NATIONAL BOARD FOR CERTIFIED COUNSELORS®

Ethical Principles for Artificial Intelligence in Counseling

April 12, 2024

INTRODUCTION

The emergence of artificial intelligence (AI) tools presents new opportunities as well as ethical challenges for counselors. The responsible and ethical integration of AI requires upholding core principles of counseling ethics and client well-being.

Accountability
While AI provides recommendations, the counselor bears responsibility for clinical decisions and is accountable for client outcomes. There must be human oversight for AI-generated treatment plans or interventions.

Client Welfare
As with any service delivered, the counselor's primary responsibility is advancing client welfare and the ethical delivery of care. AI integration must demonstrably meet this directive. If client well-being may be compromised, alternative approaches should be used.

Competence (AI)
Prior to integrating AI in their services, counselors must possess a foundational knowledge of how the technology works, its capabilities and limitations, and the ethical use of client data. This is consistent with the need to demonstrate competence with any assessment tool or intervention.

Competence (Clinical)
AI must not be used as a replacement for professional judgment or the counselor–client relationship. Counselor competence in assessment, diagnosis, and treatment is still essential. Consideration must be given to multicultural issues and carefully considered to ensure unbiased and competent delivery of services.

Confidentiality
Counselors have an ethical responsibility to protect client privacy and confidentiality when using or storing client data with AI tools. Policies and safeguards must be in place to prevent unauthorized access or use.

GENERAL TENETS

Each tenet specifies a corresponding directive from the NBCC *Code of Ethics*.

Transparency and Consent

- Disclose to clients if and when AI will be used and explain its purpose, limitations, and specific types. (NBCC directive 32)
- Obtain informed consent specifically regarding AI use. (NBCC directive 34)
- Allow client choice to opt out of AI tools. (NBCC directive 17)
- Exercise transparency about data access and sharing and retention policies. (NBCC directive 19)

Competence and Oversight

- Possess foundational knowledge of AI capabilities and limitations in the counseling context. (NBCC directive 1)
- Review AI recommendations for appropriateness before application to clients. (NBCC directive 5)
- Intervene or override AI if client welfare may be compromised. (NBCC directive 17)
- Maintain responsibility for clinical decisions and oversight. (NBCC directive 8)

Accountability

- Ensure to clients and others that the counselor remains fully accountable for client outcomes. (NBCC directive 8)
- Document instances of AI use, details of tools utilized, and justifications. (NBCC directive 35)
- Evaluate client outcomes to assess AI effectiveness. (NBCC directive 36)
- Report problems with AI tools to developers and clients as appropriate. (NBCC directive 10)

CLINICAL TENETS

Each tenet specifies a corresponding directive from the NBCC *Code of Ethics*.

Client Welfare

- Prioritize client welfare over efficiency or financial incentives for using AI. (NBCC directive 17)
- Intervene or override AI recommendations if client well-being may be compromised. (NBCC directive 17)
- Ensure competence to determine when AI use is and is not appropriate. (NBCC directive 1)
- Provide alternative services if a client declines consent for AI use. (NBCC directive 17)
- Discuss and address any client concerns regarding AI tools and data usage. (NBCC directive 32)
- Ensure clients have access to compatible technology before engaging AI. (NBCC directive 99)

Confidentiality

- Encrypt all client data used by or stored in AI systems. (NBCC directive 27)
- Restrict internal access to AI tools only to necessary personnel. (NBCC directive 18)
- Develop and implement policies for secure deletion of client data if no longer needed. (NBCC directive 30)
- Meet and maintain all laws and regulations for digital security and privacy. (NBCC directive 13)

Telemental Health

- Disclose use of AI tools in telemental health informed consent. (NBCC directive 103)
- Ensure AI systems meet security standards for digital transmission with HIPAA and other state regulatory requirements. (NBCC directive 93)
- Review automatically generated client risk assessments carefully before application. (NBCC directive 5)
- Cease use and report problems with AI telemental health tools to developers or organizations. (NBCC directives 13, 17)

Supervision and Consultation

- Supervisors should be knowledgeable about AI tools used by supervisees. (NBCC directive 41)
- Supervisors should explain if and how AI will be used in supervision. (NBCC directive 42)
- Address ethical use of AI in supervision agreements. (NBCC directive 42)
- Consult experts if unfamiliar with details of an AI system and discuss with supervisees. (NBCC directive 5)
- Document consultations about AI tools in supervision notes. (NBCC directive 50)

Counselor Education

- Possess competence to instruct students on AI systems. (NBCC directive 83)
- Teach critical thinking on appropriate AI integration. (NBCC directive 83)
- Advise students that some tools are unproven or experimental. (NBCC directive 83)
- Report AI biases and problems to developers. (NBCC directive 10)

Testing and Appraisal

- Ensure competence with AI tests/assessments used. (NBCC directives 60, 63)
- Explain AI testing purposes, limitations, and rights. (NBCC directives 64, 65)
- Review AI test/assessment recommendations carefully before applying for client care. (NBCC directive 62)
- Ensure secure storage of sensitive client data. (NBCC directive 58)

Research and Reporting

- Anonymize client data before use in research. (NBCC directive 70)
- Disclose AI use in securing informed consent for research. (NBCC directive 74)
- Validate AI research findings with transparency. (NBCC directives 75, 76)
- Credit contributors appropriately, including data sources. (NBCC directive 79)

Gatekeeping and Advocacy

- Ensure counselors are well prepared and competent to consider and use AI tools in clinical practice. (NBCC directive 88)
- Advocate for the ethical use of AI in counseling and address programmatic barriers related to AI technology. (NBCC directive 89)
- Monitor supervisees'/students' ethical AI use as part of gatekeeping role. (NBCC directive 88)
- Intervene if supervisees/students use AI in concerning ways. (NBCC directive 45)
- Report AI systems causing possible client harm and advocate for improved safeguards to AI developers. (NBCC directives 10, 17)
- Stay informed on AI risks to maintaining cultural competence. (NBCC directive 7)

From: **NBCC Ethics Department** ethics@nbcc.org
Subject: Ethical Principles for Artificial Intelligence in Counseling
Date: June 3, 2024 at 2:16 PM
To: Me@anorton.com

ND

NATIONAL BOARD FOR CERTIFIED COUNSELORS.

Dear NCC,

NBCC is pleased to release the *Ethical Principles for Artificial Intelligence (AI) in Counseling*. This supplementary document helps counselors to apply the existing principles of the NBCC *Code of Ethics* to matters of AI use. It is the result of collaboration between NBCC and subject matter experts who are also members of the NBCC Ethics Advisory Council.

We encourage all NCCs to become familiar with this document, regardless of their current plans for AI. You can read the *Ethical Principles for Artificial Intelligence in Counseling* here.

