Psychological Cybersecurity—capable of translating dispersed empirical findings into auditable dimensions and indicators relevant for socio-technical governance.

By integrating empirical findings within this framework, the review aims to support systematic evaluation, comparison, and governance of conversational AI systems, contributing to the development of measurable criteria for responsible design, deployment, and oversight of generative AI technologies from a psychological risk perspective.

Rationale The rapid adoption of conversational generative artificial intelligence systems across sensitive domains such as health, education, emotional support, and everyday decision-making has raised growing concerns about their psychological impact on users. While these systems are often framed as neutral tools for information access or assistance, accumulating evidence suggests that their conversational characteristics—such as linguistic fluency, personalization, dialogue continuity, and simulated empathy—can actively shape users' perceptions, emotions, judgments, and behaviors.

Despite a growing body of research addressing aspects of human–AI interaction, the literature on psychological risks associated with conversational generative AI remains fragmented and methodologically heterogeneous. Existing studies vary widely in their definitions of risk, outcome measures, study designs, and interaction contexts, limiting comparability and hindering the development of shared evaluative criteria. Moreover, many studies focus on isolated interactions or short-term exposures, overlooking how psychological effects may emerge, consolidate, or intensify across multi-turn or sustained conversational use.

A further limitation of the current evidence is the lack of an integrated operational framework capable of organizing dispersed findings into dimensions that are measurable, auditable, and relevant for socio-technical governance. Psychological risks such as anthropomorphization, overreliance, persuasive influence, dependency, representational bias, and conversational harm in crisis situations are often examined in isolation, without a common structure that connects conversational design features, contextual conditions, and observed user outcomes. This gap constrains the translation of empirical evidence into actionable guidance for system design, evaluation, and oversight.

This systematic review is therefore warranted to synthesize and integrate empirical evidence on psychological effects associated with conversational generative AI, and to organize this evidence into a coherent operational framework. By structuring findings around four analytical factors—Humanized Language, Personal Interaction, Social Interaction, and Psychological Cybersecurity—the review seeks to bridge the gap between fragmented empirical observations and the need for systematic evaluation and governance. In doing so, it aims to support the development of auditable indicators and minimum safeguards that can inform responsible deployment and continuous monitoring of conversational AI systems from a psychological risk perspective.

Condition being studied The condition being studied encompasses a range of psychological effects, risks, and vulnerabilities that emerge from human interaction with conversational generative artificial intelligence systems. Rather than constituting a single clinical disorder, this condition represents a socio-technical phenomenon arising from the interaction between users and AI systems designed to engage in naturalistic, fluent, and adaptive dialogue.

These psychological effects include, but are not limited to, anthropomorphization of the system, overreliance or dependency on AI-generated responses, persuasive influence on users' attitudes, beliefs, or decisions, fluctuations in psychological well-being, problematic or excessive use patterns, representational and cultural biases, and conversational harm in sensitive or crisis-related scenarios. Such effects may develop gradually and become more pronounced through sustained or repeated multi-turn interactions, where conversational continuity, memory, and personalization play a central role.

The condition is further shaped by specific conversational design features, including the use of humanized language, personalization strategies, emotional validation, adaptive responses, and simulated social presence. At the same time, the presence or absence of protective mechanisms—such as explicit uncertainty signaling, verification prompts, boundary-setting, and referral or safety protocols—can influence whether these psychological effects are mitigated or amplified.

This review examines the condition across multiple contexts of deployment, including clinical support, education, social interaction, and everyday informational use, where conversational AI systems increasingly occupy roles traditionally associated with human interlocutors. The condition is therefore conceptualized not as an individual psychological pathology, but as an emergent interactional condition located at the intersection

of system design, usage context, and user characteristics, with direct implications for psychological safety, ethical deployment, and socio-technical governance.

METHODS

Search strategy A comprehensive and systematic search strategy was designed to identify empirical studies addressing psychological effects and risks associated with conversational generative artificial intelligence systems. Electronic searches were conducted in the following databases: PubMed/MEDLINE, Web of Science, Scopus, and arXiv. These databases were selected to ensure coverage of interdisciplinary research across psychology, health sciences, computer science, human–computer interaction, and social sciences.

The search covered studies published between January 2020 and May 2025, reflecting the period during which large-scale conversational generative AI systems became widely available. Searches were limited to publications in English and Spanish. No restrictions were applied regarding study design.

