INFOTECH
Journal of Applied Sciences, Informatics, Computing and Knowledge Society

Volume 4, No. 1, January/June, 2023

Revolutionizing Mental Health Care: The Impact of AI Chatbots

Olivia Nelson[1]

[1] University of the South Pacific - Fiji, olivianelson83@outlook.com, https://orcid.org/0009-0006-7594-2029

ABSTRACT

Since 2022, AI chatbots have attracted attention for their use of big data, NLP, and ML algorithms. These entities enhance capabilities, increase productivity, and provide guidance across diverse domains. Human–Artificial Intelligence (HAI) integrates human values into AI, addressing limitations and improving effectiveness. In mental health, AI and ML are integrated into digital solutions to tackle access, stigma, and cost challenges. However, ethical and legal uncertainties surround these technologies. This literature review explores AI chatbots' potential in transforming digital mental health, emphasizing the need for ethical, responsible, and trustworthy AI algorithms. Three key research questions focus on the impact of AI chatbots on technology integration, the balance between benefits and drawbacks, and addressing bias in AI applications. Methodologically, the review involves thorough searches of databases using keywords related to AI chatbots and digital mental health. Rigorously selected peer-reviewed articles and media sources contribute to a comprehensive analysis. In summary, while AI chatbots hold promise in reshaping digital mental health, navigating ethical and practical challenges is crucial. Incorporating HAI principles, responsible regulations, and scoping reviews are essential for maximizing benefits and minimizing risks, with collaborative strategies and contemporary education promoting responsible use and mitigating biases in the digital mental health landscape.

Keywords: *AI-driven mental health; HAI principles; Responsible AI; Digital mental health landscape*

Received:
24 February, 2023

Accepted for publication:
07 April, 2023

INFOTECH
Journal of Applied Sciences, Informatics, Computing and Knowledge Society

Volume 4, No. 1, January/June, 2023

Transformando la Atención de la Salud Mental: El Impacto de los Chatbots de IA

RESUMEN

Desde 2022, los chatbots de inteligencia artificial (IA) han llamado la atención por su uso de grandes conjuntos de datos, procesamiento de lenguaje natural (NLP) y algoritmos de aprendizaje automático (ML). Estas entidades mejoran las capacidades, aumentan la productividad y proporcionan orientación en diversos ámbitos. La Inteligencia Artificial Humana (HAI) integra valores humanos en la IA, abordando limitaciones y mejorando la efectividad. En la salud mental, la IA y el ML se integran en soluciones digitales para abordar desafíos de acceso, estigma y costos. Sin embargo, existen incertidumbres éticas y legales en torno a estas tecnologías. Esta revisión de la literatura explora el potencial de los chatbots de IA en transformar la salud mental digital, haciendo hincapié en la necesidad de algoritmos de IA éticos, responsables y confiables. Tres preguntas clave de investigación se centran en el impacto de los chatbots de IA en la integración tecnológica, el equilibrio entre beneficios y desventajas, y abordar el sesgo en las aplicaciones de IA. Metodológicamente, la revisión implica búsquedas exhaustivas en bases de datos utilizando palabras clave relacionadas con chatbots de IA y salud mental digital. Artículos revisados por pares y fuentes de medios rigurosamente seleccionadas contribuyen a un análisis integral. En resumen, aunque los chatbots de IA prometen remodelar la salud mental digital, es crucial abordar los desafíos éticos y prácticos. La incorporación de principios HAI, regulaciones responsables y revisiones de alcance son esenciales para maximizar beneficios y minimizar riesgos, con estrategias colaborativas y educación contemporánea que promueven el uso responsable y mitigan sesgos en el panorama de la salud mental digital.

Palabras clave: *Salud mental impulsada por IA; Principios HAI; IA responsable; Panorama de salud mental digital*

INTRODUCTION

AI chatbots, characterized as intelligent conversational computer systems, have evolved to think, learn, and execute tasks collaboratively with or independently of humans. Employing big data, natural language processing (NLP), and machine learning (ML) algorithms, these conversational agents, also known as generative AI utilizing large language models, have seen significant advancements over the past 15 years in robotics, ML, AI models, and NLP. Their prominence heightened with the launch of ChatGPT in November 2022 [3].

While AI chatbots present opportunities for insightful responses beyond human capacity, they often lack a personalized and empathetic touch. To address these limitations, the concept of human–artificial intelligence (HAI) is proposed, emphasizing a collaborative approach where humans and AI leverage each other's strengths for efficient, safer, sustainable, and enjoyable work and lives. This aligns with the Center of Humane Technology's efforts to integrate human values, such as empathy, compassion, and responsibility, into AI [4].

Mental health, a critical global issue impacting millions worldwide, faces challenges like limited access to services, stigma, and cost barriers. In Australia, for instance, around 20% of adults experience a mental disorder, rising to 44% over a lifetime [6]. The economic impact is substantial, with billions of dollars lost due to diminished health and reduced life expectancy [7]. Digital mental health solutions, targeting young people, employ technology for assessment, support, prevention, and treatment, incorporating AI and ML models for predicting mental illness and AI chatbots for psychological support [13–15]. However, ethical and legal uncertainties surround these tools.