Sincerely,

The NBCC Ethics Department

NBCC
3 Terrace Way
Greensboro, NC 27403

More Info
NBCC Code of Ethics
Ethics Disclosure

Reach Out
tel: 336-547-0607
fax: 336-547-0017
email: ethics@nbcc.org

AMHCA Code of Ethics

Addendum December 2023

AMERICAN MENTAL HEALTH COUNSELORS ASSOCIATION

The only organization working exclusively for the mental health counseling profession

AMHCA CODE OF ETHICS
Addendum December 2023

Ethical Priorities
for Clinical Mental Health Counseling

Addendum to AMHCA Code of Ethics (Revised 2020)

I. Commitment to Clients

B. Counseling Process

6. The Use of Technology Supported Counseling and Communications (TSCC)

   l. CMHCs clearly disclose to clients when clinical services are provided by any modality other than direct person-to-person contact. In order to support client autonomy, CMHCs must disclose the extent, if any, to which clinical services are provided through, or in conjunction with, the use of artificial intelligence, machine learning, deep learning, or any other human simulation modality. CMHCs must obtain client consent before initiating these clinical services.

   CMHCs remain ultimately responsible for all clinical services they provide including their choices about client evaluations, treatment planning, interventions, and assessments. When CMHCs use technology to support clinical services, they are also responsible for ensuring that such technology is sufficiently safe, appropriate, and effective.

The unabridged version of *AMHCA Code of Ethics* appears in Appendix C of "Essentials of the Clinical Mental Health Counseling Profession," and is also available at no cost from *www.amhca.org/publications/ethics*.

American Mental Health Counselors Association
107 S. West St, Suite 110
Alexandria, VA 22314
703-548-6002
www.amhca.org

Portal Form: Artificial Intelligence (AI) Companion Consent Form

About Artificial Intelligence (AI) Companion Consent Form

Edit

Sharing on the Portal: Shareable on Demand

Staff Access: Administrative

Artificial Intelligence (AI) Companion Consent Form

Edit

Patient: Patient Name , DOB Patient DOB

Date: Date Submitted

⚠ This is a preview of the form and information here will not be saved.

Integrity Counseling uses Zoom for Healthcare, a HIPAA-compliant platform, for telehealth sessions.

Your therapist may offer you the option of using Zoom's AI Companion, a generative artificial intelligence (AI) tool that can answer questions, generate a transcsipt of your session, and access an AI-generated summary of the session, highlighting important content from the session that you can review afterwards. Some clients find it helpful to review and apply this information in their lives between sessions.

When humans take notes, they sometimes make errors, and AI does, too. As is the case with all generative AI tools, Zoom for Healthcare's AI Companion is imperfect, and AI-generated summaries and transcripts sometimes contain errors. You should not assume that everything in your transcript and summary is accurate, and if you have any questions about the transcript or summary, please discuss them with your therapist.

Some generative AI tools use information from meetings and sessions to "train" an AI tool. In healthcare, such use of information would be inappropriate, as the content of your sessions is confidential. Fortunately, Zoom does not use any customer audio, video, chat, screen sharing, attachments, or other communications like customer content (such as poll results, whiteboard, and reactions) to train Zoom's or its third-party artificial intelligence models.

If you choose to activate Zoom's AI Companion during your session(s), we ask that you agree to refrain from sharing the AI-generated summary and transcript of your session without your therapist's consent.

Do you consent to your therapist activating Zoom's AI Companion for each session, including generating a transcript and summary that you can refer to in between appointments. *

- ○ Yes, I consent to activating Zoom's AI Companion during sessions.
- ○ No, I do not consent to activating Zoom's AI Companion during sessions.

Do you understand that AI-generated transcripts and summaries may contain errors, that you should not assume them to be fully accurate, and that you should discuss any questions you have about them with your therapist? *

- ○ Yes, I understand that AI-generated text sometimes contains errors, and I'll discuss questions about AI-generated text with my therapist.
- ○ No, I do not understand and would like to discuss this further with my therapist.

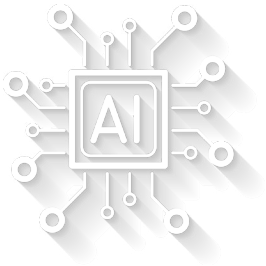
Do you agree to refrain from sharing your AI-generated transcript and summary with others without your therapist's consent? *

- ○ Yes, I agree to refrain from sharing AI-generated transcripts and summaries from my sessions with others without my therapist's consent.
- ○ No, I do not agree to refrain from sharing AI-generated content from my sessions with others without my therapist's consent.

Practice Acknowledgment

☐ **Sign This Form:** I, Practice User, have reviewed this document on Date.

Share on Portal



Making Sense of ChatGPT and Other AI Tools

Terminology

Adversarial learning:

Adversarial learning is a machine learning technique that involves training models to defend against adversarial attacks. It improves the model's robustness and reliability in applications where security and safety are critical, such as in computer vision, natural language processing, and autonomous systems.

Analogical reasoning:

Analogical reasoning is a cognitive process that involves identifying similarities and relationships between different concepts or ideas. It involves using knowledge from one domain to reason about a different domain, and it is used in many fields, including artificial intelligence, philosophy, and mathematics.

Artificial intelligence:

Artificial intelligence is the theory and development of computer systems able to perform tasks that normally require human intelligence, such as visual perception, speech recognition, decision-making, and translation between languages.

Burstiness:

Burstiness refers to the phenomenon where certain words or phrases occur more frequently within a short period of time, followed by longer periods of no occurrence. This uneven distribution of word frequency can affect the performance of statistical models and algorithms used in NLP, and it is an important factor to consider when analyzing language data.

ChatGPT:

ChatGPT is an AI language model developed by OpenAI that can understand and generate human-like responses to text-based inputs.

Natural language processing:

Natural Language Processing (NLP) is a field of study in artificial intelligence that focuses on enabling computers to understand, interpret, and generate human language. NLP is used in many applications, such as chatbots, language translation, sentiment analysis, and speech recognition.

Natural language generation:

Natural Language Generation (NLG) is a subfield of Natural Language Processing (NLP) that focuses on generating human-like language from structured data or other inputs. NLG is used in many applications, such as chatbots, automated journalism, and data-to-text systems.

Exploring AI and ChatGPT: Impact and Possibilities

Natural language inference task:

A national language inference task is a common task in NLP processing that involves determining the relationship between two pieces of text. The goal of NLI is to determine whether a given hypothesis can be inferred from a given premise, based on the meanings of the two texts.