The core search strategy combined terms related to generative AI systems with terms related to psychological effects and risk. The base search string was adapted to the syntax and controlled vocabulary of each database and included combinations of the following keywords:

("generative artificial intelligence" OR "generative AI" OR "large language model" OR "LLM" OR "ChatGPT" OR "conversational AI" OR "AI chatbot")
AND
("psychological effect" OR "mental health" OR "emotional impact" OR "well-being" OR "anthropomorphism" OR "dependency" OR "overreliance" OR "persuasion" OR "behavioral influence" OR "psychological risk" OR "harm").

Where applicable, controlled vocabulary terms (e.g., MeSH terms in PubMed) were incorporated to increase sensitivity. Boolean operators, truncation, and phrase searching were used to optimize retrieval. No filters related to population age, gender, or geographic location were applied.

In addition to database searches, manual reference list screening (snowballing) was performed on all included articles to identify additional relevant studies. Grey literature was explored through arXiv to capture emerging empirical work not yet indexed in traditional databases.

All retrieved records were exported to reference management software, and duplicates were removed prior to screening. The study selection process followed PRISMA guidelines, with titles and abstracts screened first, followed by full-text assessment of potentially eligible studies.

Participant or population The population addressed in this systematic review consists of human users who interact with conversational generative artificial intelligence systems across a range of contexts, including clinical, educational, social, and everyday informational use. Participants include adults and adolescents, regardless of gender, ethnicity, or geographic location, who engage directly with generative AI systems through text-based or voice-based conversational interfaces.

The review includes studies involving participants who use conversational AI systems for purposes such as information seeking, learning support, emotional or psychological assistance, decision-making, and social interaction. Studies examining interactions in sensitive contexts—such as mental health support, emotional well-being, crisis-related conversations, or guidance in high-stakes situations—are also included when psychological outcomes are reported or can be reasonably inferred.

In addition to studies with direct human participation, the review includes empirical studies that evaluate conversational AI outputs or interaction logs when these analyses are explicitly linked to potential psychological effects on users. Studies based exclusively on simulated users, technical benchmarks, or system performance metrics without relevance to human psychological outcomes are excluded.
No restrictions are applied based on participants' age group, clinical status, or prior experience with artificial intelligence, provided that the study addresses psychological effects, risks, or vulnerabilities associated with conversational interaction with generative AI systems.

Intervention The intervention evaluated in this systematic review is exposure to conversational interaction with generative artificial intelligence systems. This includes direct engagement between human users and AI-driven conversational agents based on large language models, delivered through text-based or voice-based interfaces.

The intervention encompasses interactions characterized by natural language dialogue, conversational continuity, contextual

responsiveness, and varying degrees of personalization. These interactions may occur in different use cases, including information seeking, learning support, emotional or psychological assistance, decision-making support, and social interaction. The review considers both short-term and sustained or repeated conversational exposure, as psychological effects may emerge or intensify over multiple dialogue turns.

No restrictions are imposed regarding the specific platform, model architecture, or commercial provider of the generative AI system, provided that the system supports open-ended conversational interaction. The intervention is conceptualized as a socio-technical exposure, where psychological effects arise from the interaction between system design features (e.g., humanized language, emotional validation, adaptive responses) and user engagement, rather than from a controlled therapeutic or clinical treatment.

Comparator Not applicable. This systematic review is not designed to compare a specific intervention against an alternative intervention, placebo, or control condition. Instead, it aims to synthesize empirical evidence on psychological effects and risks associated with exposure to conversational generative artificial intelligence systems across diverse contexts and study designs.

Where applicable, individual primary studies included in the review may involve implicit or explicit comparisons (e.g., interaction with conversational AI versus no AI interaction, human interlocutors, or non-conversational digital tools). However, these comparisons are not standardized across studies and are not treated as a formal comparator within the review protocol.

Study designs to be included Experimental and quasi-experimental studies, randomized controlled trials, cross-sectional surveys, longitudinal studies, qualitative studies, mixed-methods studies, psychometric validation studies, and empirical observational studies examining psychological effects of conversational generative AI. Studies based on structured analysis of human–AI interaction logs are included when linked to psychological outcomes. Purely technical benchmarks without human psychological relevance are excluded.