This narrative literature review aims to illustrate the potential of AI chatbots in providing accessible digital mental health for diverse populations. It emphasizes the challenges posed by the imperative to develop ethical, responsible, and trustworthy AI algorithms.

METHODS

Adapting the approach outlined by Demiris et al. [16], this narrative literature review follows four steps: (1) Conducting a comprehensive search across various databases and search engines, (2) Identifying and incorporating pertinent keywords from relevant articles, (3) Reviewing abstracts and texts of

selected articles addressing the research aim, and (4) Documenting results by summarizing and synthesizing findings into the review.

Due to the heterogeneity of the topic and its evolving nature, a systematic review was not feasible. The interdisciplinary nature of digital mental health, spanning psychology, technology, and healthcare, further complicates matters, leading to diverse research approaches and methodologies. Consequently, a purposive selection of articles was chosen, adopting an educational approach to showcase how AI chatbots may impact digital mental health. This flexibility allowed for the inclusion of various perspectives and ideas, contributing to a more holistic understanding of the subject matter.

The retrieval of peer-reviewed journal articles, media pieces, and conference proceedings involved searches across computerized databases, targeted online investigations, and authoritative texts, guided by three research questions presented in the Editorial, "AI Chatbots: Threat or Opportunity?" [3]. These questions provided a framework for exploring the subject, acknowledging the ongoing evolution of our understanding of the science:

1. The emergence of AI chatbots is asserted to usher in a new era, presenting substantial advancements in integrating technology into people's lives and interactions. The investigation aims to determine the likelihood of this occurrence and identify where these impacts might be most prevalent and effective.
2. The research seeks to ascertain whether it is feasible to strike a balance in the impact of these technologies, minimizing potential harms while maximizing and sharing potential benefits.
3. With a growing body of evidence indicating bias and prejudice in the design and implementation of various AI applications, particularly algorithms, the study addresses strategies to counter and correct these issues.

The exploration involved scouring databases like Scopus, ScienceDirect, Sage, and the Association for Computing Machinery (ACM) Digital Library, in addition to leveraging search engines such as PubMed, Google Scholar, and IEEE Xplore. The search was conducted using terms like "AI chatbots" OR "generative artificial intelligence" OR "conversational agents" AND "digital mental health" OR "mental health care."

The criteria for selection were outlined as follows:

Inclusion Criteria

1. Studies available in peer-reviewed journals, media articles, and conference proceedings.
2. Studies published in the English language.
3. Studies released between 2010 and 2023.
4. Studies exploring the utilization of AI chatbots, generative artificial intelligence, or conversational agents in digital mental health or mental health care.
5. Studies presenting findings on the effectiveness of AI chatbots, generative artificial intelligence, or conversational agents in digital mental health or mental health care.

Exclusion Criteria

1. Studies not appearing in peer-reviewed journals, media articles, and conference proceedings.
2. Studies not published in the English language.
3. Studies published before 2010 or after 2023.
4. Studies not investigating the use of AI chatbots, generative artificial intelligence, or conversational agents in digital mental health or mental health care.
5. Studies not providing insights into the effectiveness of AI chatbots, generative artificial intelligence, or conversational agents in digital mental health or mental health care.

Boolean operators, namely AND and OR, were employed to merge search terms and streamline search outcomes. For instance, the use of the OR operator between "AI chatbots" and "generative artificial intelligence" retrieved articles containing either term. Similarly, the AND operator between "conversational agents" and "digital mental health" gathered articles featuring both terms. The integration of Boolean operators served to refine search results, making them more pertinent to the research question.

Exploration of relevant articles and their reference lists was conducted based on three criteria: (1) relevance to the guiding research questions, (2) exemplification of theoretical and empirical research and development, and (3) elucidation of issues and potential solutions. A best-evidence synthesis was applied to ensure a comprehensive, critical, and objective analysis of the current knowledge on the

topic. While this method demonstrates a systematic and transparent approach, minimizing bias by ensuring a thorough search and focusing on pertinent articles, it is essential to acknowledge that bias may still exist in the literature itself. Consequently, the selected articles were critically evaluated, with any potential limitations or biases within them acknowledged.

RESULTS

The Impact of AI Chatbots on Technology Integration

Research Inquiry 1: The emergence of AI chatbots is posited as a transformative force, heralding substantial progress in the integration of technology into human lives and interactions. To what extent is this assertion plausible, and if so, in what domains will these impacts be most pronounced and effective [3]?

The application of AI chatbots holds the potential to catalyze noteworthy advancements with far-reaching effects on diverse facets of human existence and interaction [17]. Especially in situations where direct human-to-human engagement is impractical or undesirable [18], AI chatbots can provide valuable services in customer support, healthcare and mental health aid, education and e-learning, personal assistance and productivity, language translation and communication, as well as social companionship and entertainment [19,20]. Given the broad spectrum of applications and the extensive body of empirical literature, concentrating on a specific domain appears judicious.

Mental health care serves as an illustrative case, with AI chatbots having garnered recognition as a valuable asset in this realm for more than a decade [21]. Encouraging clinical outcomes have been observed, ranging from AI chatbots delivering pertinent and continuously accessible support [22,23] to addressing conditions such as depression in adults [24], anxiety in university students [25,26], and symptoms of attention-deficit/hyperactivity disorder in adults [27].