Neural network:

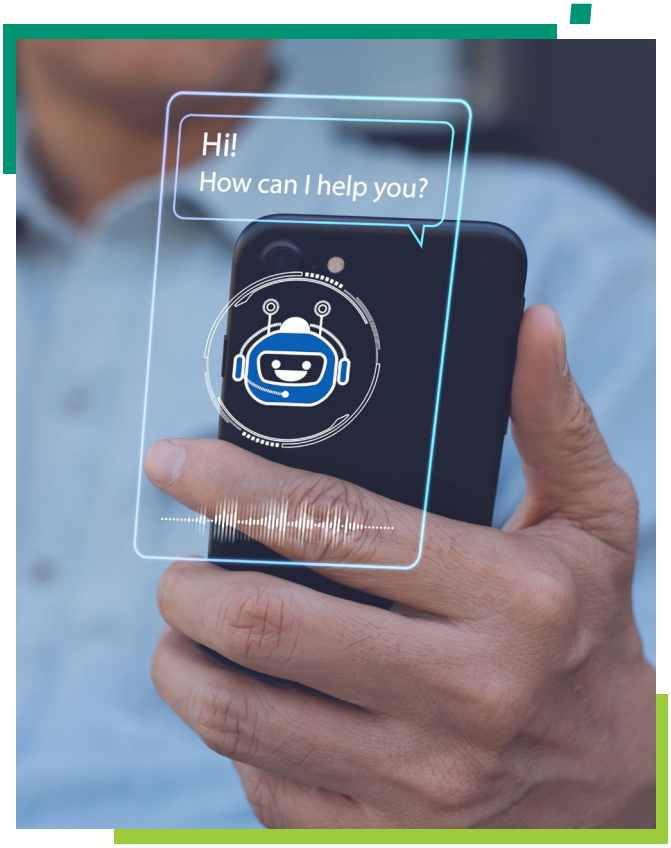
A neural network is an artificial intelligence model that consists of interconnected processing nodes or “neurons”. Information is passed between the nodes in a series of layers to process and transform the information. Neural networks are used in a wide range of applications, such as image and speech recognition, natural language processing, and predictive modeling.

OpenAI:

OpenAI is an AI research laboratory consisting of a for-profit company and a non-profit organization, based in San Francisco, California. It conducts cutting-edge research in many areas of AI and advocates for transparency, safety, and responsible use of AI.

Perplexity:

Perplexity is a measure of how well a statistical model predicts a sequence of words. It is commonly used in natural language processing to evaluate the quality of language models and the probability of a sequence of words.



Standard benchmark tasks:

Standard benchmark tasks are commonly used tasks in a field of study that are widely recognized and established, allowing researchers to compare the performance of different models or methods in a consistent and standardized way. In natural language processing, examples of standard benchmark tasks include language modeling, machine translation, sentiment analysis, and named entity recognition. These tasks provide a way to measure the effectiveness and progress of NLP models over time.

frontiers | Frontiers in Psychiatry

TYPE Conceptual Analysis
PUBLISHED 08 March 2024
DOI 10.3389/fpsyt.2024.1346059

OPEN ACCESS

EDITED BY
Toni Mancini,
Sapienza University of Rome, Italy

REVIEWED BY
Georg Starke,
Swiss Federal Institute of Technology Lausanne, Switzerland
Howard Ryland,
University of Oxford, United Kingdom
Lena Machetanz,
Psychiatric University Hospital Zurich, Switzerland

*CORRESPONDENCE
Leda Tortora
ltortora@tcd.ie

RECEIVED 28 November 2023
ACCEPTED 31 January 2024
PUBLISHED 08 March 2024

CITATION
Tortora L (2024) Beyond Discrimination: Generative AI Applications and Ethical Challenges in Forensic Psychiatry.
Front. Psychiatry 15:1346059.
doi: 10.3389/fpsyt.2024.1346059

Beyond Discrimination: Generative AI Applications and Ethical Challenges in Forensic Psychiatry

Leda Tortora*

School of Nursing and Midwifery, Trinity College Dublin, Dublin, Ireland

The advent and growing popularity of generative artificial intelligence (GenAI) holds the potential to revolutionise AI applications in forensic psychiatry and criminal justice, which traditionally relied on discriminative AI algorithms. Generative AI models mark a significant shift from the previously prevailing paradigm through their ability to generate seemingly new realistic data and analyse and integrate a vast amount of unstructured content from different data formats. This potential extends beyond reshaping conventional practices, like risk assessment, diagnostic support, and treatment and rehabilitation plans, to creating new opportunities in previously underexplored areas, such as training and education. This paper examines the transformative impact of generative artificial intelligence on AI applications in forensic psychiatry and criminal justice. First, it introduces generative AI and its prevalent models. Following this, it reviews the current applications of discriminative AI in forensic psychiatry. Subsequently, it presents a thorough exploration of the potential of generative AI to transform established practices and introduce novel applications through multimodal generative models, data generation and data augmentation. Finally, it provides a comprehensive overview of ethical and legal issues associated with deploying generative AI models, focusing on their impact on individuals as well as their broader societal implications. In conclusion, this paper aims to contribute to the ongoing discourse concerning the dynamic challenges of generative AI applications in forensic contexts, highlighting potential opportunities, risks, and challenges. It advocates for interdisciplinary collaboration and emphasises the necessity for thorough, responsible evaluations of generative AI models before widespread adoption into domains where decisions with substantial life-altering consequences are routinely made.

KEYWORDS

forensic psychiatry, forensic AI, generative AI, generative artificial intelligence, discriminative AI, ethical AI, large language models, large generative AI models

1 Introduction: discriminative vs generative AI

Generative Artificial Intelligence (GenAI) is a subfield of artificial intelligence which uses machine learning and deep learning techniques to generate 'seemingly new' human-like content, such as text, images, audio, and video, in response to various prompts, which are specific instructions provided to the AI system to execute a particular task or achieve a specific outcome.

Unlike the previously prevalent paradigm, known as discriminative artificial intelligence, which primarily focuses on discrimination tasks, such as classifying or differentiating between classes in a given dataset, generative AI models distinguish themselves by their capacity to both discriminate and generate new information based on the input data (1).

Discriminative AI models, mainly used for supervised machine-learning tasks like classification or regression, are algorithms designed to classify data instances by learning the decision boundaries that separate different classes or labels within a dataset. Examples of discriminative models include Support vector machines (SVMs), Decision Trees, Random Forests and Logistic Regression. On the other hand, generative AI models, mostly used in unsupervised and semi-supervised machine learning tasks like clustering and dimensionality reduction, are statistical models that learn regularities and patterns within input data and then use this acquired knowledge to generate novel data instances that share similarities with the original training data. Common examples of generative models include Generative Adversarial Networks (GANs), Hidden Markov models, Bayesian Network Autoregressive models and Latent Dirichlet Allocation (LDA) (2).

From a mathematical perspective, a discriminative machine learning approach trains a model by optimising parameters to maximise the conditional probability P(Y|X). In contrast, a generative model learns parameters by maximising the joint probability P(X, Y), relying on Bayes' Theorem (3). Consequently, unlike discriminative algorithms that focus on discerning decision boundaries, generative models produce artefacts with a wide range of variety and complexity (4). Additionally, while discriminative models aims for deterministic outcomes, the outputs of generative models are probabilistic and exhibit intrinsic variability (5).

The development of powerful generative AI models was prompted by the introduction of the Transformer neural network architecture in 2017 (6), which marked a significant milestone in machine learning research. Moreover, recent years have witnessed a surge in popularity and a growing interest in the application of generative models, especially since the release of ChatGPT, the popular conversational chatbot launched by OpenAI in November 2022 (7), which brought the concept of generative AI to the general public.