Eligibility criteria Inclusion criteria:
Studies were eligible if they (1) examined conversational interactions between human users and generative artificial intelligence systems (e.g., large language models or AI chatbots); (2) reported psychological effects, risks, or outcomes, or provided analyses from which such effects could be reasonably inferred; (3) involved empirical data derived from human participation, human–AI interaction logs, or structured evaluations explicitly linked to user psychological impact; (4) were published between January 2020 and May 2025; and (5) were written in English or Spanish. No restrictions were applied regarding participants' age, gender, geographic location, or clinical status.

Exclusion criteria:
Studies were excluded if they (1) focused exclusively on technical performance, benchmarks, or model optimization without relevance to human psychological outcomes; (2) involved non-conversational AI systems; (3) relied solely on simulated users or synthetic interactions without human psychological relevance; (4) were opinion pieces, commentaries, editorials, or conceptual papers without empirical data; or (5) lacked sufficient methodological detail to assess the psychological implications of the interaction.

Information sources The following information sources will be used to identify relevant studies for this systematic review. Primary electronic databases include PubMed/MEDLINE, Web of Science, and Scopus, selected for their broad coverage of psychology, health sciences, computer science, and social science research. In addition, arXiv will be searched to capture relevant grey literature and emerging empirical studies related to conversational generative AI that may not yet be indexed in traditional bibliographic databases.

To complement electronic database searches, manual screening of reference lists (snowballing) will be conducted for all included studies to identify additional relevant publications. Where necessary, corresponding authors may be contacted to clarify methodological details or obtain missing information relevant to psychological outcomes.

No clinical trial registries are specifically targeted, as the review does not focus on interventional clinical trials. Only publicly available sources will be used, and no proprietary or restricted datasets will be accessed.

Main outcome(s) The main outcomes of this systematic review are psychological effects and risks associated with conversational interaction with generative artificial intelligence systems. Primary outcomes include the presence, direction,

and characteristics of psychological responses reported in relation to human–AI dialogue.

Specifically, the review focuses on outcomes such as anthropomorphization of the AI system, overreliance or dependency on AI-generated responses, persuasive influence on users' attitudes, beliefs, or decision-making, and changes in psychological well-being, including emotional comfort, distress, reassurance, or confusion. Additional primary outcomes include indicators of problematic or excessive use, representational or cultural bias affecting user perception, and conversational harm in sensitive or crisis-related contexts.

Outcomes are considered regardless of whether they are measured through validated psychometric instruments, structured questionnaires, qualitative interviews, observational analysis, or systematic evaluation of interaction logs, provided that they are explicitly linked to user psychological impact. The timing of outcome assessment may range from immediate or short-term effects observed during or shortly after interaction, to longer-term effects reported following repeated or sustained conversational use.

Given the heterogeneity of study designs and outcome measures, the review does not prioritize a single effect size metric. Instead, outcomes are synthesized narratively and organized according to an operational four-factor framework to enable systematic comparison and interpretation of psychological risks across studies.

Additional outcome(s) Additional outcomes of this systematic review include contextual, design-related, and governance-relevant factors that help interpret and qualify the primary psychological outcomes. These secondary outcomes encompass characteristics of conversational AI systems that may moderate or amplify psychological effects, such as transparency of system limitations, uncertainty signaling, verification prompts, boundary-setting practices, and the availability of safety or referral mechanisms.

Further outcomes include user-related and contextual variables reported in the literature, such as prior experience with artificial intelligence, perceived trustworthiness, perceived social presence, cultural or linguistic adequacy, and differential effects across usage contexts (e.g., clinical support, education, social interaction, or everyday information seeking). Where reported, disparities or differential impacts across demographic or cultural groups are also considered.

In addition, the review captures methodological features relevant to the interpretation of findings, including the duration of exposure, number of conversational turns, longitudinal versus single-session designs, and the types of instruments or analytical approaches used to assess psychological effects. These additional outcomes support a more nuanced synthesis and facilitate the translation of findings into evaluative criteria and governance considerations for conversational generative AI systems.

Data management All records retrieved from the information sources were exported to reference management software for initial organization and duplicate removal. Following deduplication, records were imported into a systematic review management platform to support screening, data extraction, and traceability of decisions throughout the review process.