AI chatbots have the potential to surmount obstacles in seeking help for mental health issues by offering personalized, accessible, cost-effective, and stigma-free assistance. This facilitates early intervention and contributes valuable insights for both research and policy formulation [28–31]. Their efficacy is particularly pronounced in tasks such as monitoring, communication, memory support, screening, and diagnosis, with a focus on understanding a patient's emotional state and assisting in the analysis of

extensive datasets. For example, algorithms can discern patterns and trends that may elude human analysts. By scrutinizing a patient's medical history, genetic data, and other pertinent factors, algorithms could formulate customized symptom assessments and treatment recommendations that account for the individual's unique needs and circumstances.

However, the opportunities presented by AI chatbots should be considered alongside challenges such as a lack of human connection, dependence on technology, the accuracy and reliability of information, ethical and privacy considerations, as well as concerns related to misdiagnosis and limited understanding [28–31].

A comprehensive assessment of mental health chatbots in 2023 identified 10 applications available in the market catering to diverse mental health concerns (e.g., anxiety and depression) and tailored for various user groups (e.g., rural residents, shift workers, students, veterans, and adolescents). These applications served a range of purposes, such as enhancing social or job interviewing skills [18]. The overview specifically focused on AI chatbots due to their accessibility, affordability, and convenient provision of social and psychological support. However, it highlighted concerns about vulnerable users potentially overestimating the benefits and facing risks, particularly in crisis situations, as AI chatbots were perceived as lacking the capability to identify such scenarios. The inadequacy of semantics was identified as a critical issue, with AI chatbots struggling to comprehend the context of users' words and responding ineffectively or not at all.

A key challenge lies in users potentially not distinguishing between humans and humanlike chatbots. Addressing these limitations requires a human-centric approach, with education playing a pivotal role in fostering effective collaboration to develop sustainable solutions [32]. Users and practitioners alike need guidance on the proper utilization of AI chatbots, a need comparable to what is generally required for digital mental health platforms and interventions [33].

The interdisciplinary nature of mental health, involving psychology, psychiatry, AI, and healthcare, along with the contributions of educators, policymakers, computer scientists, and technology developers, presents significant hurdles to overcome for realizing overall benefits [34]. Mental health professionals and policymakers are instrumental in unlocking the potential of AI chatbots as valuable

tools in the intelligent system toolbox. However, it is evident that the impetus for change may be driven most effectively by graduate students and research scientists, given their willingness and ability to collaborate effectively with computer scientists and technology developers.

AI chatbots hold promise as complementary tools rather than replacements for human mental health professionals [18,20]. A review of digital mental health interventions in 2021 speculated that AI chatbots could assist mental health professionals in meeting the overwhelming demand for services [34]. A systematic review and meta-analysis of randomized controlled trials in 2023 confirmed the acceptability of AI chatbots for a wide range of mental health problems [35]. For instance, an RCT demonstrated the feasibility, engagement, and effectiveness of a fully automated conversational agent, Woebot, in delivering cognitive-behavioral therapy (CBT) for anxiety and depression in young adults [25]. Woebot [36] and Wysa [37] show promise in establishing therapeutic bonds with users.

Despite the feasibility of AI chatbots as engaging and acceptable therapeutic delivery tools, additional studies are needed to understand the factors contributing to a digital therapeutic alliance [36,37] and to minimize misunderstandings [38]. While mental health chatbots exhibit lower attrition rates compared to other digital interventions [24,39], attention is required to address dropout rates and clarify their efficacy for specific disorders [40]. Some reviews highlight the high potential of AI chatbots in identifying patients at risk of suicide [41–43] and in triage and treatment development through the integration of natural language processing (NLP) into social media in real-time [44–46].

Advancements in Generative Pre-Trained Transformer (GPT) programs, exemplified by ChatGPT 4, signal the potential use of AI chatbots in suicide prevention [47]. However, there is a critical necessity to deepen our comprehension of AI chatbot limitations, including negative sentiment, constrictive thinking, idioms, hallucinations, and logical fallacies. A study focusing on messages related to suicidal thoughts sought insights from the arrangement of words, sentiment, and reasoning [48]. While addressing AI chatbot hallucinations and fallacies requires human intervention, it is feasible to detect idioms, negative sentiment, and constrictive language using off-the-shelf algorithms and publicly available data. Safety concerns were highlighted when a chatbot named Eliza was implicated by a Belgian man's widow for her husband's suicide [49].

Qualitative studies are imperative to mitigate poor semantics and errors while fostering trust in AI chatbots. Thematic analysis based on retrospective data is essential to identify common message themes sent to mental health chatbots, enhancing their effectiveness as a support source. AI chatbots can contribute to addressing problem areas through Natural Language Processing (NLP) for sentiment analysis—an expeditious and effective qualitative data analysis method—to aid in comprehending multidimensional online feedback.