ChatGPT is an example of large language models (LLMs), which are deep learning models programmed to understand and generate natural language; these models, having been trained on a massive corpus of textual data, are able to produce human-like text and perform a range of language- related-tasks (i.e. text generation, question answering, language translation and more), interacting with the user conversationally (8, 9).

Nevertheless, it is important to note that generative AI can generate a wide array of outputs beyond text. For this reason, throughout this paper, the broader term 'large generative AI models (LGAIMs)' will be adopted to encompass all the different types of generative AI models, of which large language models (LLMs) are only a subset (10).

2 Types of large generative AI models

Large generative AI models (LGAIMs) comprise several subsets of generative AI models designed to generate realistic content across different modalities, such as large language models (LLMs), producing text (e.g., GPT-4, ChatGPT, Bard, Bing) and unimodal and multimodal models generating other media, such as images (e.g., Stable Diffusion, DALL·E 3, Bing Image Creator), videos (e.g., Synthesia, Imagen Video), audio (e.g., MusicLM, Musenet) and more.

Large generative AI models comprise several billion parameters, are trained on large datasets, and rely on significant computational resources. Many large generative AI models are currently in use, and their numbers continue to grow as AI experts experiment with existing models. LGAIMs can be classified according to several criteria, one of which is to categorise them by their underlying architecture. Generative AI comprises a variety of models employing different training mechanisms and output generation processes. At present, the most prevalent generative AI models are:

2.1 Generative adversarial networks

GANs are a class of models introduced in 2014 by Ian (11) A Generative Adversarial Network (GAN) consists of two neural networks: a generative model, known as the Generator (G), and a discriminative model, known as the Discriminator (D), working jointly in an adversarial manner to generate realistic data (12, 13).

GANs are best suited for tasks requiring the creation of authentic-looking data, such as images (14) and videos (15), favouring their use in industries such as entertainment and advertising, but also exposing them for potential malicious uses, such as deepfakes generation (16).

2.2 Transformer-based models

Also called 'foundation models' (17) because they serve as the foundation upon which many other AI models are built, Transformers were introduced in 2017 (6) by a team of Google researchers.

A Transformer model is a type of neural network relying on a set of mathematical techniques called attention mechanisms or self-attention mechanisms; these mechanisms assign weights to each input representation and dynamically learn the most relevant

information from the input data. The resulting output is obtained by computing a weighted sum of the input values, determining the weights through a compatibility function relating the query with its corresponding key (6, 18).

Those features allow transformer models to learn context by capturing relationships within sequential data, like the words in a sentence, making them ideal for tasks like text generation, and content and code completion. As a result, they have been highly successful in natural language processing (NLP) applications, being the foundation upon which the most popular type of generative AI models, large language models (LLMs), are built. Common subsets of Transformer-based models include Bidirectional Transformers (BERT) Models (19) and Generative Pre-Trained Transformers (GPTs) (20), such as GPT-4, GPT-3, T5 (Text-To-Text Transfer Transformer) and more.

2.3 Diffusion models

Diffusion models were developed by Stanford researchers in 2015 (21). They are probabilistic generative models that work by iteratively injecting Gaussian noise into the data. Then, a series of probabilistic denoising steps are applied to reverse this procedure and generate new data samples (22).

Diffusion models have found applications, especially in image generation (23), synthesis (24), and image super-resolution (25). They are the architecture of popular image generation services, such as Dall-E 2, Stable Diffusion, and Midjourney. In addition, they showed promising results in text-to-speech (26), text-to-video (27) and text-to-3D (28).

2.4 Variational autoencoders

Variational Autoencoders (VAEs) were introduced in 2013 by Kingma & Welling (29); they are generative models that encode input data into a lower-dimensional latent space and subsequently reconstruct it to its original form. This process involves three components: an Encoder, compressing input data into a probabilistic latent space, the Latent Space, retaining the compressed knowledge, and a Decoder, reconstructing the input data from the compressed latent space (30).

VAEs have found wide applications in several tasks, including image (31), text (32) and music generation (33). Furthermore, VAEs also excel at data compression (34), anomaly detection (35) and missing data imputation (36), and carry the potential for innovation in areas such as finance, speech/audio source separation, and bio signal applications (30).

2.5 Neural radiance fields

Neural Radiance Field (NeRF) is a novel approach in computer graphics and computer vision introduced by Mildenhall et al. in 2020 (37).

NeRFs are novel view synthesis methods, mainly applied to create highly detailed and photo-realistic 3D reconstructions of scenes based on 2D images; they achieve this through volume rendering techniques and implicit neural scene representations, often employing multi-layer perceptrons (MLPs) to synthesise novel views of 3D scenes by learning both their geometry and lighting characteristics (38). Therefore, NeRF models have found diverse applications across fields such as photo-realistic 3D editing (39), medical 3D image reconstruction (40), and neural scene representations for world mapping (41). NeRFs have also shown potential in areas like the industry and robotics domain (42), autonomous navigation (43), and augmented and virtual reality (44), where they carry the potential to lead to more efficient techniques for capturing and generating human 3D avatars and objects in the metaverse (45, 46).

Finally, large generative AI models (LGAIMs) can be broadly categorised into two main types: unimodal models and multimodal models. Unimodal models are designed to process just one type of input and generate content based on prompts from the same data format; examples of unimodal models are OpenAI's GPT-3, NVIDIA's StyleGAN2 or Google's BERT. On the other hand, multimodal models are designed to accept inputs and prompts from different modalities and generate content that combines information from different sources and data formats, such as text and images, resulting in more comprehensive outputs (47); examples of multimodal LGAIMs are OpenAI's GPT-4, ImageBind by Meta AI, and PaLM 2 by Google.

3 Discriminative AI's applications in forensic psychiatry

Before the recent progress and growing popularity of generative AI, discriminative AI was the dominant paradigm in artificial intelligence applications. In forensic psychiatry and criminal justice, discriminative models were developed to assist forensic psychiatrists and legal professionals in assessment and decision-making processes, for instance, informing decisions about pretrial risk assessment, sentencing, bail, parole, probation, allocation to rehabilitation programmes, timing and discharge conditions, and the need for further evaluations.

The most popular and debated application of AI in forensic psychiatry is violence and recidivism risk assessment. Discriminative AI models have been developed to evaluate and predict the likelihood of violence, recidivism, or other unlawful or harmful outcomes in individuals with a psychiatric or criminal history. Within risk assessment, discriminative algorithms feature many applications, such as predicting the risk of general, violent and sexual recidivism (48–52), forecasting future offences (53, 54) and evaluating risk of violence and aggression in psychiatric settings (55, 56), especially amongst individuals labelled as having an enhanced risk of engaging in violent conducts, such as patients diagnosed with schizophrenia (57–60).

These models classify individuals into different risk levels by analysing a vast range of data, including clinical assessments, patient history, demographic factors, and clinical notes.

Additionally, they can incorporate personalised data derived from physiological metrics, such as movement sensors and electronic health records (61).