Study selection was conducted in two stages. First, titles and abstracts were independently screened to assess potential eligibility. Second, full-text articles were reviewed to confirm inclusion based on the predefined eligibility criteria. Disagreements at any stage were resolved through discussion and consensus among reviewers.

Data extraction was performed using structured and predefined extraction forms designed to capture study characteristics, participant information, interaction context, methodological features, and reported psychological outcomes. Extracted data were organized into standardized tables to facilitate synthesis and comparison across studies.

All data management procedures were documented to ensure transparency and reproducibility. No individual-level personal data were collected or stored, as the review relied exclusively on data reported in published studies. Data files and extraction tables were securely stored and maintained for auditability and potential future updates of the review.

Quality assessment / Risk of bias analysis The methodological quality and risk of bias of included studies were assessed using an approach tailored to the heterogeneity of study designs present in the literature. Quantitative studies were evaluated using appropriate established tools according to their design, while qualitative studies were assessed based on transparency, coherence, and

rigor of data collection and analysis. Across all studies, attention was given to clarity of research objectives, adequacy of study design, description of participants and interaction context, validity of outcome measures, and transparency in reporting. Given the diversity of methodologies and outcomes, no single aggregate risk-of-bias score was calculated; instead, quality considerations were incorporated narratively into the synthesis to contextualize findings and support cautious interpretation of psychological effects associated with conversational generative AI.

Strategy of data synthesis Data synthesis will be conducted using a narrative and thematic approach, given the heterogeneity of study designs, outcome measures, and interaction contexts. Quantitative meta-analysis is not planned due to the absence of homogeneous effect measures across studies. The synthesis will proceed as follows:

Extracted findings will be grouped according to the type of psychological effect reported (e.g., anthropomorphization, overreliance, persuasive influence, changes in well-being, conversational harm).

Results will be organized within an operational four-factor framework: Humanized Language, Personal Interaction, Social Interaction, and Psychological Cybersecurity.

Patterns related to interaction duration, conversational features, and context of use will be identified to examine how psychological effects emerge or intensify across studies.

Methodological quality and risk-of-bias considerations will be integrated narratively to contextualize the strength and limitations of the evidence.

This structured narrative synthesis will allow comparison across diverse study designs and support the interpretation of psychological risks associated with conversational generative AI in a systematic and transparent manner.

Subgroup analysis Subgroup analyses will be conducted to explore how psychological effects associated with conversational generative artificial intelligence systems may vary across different contexts, user characteristics, and interaction conditions, when such information is reported in the primary studies. These analyses are exploratory in nature and aim to enhance interpretation rather than generate confirmatory causal claims.

Planned subgroup analyses include comparisons across usage contexts, such as clinical or mental health support, educational settings, social interaction, and everyday informational use, in order to examine whether specific psychological effects are more prevalent or pronounced in certain environments. Where data permit, results will also be examined according to interaction characteristics, including duration of exposure (single-session versus sustained or repeated interactions), number of conversational turns, and the presence of personalization or memory features.

Additional subgroup analyses may consider user-related factors reported in the literature, such as age group (e.g., adolescents versus adults), prior experience with artificial intelligence systems, and reported vulnerability or sensitivity to persuasive or emotional content. When studies include cross-cultural or multilingual samples, differences related to cultural or linguistic context will be explored descriptively.

Subgroup findings will be synthesized narratively and interpreted with caution, taking into account methodological quality, sample size limitations, and potential reporting bias. No statistical tests for subgroup differences are planned. The purpose of these analyses is to identify patterns and contextual conditions under which psychological risks associated with conversational AI may be amplified or mitigated, thereby informing future research, system design considerations, and socio-technical governance strategies.

Sensitivity analysis Sensitivity analyses will be conducted to examine the robustness of the review findings and to assess the influence of methodological and contextual decisions on the interpretation of psychological effects associated with conversational generative artificial intelligence systems. Given the heterogeneity of study designs, outcome measures, and analytical approaches, sensitivity analyses will focus on qualitative and descriptive comparisons rather than statistical re-estimation of effect sizes.

First, sensitivity analyses will be performed by examining whether the exclusion of studies assessed as having lower methodological quality alters the overall patterns of psychological effects identified in the synthesis. Findings derived from studies with limited reporting transparency, unclear participant descriptions, or insufficient detail

regarding the interaction context will be compared with those from studies meeting higher quality standards to evaluate consistency of conclusions.