Recommendations for identifying and assessing the impact of AI chatbots include:

- Conducting qualitative studies employing AI chatbots to illustrate their contributions to accessibility, engagement, and effectiveness by (1) identifying user needs, (2) understanding barriers to usage, (3) evaluating user experience and AI chatbot impact, and (4) integrating human–AI approaches to address challenges.
- Contributing to empirical evidence through longitudinal studies and Randomized Controlled Trials (RCTs) to determine the mental health conditions and populations for which AI chatbots may be recommended.
- Establishing a practical attrition prediction method to identify individuals at high risk of dropping out, utilizing advanced machine learning models (e.g., deep neural networks) by leveraging analyses of feature sets (e.g., baseline user characteristics, self-reported user context, AI chatbot feedback, passively detected user behavior, and clinical functioning of users).

Striking a Balance between the Advantages and Drawbacks of AI Chatbots

Providing a comprehensive global assessment proves challenging due to the absence of widely collaborative international standards and the varied applications of AI chatbots. Current investments in AI research, education, societal adaptation, innovation, employment opportunities, and job creation seem inadequate when considering the scale of imminent changes.

Given the novelty and intricacy of AI in mental health, a timely focus on cutting-edge education, such as specialized university courses in digital mental health and informatics utilizing regularly updated textbooks and modules, is crucial. The objective should be to foster discerning skills and critical thinking across diverse subjects, paving the way for innovative benefits in mental health care and the

AI technology industry while also mitigating the escalating costs associated with mental illness. Despite the prominence of AI chatbots, they have not fully realized their potential in assisting with mental health problems among digital users, predominantly young people [18].

While quality, effective, and user-friendly chatbots like Woebot and Wysa are available to support mental health [36,37], additional studies are essential to provide evidence across a broader spectrum of mental health disorders and symptoms. Moreover, the predominant driving force behind development is rooted in technology rather than being led by mental health professionals with technological expertise. Communication style and methodologies differences between technology and mental health care researchers, such as pattern-based versus hypothesis-derived approaches, have constrained the integration of these two realms. Another obstacle is the limited opportunities for high-level researchers capable of comprehending and implementing hybrid methods.

Nonetheless, there is considerable potential for mental health care to serve as an example where AI chatbots can contribute to providing cost-effective solutions for a diverse range of users and objectives [21–24,36–38]. Mental health care professionals may need to embrace AI chatbots to enhance productivity [50]. Efforts should also be directed toward broadening the definition of productivity if substantial advancements in integrating technology into people's lives and interactions are to be achieved. For instance, accurately measuring AI chatbots' contributions to the economy and people's health poses a challenge. While there might be gains to the gross domestic product (GDP) of developed countries, job losses due to AI disruption are also plausible. Besides productivity, affordability, and accessibility, policies should consider mental health and human capital.

Assessing the impact of AI chatbots on productivity necessitates consideration within the frameworks of national and international economics, standards, and regulations. It is evident that not all governments share the same principles, and the digital divide raises concerns about further marginalizing the underserved and unserved populations [34]. Consequently, productivity and humanity must be weighed against global risks such as war and the costs associated with the physical effects of climate change [50]. As some governments heavily invest in defense, decarbonization of heavy industries, and transitions among energy systems, there will be competing demands for investment in

AI technologies. Meanwhile, the emergence of ChatGPT serves as an example of stakeholders grappling with the challenges posed by the rapid pace of technological development.

The methodology behind the Productivity Commission's projections of AI contributing to a substantial boost in the Australian economy, predicting a potential increase in GDP ranging between 66.67% and 266.67% over the next decade, remains unclear [7]. In 2023, the Australian Government outlined a 40-year outlook on intergenerational equity, forecasting increased financial burdens on younger generations [51]. This raises questions about how countries, including Australia, navigate and optimize significant shifts in their national economy while effectively integrating the impacts of AI technologies. Examining the context of mental health care in Australia underscores the importance of evaluating AI's alignment with existing safety and quality structures before delving into its economic potential. The National Standards in mental health services in Australia establish a framework for safety and quality in both hospital and community services, primarily regulating the practices of mental health professionals [52]. However, with a surge in demand and constrained supply in mental health care, especially exacerbated by the COVID-19 pandemic [53,54], digital mental health emerged as a crucial service gap filler. This led to the development of the National Safety and Quality Digital Mental Health Standards in 2020, aiming to enhance the safety and quality of digital mental health service provision [55]. Yet, mental health professionals and policymakers grapple with the opportunities and challenges posed by AI [56]. For instance, prompt engineering techniques employed with ChatGPT to bypass content filters in social media present potential risks, including harm and exploitation of vulnerability.

In 2018, the Australian Government adopted a voluntary ethics framework for "responsible" AI to guide businesses and governments in designing, developing, and implementing AI responsibly [57]. Despite this, mainstream AI chatbots are predominantly developed in the US. Australia, along with various other countries, is either seeking input or planning on regulating AI chatbots [58]. The European Union implemented the Digital Services Act and the Digital Market Act to create a safer digital space, protecting users' fundamental rights and establishing a level playing field for businesses [59]. Ensuring that AI algorithms are developed and trained using diverse and representative datasets, and rigorously validating and verifying insights generated by AI through human experts, is deemed essential. OpenAI,

the owners of ChatGPT, proposed proactive risk management for these "frontier AI" models, including pre-deployment risk assessments, external scrutiny of model behavior, and ongoing monitoring post deployment [60].