In recent years, there has been a growing interest in integrating genetic, electrophysiological, and neuroimaging data into algorithmic risk assessment models in psychiatry (62). For instance, AI has been coupled with neuroimaging in a technique defined as 'AI Neuroprediction', which is the use of structural or functional brain variables coupled with machine learning techniques to identify neurocognitive markers for the prediction of recidivism (63).

In addition to risk assessment, discriminative AI tools have also been applied to improve diagnostic support, aiming to enhance clinical decision-making and diagnostic accuracy. Discriminative algorithms can analyse various types of data, such as behavioural patterns, speech, and textual data, like patient interviews or questionnaires, through several techniques, like natural language processing (NLP), to acquire diagnostic insights; for instance, they can perform machine-learning-based sentiment analyses to examine the patient's psychological condition and identify potential risks for harmful behaviours, such as risk factors associated to suicide in youth (64). AI-based decision support system (AI-based DSSs) have been applied to various tasks, from the prediction of mental health disorders (65) to risk assessment and management in patients discharged from medium secure services (MSS) (66).

The aforementioned capabilities also find application in personalised treatment planning; by examining patient histories, symptoms, physiological data and responses to previous treatments, discriminative AI algorithms can provide treatment recommendations, uncovering previously unnoticed targets for intervention and aiding in developing more individualised rehabilitation programs for individuals transitioning the criminal justice system. Furthermore, discriminative models, by predicting the potential treatment's effectiveness for each individual, could help optimise resource allocation. This issue is particularly relevant in forensic psychiatry, where institutions often grapple with acquiring sufficient resources to meet patients' needs and specialised service demands due to limited staff and financial support.

Nonetheless, it is crucial to highlight that the applications of AI in forensic psychiatry raise several legal and ethical issues that, since their advent, have largely yet to be addressed. While technologies develop at an incredibly fast pace, regulatory policies about their applications struggle to keep up.

The outputs of AI forensic risk assessment tools, relying on datasets reflecting historical biases and ongoing prejudice, have been shown to discriminate against historically marginalised groups in society, perpetuating and amplifying societal systems of inequality. For instance, AI forensic risk assessment algorithms exhibit racial and gender bias, as they systematically overclassify Black defendants and women in higher-risk groups for criminal recidivism (67) and several issues have been raised about these models' lack of fairness, accuracy and transparency (68). Furthermore, AI-based decision support systems (DSSs), perpetuating biased decision-making, lead to harmful and discriminating outcomes, such as unfair allocation of resources (69), and the increased use of predictive algorithms by law enforcement for predictive policing results in increased surveillance of marginalised groups, raising concerns about privacy and civil liberties (70).

Thus, while it is evident that the criminal justice system continues to face challenges related to the implementation of emerging technologies, we are now entering a new era of AI, marked by the advent of generative artificial intelligence, which is expected to exacerbate them further.

4 Generative AI's transformative impact on forensic psychiatry

The recent advancements in generative AI and the continuous evolution of large generative AI models (LGAIMs) impact multiple societal sectors, from business and healthcare to education and science. Their influence is further extending to critical areas like courtrooms, correctional facilities, and psychiatric settings, where generative AI models hold the potential to reshape forensic mental health practices and law enforcement procedures. In this paragraph, it will be explored how generative models, through their ability to analyse unstructured data across different formats (multimodal generative AI) and generate new synthetic realistic data (data generation and data augmentation), carry the promise not only to influence traditional discriminative AI applications, like risk assessment and personalised treatment design but also to create new opportunities in areas previously underexplored, such as training and education.

Multimodal generative AI models refer to a type of artificial intelligence model designed to process and integrate a vast amount of different data types, for instance, audio recordings of patient interviews, behavioural video observations, and textual reports from psychiatric assessments, but also neuroimaging, genomic data and electronic health records.

In psychiatry, multimodal generative AI models have shown promising results through their ability to analyse multidimensional health data, aiding to predict treatment trajectories (71), improving data interpretation and assisting in the production of clinical reports (72).

The application of multimodal GenAI models in forensic psychiatry could aid in performing advanced behavioural analyses, thereby facilitating a more comprehensive assessment of the patient's condition and improving the predictive power of risk assessment tools. By their ability to incorporate the temporal information in the learning process, thus capturing the dynamic evolution of the extracted features for each patient (73), these models can integrate a wide range of contextual information, from verbal to non-verbal cues like tone, facial expressions, and body language, thus enabling the implementation of multimodal sentiment analysis and emotion detection tools, aiming to uncover individuals' emotional states and predict emotional categories. This enhanced emotion detection capacity could serve the development of advanced multimodal decision support systems (DSSs),

providing diagnostic insights and highlighting potential risk factors associated, for instance, with violence, aggression, or self-harm. Moreover, the ability of multimodal generative AI models to detect sudden changes and inconsistencies in emotional states could function as an 'early warning system', alerting mental health professionals about concerning patterns of behaviours and triggering further assessment. Finally, multimodal models could help tailor personalised interventions in treatment design and planning by integrating several data sources about a patient's profile and history.

It is important to note that this capability of enhanced behavioural and emotional analyses might also be misused for concerning applications, for instance, to build lie-detection tools to evaluate the credibility of offenders and witnesses or by attempting to reconstruct a person's mental state and memories during a specific crime. At the same time, the capacity to analyse and integrate a vast amount of personal data over time, from health records to communication history and social media posts, could also contribute to problematic AI profiling techniques.

In addition to multimodal models, generative AI could further influence forensic psychiatry practices by employing data generation and data augmentation techniques, referring to the ability to synthesise new data samples that share similarities with a given dataset.

The potential of generative AI to generate new data instances can impact treatment design and planning by facilitating the creation of personalised treatment simulations. These simulations involve AI-generated scenarios resembling patient profiles and treatment trajectories, enabling forensic clinicians and professionals to virtually test different treatment approaches before implementation and provide insights into their effectiveness. These simulations hold particular promise in addressing complex cases where the efficacy of treatment is uncertain, helping to optimise resource allocation and to evaluate new policies and interventions for individuals transitioning the criminal justice system. For instance, generative AI models could enhance the development of Digital Twins (DTs), virtual models simulating clinical patient trajectories and treatment effects (74), with the potential to assist in tailoring treatment plans, accelerate drug discovery and improve the efficiency of clinical trials (75).

A newly envisioned application leveraging generative AI's scenario simulation capabilities extends to the often overlooked dimension of training and education, where realistic synthetic scenarios simulating various forensic psychiatric case studies and patient interactions could allow forensic mental health professionals to practice decision-making and assessment skills.

By employing GenAI-powered virtual simulations, current and prospective forensic psychiatrists could practice and refine their diagnostic skills in controlled environments, where interactions with AI-generated patients simulating different psychiatric conditions could enable them to gain insights into different behavioural patterns and identify critical risk factors. These simulations could extend to various environments, such as virtual courtrooms and psychiatric settings, where GenAI-created scenarios could simulate ethical dilemmas to help forensic psychiatrists test and navigate ethically challenging situations they might encounter in practice. In the legal realm, they could also assist in defence by generating counterfactual scenarios to explore how patient outcomes might have unfolded under different circumstances.