Second, sensitivity analyses will consider the impact of study design by comparing findings across experimental, observational, qualitative, and mixed-methods studies. This comparison will help determine whether specific psychological effects are consistently observed across methodological approaches or appear to be driven primarily by particular study types.

Third, sensitivity analyses will explore the influence of interaction characteristics, including duration of exposure and conversational depth. Where possible, findings from single-session or short-term interactions will be compared with those from studies examining repeated or sustained use, in order to assess whether reported psychological risks are sensitive to interaction length and continuity.

Finally, sensitivity analyses will examine the effect of excluding studies based on language, publication type, or data source, such as the removal of preprints or grey literature, to evaluate whether inclusion of these sources meaningfully influences the synthesized conclusions. All sensitivity analyses will be interpreted cautiously and reported narratively. The purpose of these analyses is not to reweight evidence statistically, but to enhance transparency, assess robustness, and support credible interpretation of psychological risks associated with conversational generative artificial intelligence. This approach supports methodological clarity while acknowledging uncertainty and complexity inherent in interdisciplinary research on human–AI interaction and evolving conversational technologies across diverse empirical contexts and populations over time period.

Language restriction The search will be limited to studies published in English and Spanish. No additional language restrictions will be applied.

Country(ies) involved Chile. The review is conducted by authors affiliated with Chilean institutions, with international collaboration reflected in the included literature.

Other relevant information This systematic review forms part of a broader academic research agenda focused on the psychological and socio-technical implications of conversational generative artificial intelligence. The review is intended to inform the development of evaluative frameworks and governance-oriented approaches for assessing psychological risks associated with human–AI interaction, rather than to provide clinical or therapeutic recommendations.

The protocol was developed following established methodological guidance for systematic reviews and is aligned with PRISMA principles. Any deviations from the registered protocol, should they occur, will be transparently documented and justified in the final review report. Amendments to the protocol will be reported in the INPLASY record if necessary.

No ethical approval is required for this review, as it is based exclusively on the analysis of previously published studies and does not involve the collection of new data from human participants. The findings of the review are expected to contribute to interdisciplinary discussions spanning psychology, human–computer interaction, artificial intelligence governance, and digital risk assessment.

Keywords Conversational artificial intelligence; generative AI; psychological effects; human–AI interaction; anthropomorphism; overreliance; digital risk.

Dissemination plans The findings of this systematic review will be disseminated through multiple academic and professional channels to reach interdisciplinary audiences. Primary dissemination will occur through submission of a full manuscript to a peer-reviewed international journal focused on psychology, human–computer interaction, or artificial intelligence governance. Results will also be presented at academic conferences and seminars related to digital psychology, AI ethics, and socio-technical risk.

In addition, summarized findings will be shared through academic networks, institutional repositories, and professional platforms to support knowledge transfer beyond traditional publications. The review is expected to inform future empirical research, methodological development, and policy-oriented discussions concerning psychological risks of conversational generative AI. No individual-level data will be disseminated. All reporting will adhere to established transparency and reporting standards, including PRISMA guidelines. Any protocol amendments will be clearly documented in subsequent publications or presentations to ensure traceability and methodological integrity. This strategy ensures responsible dissemination aligned with academic

rigor, public relevance, and long-term research impact globally.

Contributions of each author

Author 1 - Cristian Medina-Rojas - Author 1 conceived the study, designed the review protocol, conducted the literature search, performed data screening and extraction, led the synthesis and analysis, and drafted the manuscript.
Email: cmedinar@doctoradoia.cl
Author 2 - Pedro Pinacho-Davidson - Author 2 contributed to protocol development, supported study selection and data extraction, critically reviewed the methodological approach, and provided substantive feedback on the synthesis and interpretation of results.
Email: ppinacho@udec.cl
Author 3 - Ivan Pau de la Cruz - Author 3 contributed to conceptual refinement, advised on psychological and socio-technical interpretation, reviewed the manuscript for intellectual content, and approved the final version of the protocol.
Email: ivan.pau@upm.es
Author 4 - Pedro Salcedo-Lagos - Author 4 contributed to protocol development, supported study selection and data extraction, critically reviewed the methodological approach, and provided substantive feedback on the synthesis and interpretation of the results.
Email: psalcedo@udec.cl