The responsibility for transparency in AI use, safeguarding privacy and confidentiality, and responsible utilization for optimal performance ultimately lies with users [61]. For example, recruits using AI chatbots to fulfill their duties may raise questions about the value of collaborative work if trust cannot be established and maintained [62]. An inherent challenge with AI chatbots is their novelty as a technology, carrying the potential to become fundamentally pervasive and pose cybersecurity risks due to their ability to generate various malicious codes and algorithms that can disrupt infrastructure or financial systems [63].

The utilization of AI in mental health research has been widely acknowledged for its potential to provide valuable insights and enhance outcomes for individuals with mental health disorders [64,65]. However, it is crucial to carefully categorize and oversee "high" risks, giving precedence to ethical considerations at every stage. The growing use of AI chatbots for mental health and crisis support necessitates increased attention and education among stakeholders to effectively harness these tools [18,66]. Initiatives like fair-aware AI in digital mental health have been advocated to promote diversity and inclusion [67], while explainable AI has been proposed as a tool to establish transparency and trust between users and practitioners [68].

The notion of Human–Artificial Intelligence (HAI) emerges as a supplementary concept in the evolution of AI systems, advocating for a collaborative approach where multiple AI models collaborate with human input to produce recommendations and predictions, moving away from a reliance on a singular algorithm. The subsequent phase involves determining optimal combinations of humans and AI chatbots for diverse tasks in research, practice, and policy realms [69]. However, a comprehensive strategy towards AI technologies is imperative, particularly with respect to broad-ranging regulation.

Australia's AI Ethics Principles [70] posit that regulation plays a pivotal role in achieving safer, more reliable, and fairer outcomes for all Australians. It serves to mitigate the risk of adverse impacts on individuals affected by AI applications and encourages businesses and governments to adhere to the

highest ethical standards in the design, development, and implementation of AI. A subsequent position statement on generative AI underscores that regulation can effectively address concerns about potential harms, encompassing algorithmic bias, errors, the proliferation of misinformation, inappropriate content, and the creation of deepfakes [71].

Regulation, achieved through mechanisms like transparency, accountability, and risk mitigation strategies, ensures the responsible and ethical use of AI [72]. Furthermore, it has the potential to cultivate public trust in AI technologies by aligning their development and utilization with societal values and expectations [73]. This approach paves the way for the widespread adoption of AI technologies, allowing society to fully harness their potential benefits.

The regulatory framework for AI should encompass defining the parameters of "unsafe" AI and identifying the facets of AI subject to regulation [74]. This necessitates a clear comprehension of the anticipated global risks and benefits of AI technologies, coupled with an understanding of public trust and acceptance of AI systems. While individuals in Western countries tend to exhibit caution towards AI, expressing reservations about the benefits outweighing the risks, those in emerging economies (such as Brazil, India, China, and South Africa), as well as young, university-educated individuals and those in managerial roles, display greater trust and enthusiasm for AI [75].

Overly stringent regulations could impede innovation and hinder the development of AI technologies [76]. Therefore, effective regulation necessitates international cooperation. Without a global consensus, companies might relocate their AI development activities to less regulated jurisdictions, leading to a regulatory race to the bottom. It is imperative to secure AI models and associated systems using industry-standard security protocols. Regular updates and patches should be applied to address any identified vulnerabilities in AI models and systems.

Recommendations for supervising and promoting the responsible utilization of AI applications encompass the following:

- **Invest in Research:** Allocate resources for research to evaluate the effectiveness and potential risks associated with AI applications. Develop systems dedicated to monitoring and auditing AI systems, enabling the identification of unusual or suspicious activities.

- **Enforce Stringent Safety Measures:** Implement rigorous safety protocols to safeguard against potential risks. Enact robust regulations and collaborative standards to ensure the responsible and ethical use of AI technologies.
- **Validate Human–Artificial Intelligence (HAI) Models:** Validate the effectiveness of models combining AI chatbots with human experts (HAI models) in the realms of research, practice, and policy. This integration aims to optimize mental health care assistance by leveraging the strengths of both AI and human input.

Addressing Bias and Prejudice in AI Applications

The World Health Organization has issued a cautionary statement, emphasizing the necessity for careful consideration when employing generative AI in healthcare [77]. The efficacy of AI algorithms hinges on the quality of the training data, and biases within this data can lead to skewed outcomes [78]. Furthermore, the application of AI in mental health care introduces significant risks and ethical considerations [79], including concerns related to security, bias, and privacy, especially regarding the storage and utilization of sensitive medical and personal data [80].

On a broader scale, there are applications of AI and automated decision-making deemed "high-risk," prompting warnings about potential harms, such as the creation of deepfakes and algorithmic bias [81]. Concerns also exist about AI perpetuating biases or reinforcing limited perspectives [82,83], along with the potential for automation to replace humans in certain roles [84]. Nevertheless, AI can be deployed to counter disinformation and enhance the accuracy and reliability of reporting [85,86]. The challenge lies in defining and determining what qualifies as "unsafe" AI. Numerous Australian science experts have advocated for the implementation of rigorous safety measures, robust regulations, and standards for such "unsafe" AI [76]. It is apparent that proactive measures to mitigate the risks associated with high-risk AI need to be promptly pursued to avoid impeding progress in AI.