Alongside AI-powered simulations, generative AI can employ data augmentation methods, commonly used to expand existing datasets by creating variations of the original data samples. Generative data augmentation techniques have been used to address data scarcity by generating synthetic data to train more robust predictive models for medical diagnosis of multiple mental health conditions (76, 77). By generating synthetic patient profiles, data augmentation tools provide supplementary data for analysis, expanding the training dataset for predictive models facing challenges related to insufficient or unbalanced data. This is an issue notably prevalent in forensic psychiatry, where datasets are often limited due to the sensitive nature of information, potentially resulting in an unbalanced representation of various mental health conditions or behavioural patterns.

Finally, it is crucial to highlight the role of generative AI models as decision-making support tools, considering the increasing number of experts who are consulting these models, especially large language Models (LLMs), looking for guidance on a variety of tasks, such as reviewing mental health evaluations in criminal cases, communicating findings in court, and accessing relevant case studies. The increasing use of generative AI will substantially influence decision-making processes within courtrooms and forensic psychiatry settings, regulating which information is accessed and used for report completion and evaluations, as well as affect the data collection processes and diagnostic assessments, for instance, through recommendations to administer relevant tests, questionnaires or interview questions.

In conclusion, the influence of generative AI on forensic psychiatry extends far beyond its discriminative AI applications, with considerable forthcoming developments and its unique set of possibilities and challenges.

5 Differences between discriminative and generative AI applications in forensic psychiatry

Generative and discriminative AI both hold potential for applications in forensic psychiatry, but they differ in how they approach the task in several ways.

First, they have different purposes. The primary goal of generative models is to generate new data instances resembling the training data by modelling the joint probability distribution of the observed data. On the other hand, discriminative models aim to distinguish between different classes or categories in the dataset by learning the conditional probability distribution. Therefore, they produce different outputs; while generative models produce data samples drawn from the learned probability distribution, such as realistic synthetic audio, video or textual content, discriminative models directly output class labels or continuous values, making them suited for different tasks.

Accordingly, discriminative and generative AI have different applications and use cases in forensic psychiatry (Prediction vs Generation). Discriminative models are primarily used for predictions, classifications, and regression tasks (Prediction) and find application in tasks such as violence and recidivism risk assessment, diagnostic support and treatment recommendations. On the other hand, generative models are best suited for tasks requiring data generation, data augmentation and probabilistic modelling (Generation). Generative models not only enhance the effectiveness of traditionally discriminative tasks, such as enabling comprehensive behavioural analyses through multimodal models, but also unlock novel opportunities, for instance, through the development of GenAI-powered simulations tailored for personalised treatments and interventions as well as training and educational purposes.

Generative and discriminative AI models further differ regarding training data requirements; specifically, generative models employ unsupervised learning techniques and are trained on unlabelled data, while discriminative models excel in supervised learning and are trained on a labelled dataset. Consequently, generative AI models require more extensive training data compared to discriminative AI algorithms, which can often perform relatively well with smaller datasets, especially when implementing methods like transfer learning or fine-tuning pre-trained models. Another difference pertains to interpretability; while achieving interpretability is already challenging in discriminative models, it becomes even more intricate with generative ones. Discriminative models, employing labelled data, provide outputs that can be interpreted as class probabilities, providing insights into predictive feature contributions. In contrast, generative models introduce a higher level of complexity, as their outputs may not correspond directly to known classes or categories.

In conclusion, the choice between these approaches will depend on the specific task's objectives, the desired outcome and the available data. Additionally, a hybrid approach combining both methods could offer benefits from both perspectives, contributing to more comprehensive results.

Finally, it is crucial to emphasise that the applications of discriminative and generative AI in forensic psychiatry must be approached carefully and require thorough analysis and regulation before widespread adoption.

6 Ethical and legal challenges of generative AI applications in forensic psychiatry and criminal justice

As previously discussed, large generative AI models (LGAIMs) are rapidly transforming many aspects of modern life, including how we communicate, create, and work, impacting various sectors of society. Nevertheless, generative AI models, like other transformative technologies, while harbouring enormous potential, also carry significant risks, and their application raises several ethical and legal concerns.

Misuses of this technology, especially in the fields of forensic psychiatry and criminal justice, might result in significant harm spanning from discrimination to predictive policing, mass surveillance and profiling, impacting individuals' freedom, right to a fair process, allocation of resources and education of the future generation of legal and mental health professionals. This paragraph will present an overview of some of the pivotal challenges associated with generative AI applications in forensic psychiatry and criminal justice.

The discussion will begin by examining the impact of generative AI on some of the prevalent challenges in AI implementation in forensic psychiatry and criminal justice, encompassing issues such as biases and criminalisation, lack of transparency and interpretability, data privacy, and autonomy. Subsequently, the discourse will delve into GenAI-specific challenges, covering topics such as hallucinations, deepfake fabrications, and homogenisation, along with issues like overreliance. Finally, the analysis will address broader societal concerns about the implementation of generative AI in society, such as environmental impact and power imbalances.

6.1 (Gen)AI bias-driven criminalisation

Discriminative AI algorithms are well-known for embedding several sources of harmful biases and stereotypes against historically marginalised groups within society, and generative AI models are no exception. Research has shown that large language models (LLMs) tend to replicate biases in the training data (78, 79), an issue already prevalent in discriminative algorithms.

For instance, large language models (LLMs) exhibit instances of racial and gender bias when, during in-context impersonation tasks, they describe cars better when asked to impersonate a black person or a male while describing birds better when impersonating a white person or a female (80). Furthermore, an analysis of GPT-2 and GPT-3.5 revealed a propensity to generate masculine-associated pronouns more frequently than feminine-associated ones and show gender-biased association in the context of professions, considering occupations such as Doctor or Engineer as masculine more often than roles like Nurse and Teacher, often regarded as feminine (81). Language-dependent ethnic biases, involving the over-generalised association of an ethnic group to specific attributes, mostly negative, have been found in BERT, where non-toxic comments are incorrectly labelled as toxic when including Middle Eastern country names (82).

Similarly, evidence of religious bias has been found in AI text generators, where the models generate words such as violent, jihad, bomb blasts, terrorism and terrorist at a greater rate in association with the religion Muslim or Islam than with other religions (83, 84).

Biases are also present in the often overlooked dimension of disability; studies have shown that, even when disability is not discussed explicitly, pre-trained language models (PLMs) consistently assign more negative scores to sentences containing words associated with disability compared to those that do not (85). This confirms previous findings indicating that a high percentage of online comments mentioning disabilities on the Jigsaw (86) dataset

was labelled as toxic and showed an over-representation of terms related to homelessness, gun violence, and drug addiction, negatively impacting the representation of disability (87).

These systems further suffer from an intersectional bias, where the intersection of different categories of social difference results in new forms of stigmatisation (88).