Generative AI is actively utilized in the media for personalized and targeted advertising, content creation and curation automation, and analysis of audience behavior and preferences [87,88]. Tools like ChatGPT [87], when combined with social media, may contribute to misinformation or disinformation, potentially leading to the deprioritization of legitimate news outlets in favor of spam and false or

manipulative user-uploaded content [87]. Biases and errors in generative AI [67,87] highlight the questionable nature of existing information assessment guidelines, particularly concerning evidence credibility, source transparency, and limitation acknowledgment. Generative AI underscores the need for new guidelines prioritizing ethics, fairness, privacy, and transparency [76], while also recognizing the intellectual property rights of human creators and organizations [89]. This situation may be exacerbated by potentially anticompetitive practices employed by dominant technology platforms such as Google and Meta [88].

There is an imperative to address and rectify AI applications that contribute to the perpetuation of bias, harassment, and marginalization, as well as the erosion of critical thinking and independent thought. AI chatbots present innovative opportunities to respond to calls for the identification and moderation of fake news [90] and the transparent regulation of social media platforms [91–93]. For instance, an examination of YouTube's impact on loneliness and mental health revealed that its recommendation algorithms might unintentionally reinforce existing beliefs and biases, disseminate misinformation and disinformation, and facilitate the spread of unhelpful or harmful content [46]. Nevertheless, the review also acknowledged that YouTube can have positive effects on loneliness, anxiety, and depression when users actively engage with the platform for education, social connection, and emotional support.

Biased and prejudiced AI applications can be addressed and rectified through education and research, with the support of AI chatbots [94]. However, human researchers or experts with a deep understanding of the history and context of research problems may be required to guide and supervise AI chatbots in generating solutions. For example, YouTube's recommendation algorithm, introduced in 2005 and continually evolving, has undergone changes aimed at prioritizing metrics such as shares, likes, and dislikes to enhance safety and remove harmful content [95]. The emergence of AI chatbots underscores the need for ongoing adaptation through legislation, the establishment of ethical values [94], and improvements to existing AI systems [46].

YouTube has implemented various initiatives, including mental health policies, algorithm adjustments, content moderation, psychoeducation for content creators and users, resource panels for mental health and crisis support, warnings for self-harm and suicide content, and parental controls [96]. Despite

YouTube's efforts to promptly remove offensive content, the algorithm's operation can inadvertently create filter bubbles and echo chambers, exposing users to content that reinforces their existing beliefs and biases [98]. This phenomenon can lead to polarization and misinformation, negatively impacting mental health. Enhanced algorithms are advocated to detect bias, rectify errors, and moderate how videos appear in watchlists, guiding users toward safe, well-informed, and inclusive content and referring them to mental health and crisis resource panels—with AI chatbots playing a supportive role [46].

Nevertheless, problematic social media use affects one in three young individuals in Australia, extending beyond YouTube to platforms like Facebook, Twitter, Snapchat, Instagram, and TikTok [99]. Cyberbullying is a prevalent issue across these platforms [100]. Numerous studies have identified a clear association between heavy social media use and an increased risk of depression, anxiety, loneliness, self-harm, and suicidal thoughts [101–103]. While there is a scarcity of psychological studies on TikTok [104], a causal study across American colleges found that access to Facebook resulted in a 7% increase in severe depression and a 20% increase in anxiety disorder [102]. This significant correlation between Facebook's presence and a decline in mental health among young people is alarming, especially considering the substantial rise (57%) in suicides among Americans aged 10–24 between 2007 and 2017 following the advent of Facebook in 2004 [105].

In 2018, significant data breaches and the use of "psychological warfare tools" on Facebook came to light through the Cambridge Analytica files [106]. Following calls for ethical, secure, and private data usage, Australia took the lead in global social media regulation with the implementation of the Online Safety Act in 2021. Public hearings revealed potential harm and safety issues associated with Facebook's algorithms. However, in 2022, the Australian government and the tech industry recognized that an outdated classification system hindered the establishment of new codes for regulating online content [108]. In 2023, Australia expressed interest in adopting risk-based classification systems for AI chatbots, similar to those being developed in Canada and the EU [109].

Advancements in AI chatbots, predictive models, and virtual assistants enable the integration of multiple models with human expert input to address mental health challenges, enhance suicide

prevention, improve access to care, and reduce barriers to seeking help. These tools utilize natural language processing (NLP) and machine learning (ML) to analyze mental health data, understand individual needs, and offer personalized support. A theoretical framework proposes the creation of an adaptive Social Media Virtual Companion (SMVC) to educate and support adolescent students in social media interactions, aiming to achieve collective well-being [110]. This SMVC framework exemplifies the design of social media systems and embedded educational interventions through Human-AI Interaction (HAI), combining automatic processing with educator/expert intervention.

HAI mental health strategies are deemed valuable for designing a responsible social media system in educational settings. An adaptive SMVC, integrated with solutions like Viable for sentiment analysis and DataMinr for monitoring and analyzing social media, can learn from HAI feedback and recent data to adjust recommendations. However, it's crucial to acknowledge that AI-generated sentiment can influence the emotional language in human conversations, potentially impacting social relationships. Randomized experiments have shown that algorithmic recommendation systems, such as those in AI chatbots like ChatGPT, alter how people socially interact and perceive each other, often leading to more negative evaluations. Therefore, educators should proactively and transparently promote the use of AI chatbots to avoid negative perceptions. Users may need guidance on being discerning and critical of the information provided by AI chatbots, effectively leveraging these tools to solve complex problems in their studies, and using them cautiously for self-care in mental health, seeking assistance when necessary [111].

Recommendations for addressing and rectifying the shortcomings of AI applications include the following:

- Individuals in vulnerable situations should receive informed guidance on effectively self-managing their mental health when utilizing AI chatbots, enabling them to connect with appropriate resources and treatments.
- Improvements to social media mental health and crisis resource panels can be achieved by integrating AI chatbots that offer verified digital mental health and crisis services or referrals when necessary.

- Human-Artificial Intelligence (HAI) mental health strategies, particularly with the Social Media Virtual Companion (SMVC), should be explored as a cautious approach for navigating a safer, more responsible social media environment with humane, fair, and explainable system recommendations.

CONCLUSIONS

This narrative literature review delves into the diverse impacts of AI chatbots on various societal aspects, with a specific focus on their potential in the realm of mental health care. The review serves as a valuable resource for providing an overarching understanding of the subject, pinpointing gaps in existing literature, and posing new research inquiries. Through the amalgamation of both theoretical and empirical studies, this review delivers a comprehensive snapshot of the present state of AI chatbots in mental health care. The evidence presented underscores the potential of AI chatbots to revolutionize mental health support, offering accessibility, engagement, and effectiveness in aiding individuals and populations dealing with a broad spectrum of mental health concerns and objectives.

However, it is imperative to approach the implementation and regulation of AI chatbots with prudence and responsibility. The novelty of AI chatbots in mental health is evident in this narrative literature review, which provides examples of theoretical and empirical research that future studies can build upon. The development of AI chatbots presents opportunities for reaching underserved and unserved populations, blending care for well-served individuals, particularly in treating common disorders like anxiety and depression.

Yet, challenges persist in discerning the quality and efficacy of AI chatbots, necessitating further research to clarify these areas and the level of care required for crisis support. Human factors in human-computer interaction demand more empirical attention. While AI chatbots offer accessible and convenient support, addressing barriers in the help-seeking process, they exhibit limitations such as poor semantics, biases, and a need for qualitative studies to enhance user experience. It's crucial to view AI chatbots as supplementary tools, not replacements for human mental health professionals. Despite their potential, empirical evidence and advocacy are necessary to discern their quality, usability, effectiveness, uses, and the populations that would benefit.

The call for regulation and responsible AI use is emphasized due to potential biases, privacy concerns, and the propagation of misinformation. This underscores the significance of international collaboration in establishing standards and regulations to ensure ethical and transparent AI technology use. Striking a delicate balance between innovation and regulation is paramount to prevent stifling progress while guarding against potential harm.

BIBLIOGRAPHICAL REFERENCES

Team Capacity. The Complete Guide to AI Chatbots: The Future of AI and Automation. 2023. Available online: https://capacity.com/learn/ai-chatbots/ (accessed on 19 August 2023).

Caldarini, G.; Jaf, S.; McGarry, K. A Literature Survey of Recent Advances in Chatbots. *Information* **2022**, *13*, 41. [CrossRef]

Bryant, A. AI Chatbots: Threat or Opportunity? *Informatics* **2023**, *10*, 49. [CrossRef]

The Center for Humane Technology. Align Technology with Humanity's Best Interests. 2023. Available online: https://www. humanetech.com/ (accessed on 19 August 2023).

World Health Organization. Mental Health. 2023. Available online: https://www.who.int/health-topics/mental-health#tab= tab_1 (accessed on 19 August 2023).

Australian Bureau of Statistics. National Study of Mental Health and Wellbeing. 2021. Available online: https://www.abs.gov.au/statistics/health/mental-health/national-study-mental-health-and-wellbeing/latest-release (accessed on 19 August 2023).

Australian Productivity Commission. Mental Health. 2020. Available online: https://www.pc.gov.au/inquiries/completed/mental-health#report (accessed on 19 August 2023).

Queensland Brain Institute. Life Expectancy Mapped for People with Mental Disorders. 2019. Available online: https://qbi.uq. edu.au/article/2019/10/life-expectancy-mapped-people-mental-disorders (accessed on 19 August 2023).

Clement, S.; Schauman, O.; Graham, T.; Maggioni, F.; Evans-Lacko, S.; Bezborodovs, N.; Morgan, C.; Rüsch, N.; Brown, J.S.L.; Thornicroft, G. What is the impact of mental health-related stigma

on help-seeking? A systematic review of quantitative and qualitative studies. *Psychol. Med.* **2015**, *45*, 11–27. [CrossRef]

Oexle, N.; Müller, M.; Kawohl, W.; Xu, Z.; Viering, S.; Wyss, C.; Vetter, S.; Rüsch, N. Self-stigma as a barrier to recovery: A longitudinal study. *Eur. Arch. Psychiatry Clin. Neurosci.* **2017**, *268*, 209–212. [CrossRef]

Australian Institute of Health and Welfare. Mental Health: Prevalence and Impact. 2022. Available online: https://www.aihw.gov.au/reports/mental-health-services/mental-health (accessed on 19 August 2023).