Those biases are not limited to LLMs but are also visible in Text-to-image (TTI) generative models; for instance, DALL-E 2 has been shown to underrepresent women in stereotypically male-dominated fields while overrepresenting them in stereotypically female-dominated occupations, frequently portraying a higher representation of women than men wearing smiles and tilting their heads downward, particularly in stereotypically female-dominated occupations (89).

Text-to-image (TTI) models' outputs have also been found to perpetrate identity-based stereotypes, for instance, generating stereotyped images of non-cisgender identities (90) and reproducing Western-centric representations (91, 92), resulting in the reinforcement of whiteness as the ideal standard, the amplification of racial and gender disparities, and the propagation of American-centred narratives (93).

These biased representations can have a profound impact on stakeholders, particularly when integrated into systems used in the forensic domain, where discriminatory outputs and inaccurate formulations have severe implications for all parties.

In fact, forensic psychiatric patients are a population already facing high levels of stigmatisation, as mental illness and criminal history are both commonly associated with social dangerousness, a stereotyped representation widely held in the public perception and permeating society at many levels (94). As a consequence, forensic psychiatric patients are frequently exposed to experiences of rejection and alienation, contributing to a higher risk of internalising negative perceptions held towards them, known as self-stigmatisation (95). Furthermore, these negative stereotypes are used to justify, legitimise and promote legal restrictions and discriminatory practices, such as increased use of coercion (96).

Within the correctional system, pervasive racial stigma intertwines with negative portrayals of forensic psychiatric patients as dangerous and aggressive, contributing to disproportionately high incarceration rates of African Americans (97) and their systemic over-diagnosis with highly stigmatised disorders associated with incompetence, such as psychotic disorders (98). As a result, forensic psychiatric patients face the intersection of multiple stigmatised identities, with damaging effects on self-esteem, depression, therapeutic alliance, and treatment adherence (99).

Within this context, the application of generative AI models in critical tasks that encompass life-altering outcomes, such as risk assessment, sentencing recommendation and treatment and rehabilitation planning, will not only reiterate but significantly magnify existing biases, exacerbating discrimination against forensic psychiatry patients, particularly those from historically marginalised groups, and reinforcing the stigma they experience across multiple levels of society.

For instance, research has shown that, as datasets used by generative AI models expand in scale, there is a noticeable increase in the likelihood of these models classifying Black individuals as 'criminal' or 'suspicious person,' perpetuating historical and racially biased patterns of criminalisation. Additionally, the deployment of text-to-image (TTI) models in applications like 'Forensic Sketch AIrtist' (2022) (100), a forensic sketch program by EagleAI developers utilising DALL-E 2, poses a substantial risk of exacerbating existing racial and gender biases inherent in original witness descriptions while aiming to generate 'realistic' sketches of police suspects based on users' inputs.

In summary, biased AI systems generate significant harm that cannot be overlooked. Generative AI models have the potential to significantly worsen these consequences, exacerbating disproportionate criminalisation of marginalised groups, perpetuating stigmatising attitudes and reinforcing harmful links between mental health and social dangerousness.

6.2 Transparency, interpretability, accountability

Understanding and explaining the complexity of generative AI models and their decision-making process to their stakeholders and those affected by their outputs is a challenging task, unveiling significant concerns related to their transparency and interpretability. The opacity of generative models contributes to a lack of accountability, exacerbated by the proprietary nature of the software (79) and by the absence of transparent, ethical oversight during these models' development, which prioritises hype and profit over ethical and accountable work (101). Additionally, the dominance of industry in AI research, due to its control over crucial resources such as computing power, extensive datasets, and highly skilled researchers, makes it challenging for Academia and the public sector to inquire, monitor, and audit AI models or provide alternative solutions (102), while simultaneously imposing an unfair burden of responsibility on them. The need for transparency and accountability, especially following the widespread adoption of generative AI models, calls for the creation of a regulatory framework tailored to respond to the dynamically changing AI landscape and to address not only the technical aspects but also the broader ethical, societal, and economic implications, promoting their responsible and ethical use (103) while favouring critical enquiries on issues related to responsibility, accountability, and labour exploitation (78).

6.3 Data quality, privacy & security

Training large generative AI models (LGAIMs) requires extensive data, often sourced from openly available internet data. This data often contains biased and undesirable content, raising concerns about data quality (104) as well as privacy and security issues. Web-scraped datasets might contain various personally identifiable information about the data subjects, such as their names and email addresses (105); as an example, the metadata scraped by text-to-image (TTI) generative models can include names or other personal information of the authors and the subjects of the media files.

Data privacy and security risks include unauthorised data collection, the risk of re-identification of previously anonymised data, and inadequate data retention practices that could lead to data privacy violations, such as data breaches and unauthorised data sharing.

During training, generative AI models may inadvertently encode and reproduce content containing sensitive data, posing a risk of data leakage. Moreover, even when explicit personal information is absent from the training data, the content generated by generative AI models, when combined with other accessible data, might still lead to the re-identification of individuals or the disclosure of their personal information.

In forensic psychiatry, where access to sensitive data, such as medical, criminal and psychiatric records, is bound to strict legal and ethical regulations, obtaining and using these data without adequate data protection measures violates privacy laws and ethical principles. Consequently, the use of generative AI models in such environments calls for robust regulation to ensure the confidentiality and security of patients' information, including guidelines for data anonymisation and retention and strategies to prevent data misuse and unauthorised access by external parties (103).

Moreover, if individuals are unjustly detained due to cyberattacks or hacked data, AI companies' lack of transparency and legal responsibility might leave affected individuals without adequate legal resources (106).

6.4 Intellectual property rights & copyright infringements

Although generative AI models gained popularity for their ability to generate novel content, it is crucial to note that the examples used by these models are typically derived from existing human-made works, raising issues of copyright infringement and unauthorised imitation. Large language models (LLMs) are trained on an extensive corpus of data, some of which may have been acquired without proper consent, as the models usually scrape data from the internet, disregarding copyright licenses, plagiarising content, and repurposing proprietary materials without permission.

As a result, it becomes challenging to trace the lineage of the content generated by those models, and due credit is frequently not given to the original creators, potentially exposing users to copyright infringement issues (107, 108) and resulting in legal actions against companies, accused of violating intellectual property rights (109).

6.5 Autonomy and informed consent

The widespread adoption of biased and opaque generative AI tools, developed without a robust regulatory framework, which increasingly influence decisions concerning an individual's psychiatric evaluation, treatment, or legal status, raises concerns about safeguarding individuals' autonomy and their level of agency over their own information and cases.

AI-driven decision-making tools greatly challenge the principle of respect for the patient's autonomy, especially in forensic psychiatry applications. In fact, unlike public safety protocols, critical activities such as rehabilitation and forensic mental health evaluations necessitate individuals' direct and voluntary participation (110).

The lack of transparency surrounding AI algorithms highly compromises the process of obtaining informed consent. Evaluators' limited understanding of how algorithms generate assessments, including the specific data considered, their respective importance, and the model's rationale, hinders their ability to effectively communicate this process to the individuals undergoing evaluation (110). This contradicts the fundamental principle of autonomy in medical ethics, which emphasises patients' control over procedures concerning them, including the use of their data.