U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration. Key Substance Use and Mental Health Indicators in the United States: Results from the 2018 National Survey on Drug Use and Health. 2018. Available online: https://www.samhsa.gov/data/sites/default/files/cbhsq-reports/NSDUHDetailedTabs2018R2/NSDUHDetTabsSect8 pe2018.htm#tab8-28a (accessed on 19 August 2023).

Wies, B.; Landers, C.; Ienca, M. Digital Mental Health for Young People: A Scoping Review of Ethical Promises and Challenges.

Front. Digit. Health **2021**, *3*, 697072. [CrossRef]

Iyortsuun, N.K.; Kim, S.-H.; Jhon, M.; Yang, H.-J.; Pant, S. A Review of Machine Learning and Deep Learning Approaches on Mental Health Diagnosis. *Healthcare* **2023**, *11*, 285. [CrossRef]

Andreou, A. Generative AI Could Help Solve the U.S. Mental Health Crisis. Psychology Today. Available online: https://www.psychologytoday.com/au/blog/the-doctor-of-the-future/202303/generative-ai-could-help-solve-the-us-mental-health-crisis (accessed on 19 August 2023).

Demiris, G.; Oliver, D.P.; Washington, K.T. The Foundations of Behavioral Intervention Research in Hospice and Palliative Care. In *Behavioral Intervention Research in Hospice and Palliative Care*; Academic Press: Cambridge, MA, USA, 2019; pp. 17–25. [CrossRef]

Adamopoulou, E.; Moussiades, L. Chatbots: History, technology, and applications. *Mach. Learn. Appl.* **2020**, *2*, 100006. [CrossRef]

Haque, M.D.R.; Rubya, S. An Overview of Chatbot-Based Mobile Mental Health Apps: Insights from App Description and User Reviews. *JMIR mHealth uHealth* **2023**, *11*, e44838. [CrossRef] [PubMed]

Denecke, K.; Abd-Alrazaq, A.; Househ, M. Artificial Intelligence for Chatbots in Mental Health: Opportunities and Challenges. In *Multiple Perspectives on Artificial Intelligence in Healthcare: Opportunities and Challenges*; Lecture Notes in Bioengineering; Springer: Berlin/Heidelberg, Germany, 2021; pp. 115–128. [CrossRef]

Rizvi, M. AI Chatbots Revolutionize Depression Management and Mental Health Support—DATAVERSITY. 2023. Available online: https://www.dataversity.net/ai-chatbots-revolutionize-depression-management-and-mental-health-support/ (accessed on 21 August 2023).

Vaidyam, A.N.; Wisniewski, H.; Halamka, J.D.; Kashavan, M.S.; Torous, J.B. Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape. *Can. J. Psychiatry* **2019**, *64*, 456–464. [CrossRef] [PubMed]

Daley, K.; Hungerbuehler, I.; Cavanagh, K.; Claro, H.G.; Swinton, P.A.; Kapps, M. Preliminary Evaluation of the Engagement and Effectiveness of a Mental Health Chatbot. *Front. Digit. Health* **2020**, *2*, 576361. [CrossRef] [PubMed]

Inkster, B.; Sarda, S.; Subramanian, V. An Empathy-Driven, Conversational Artificial Intelligence Agent (Wysa) for Digital Mental Well-Being: Real-World Data Evaluation Mixed-Methods Study. *JMIR mHealth uHealth* **2018**, *6*, e12106. [CrossRef]

Lim, S.M.; Shiau, C.W.C.; Cheng, L.J.; Lau, Y. Chatbot-Delivered Psychotherapy for Adults with Depressive and Anxiety Symptoms: A Systematic Review and Meta-Regression. *Behav. Ther.* **2022**, *53*, 334–347. [CrossRef]

Fitzpatrick, K.K.; Darcy, A.; Vierhile, M. Delivering Cognitive Behavior Therapy to Young Adults with Symptoms of Depression and Anxiety Using a Fully Automated Conversational Agent (Woebot): A Randomized Controlled Trial. *JMIR Ment. Health* **2017**, *4*, e19. [CrossRef]

Klos, M.C.; Escoredo, M.; Joerin, A.; Lemos, V.N.; Rauws, M.; Bunge, E.L. Artificial Intelligence–Based Chatbot for Anxiety and Depression in University Students: Pilot Randomized Controlled Trial. *JMIR Form. Res.* **2021**, *5*, e20678. [CrossRef]

Jang, S.; Kim, J.-J.; Kim, S.-J.; Hong, J.; Kim, S.; Kim, E. Mobile app-based chatbot to deliver cognitive behavioral therapy and psychoeducation for adults with attention deficit: A development and feasibility/usability study. *Int. J. Med. Inform.* **2021**, *150*, 104440. [CrossRef]