Additionally, the incorporation of AI systems in medico-legal decision-making challenges the autonomy of forensic mental health professionals. As an additional factor altering the shared decision-making process between professionals and patients, algorithms undermine clinicians' perceived authority and impact their judgment. In fact, despite the increasing influence of AI recommendations, in instances where AI judgment conflicts with human judgment, the responsibility to authorise the treatment remains with the professional, who must feel empowered to make autonomous decisions (111).

Furthermore, increased reliance on AI outputs reduces professionals' use of their own ethical reasoning. Since professionals are responsible for evaluating these outputs, a weakened ethical judgment may impact the criteria used for algorithms assessment (112).

Finally, the application of AI in medico-legal decision-making poses significant challenges to both professionals and patients. If left unregulated, it undermines their authority over crucial decisions that directly influence their lives.

6.6 Overreliance

The current debate surrounding ChatGPT and generative AI is dominated by exaggerated and sensationalistic portrayals of their capabilities, resulting in overreliance on their outputs, exacerbating the risk of spreading misinformation and reinforcing biased stereotypes (113).

This overreliance carries profound implications in forensic psychiatry and criminal justice, where outputs of generative AI models increasingly influence clinical assessment and legal decision-making.

For instance, recent news reports have highlighted several instances in which judges and lawyers relied on ChatGPT's recommendations as a support for decision-making; for instance, a British Court of Appeal judge admitted using ChatGPT to summarise an area of law for a case ruling (114), and a judge in Colombia announced he consulted ChatGPT in preparing a ruling in a children's medical rights case (115). Similarly, a judge in a Pakistani court used the chatbot to render judgements in a case (116). In another instance, two layers have submitted false evidence generated through ChatGPT in an aviation injury claim (117) - a consequence of the chatbot's 'hallucination', a phenomenon discussed in the following paragraph- which also led to

the first major sanction on the use of artificial intelligence within the legal domain.

This growing trend is particularly concerning as it showcases how the widespread availability and ease of access to generative tools contrasts with the lack of awareness of the mechanisms behind their outputs. The situation is further aggravated by the marketing of these models as outstanding and infallible products, often portrayed as possessing human or even superhuman-level reasoning capabilities.

This issue highlights the necessity for AI companies to communicate the genuine potential of their products in a transparent and non-deceptive way, as well as to divulge details about the data used in the models and their analytical processes. Also, it illustrates the need to ensure the digital literacy of legal professionals in critical times of generative AI evidence (115).

6.7 Hallucinations, inaccuracy and (mis) facts fabrication

The previously mentioned episode concerning lawyers submitting fake evidence to the court is not an isolated case; in fact, large generative AI models (LGAIMs) have demonstrated tendencies to occasionally generate non-existent and false content, casting doubt on the accuracy of their outputs — a phenomenon called 'hallucination'.

'Facts fabrication' by generative AI models is not limited to the legal context but expands to various settings. For instance, ChatGPT has been shown to produce seemingly plausible but incorrect answers when asked about scientific topics (118) and to fabricate false references for scientific articles (119).

Hallucinations have been associated with disruptions in the language generation process. As large language models (LLMs) generate probabilistic outputs relying on estimations of semantic similarity, when a disruption occurs in this process, it can lead to the integration of false information alongside factual content, raising serious concerns about the trustworthiness of their outputs (120). Hallucinations are primarily associated with LLMs, but they also manifest in models generating video, images and audio; for instance, when Midjourney was tasked with generating images of people enjoying a house party, while the overall scene appeared realistic, a closer look revealed unrealistic elements such as individuals with an excessive number of teeth or hands with more fingers than usual (121).

The fabrication of (false) information risks misleading the users and, especially as a growing number of individuals rely on these tools for guidance and information, the continuous presentation of false information as a factual truth has the potential to distort the perception of reality, acting as a 'misinformation superspreader' and resulting in significant harm, especially when inaccurate outputs are used to support forensic decision-making.

6.8 GenAI deepfake evidence and the quest for reality

Progress in generative AI models resulted in the production of content that is increasingly challenging to distinguish from human-generated material.

Once evidence generated by generative AI enters the courtroom, it presents significant challenges to all parties. For instance, judges will face the complex task of ruling on an increasing number of disputes over the authenticity of evidence that might be contested as a deepfake.

The judicial system is currently unprepared to handle evidence derived from AI systems, an area in which they possess limited expertise. This compounds the complexity of ruling on digital evidence and creates a demand for technical experts knowledgeable in generative AI and deepfake technologies, further increasing costs and time duration of legal proceedings (122).

Moreover, the growing probability of encountering AI-generated evidence in courtrooms is likely to instil a sense of doubt and scepticism amongst judges, juries and the general public, fostering an environment where all parties are inclined to consider the possibility that their counterparts have submitted AI-generated evidence - a phenomenon also referred to as "the deepfake defence" (123), which ultimately pollutes the decision-making processes.

This phenomenon will create an environment characterised by an overarching sense of distrust, in which parties can weaponise scepticism and doubts to advance their own agendas, a concept also known as the "liar's dividend" (124).

Additionally, the advancement of tools for detecting AI-generated content raises questions about which content will likely be more targeted and the potential legal consequences of identifying AI-generated evidence. At present, AI-generated content detectors are insufficiently accurate and show notable inconsistency in categorising content as either AI-generated or human-written (125).

In forensic psychiatry, where research suggests that juries and judges tend to misinterpret scientific evidence in court, for instance overestimating the reliability of neuroscientific evidence (126), leading to miscarriages of justice (127), the potential introduction of genAI-fabricated evidence introduces the risk of wrongful convictions grounded in maliciously AI-generated scientific evidence.

In summary, the rise of generative AI introduces a concerning scepticism that could disrupt decision-making at an individual and societal level, underscoring the growing need to preserve our rights to reality in this evolving era of AI.

6.9 Environmental impact and sustainability

As we delve into discussing AI's impact and ethical development, it is imperative to mention that the impressive capabilities of generative AI models come at a hidden and frequently overlooked environmental cost. In fact, alongside the usage and continuous development of generative AI models, the computational power required to train them and maintain their physical infrastructure grows together with their carbon emissions, raising concerns from a climate policy perspective (128–130).

Although these tools are currently in the early stages of gaining mainstream adoption, it is reasonable to anticipate that their

environmental costs will grow significantly in the coming years. Consequently, it is crucial to develop metrics to evaluate the environmental impact of AI development to identify strategies to mitigate it (131).

6.10 Power, homogeneity and 'bias-in-the-loop'

A discussion about AI Ethics must encompass an analysis of power dynamics; understanding the positionality of the stakeholders and their respective levels of influence is, in fact, crucial for gaining insight into the potential hazards of AI.

Algorithmic bias is a symptom of a broader issue about power imbalances and historical inequities that influence AI technologies' creation, deployment, and objectives, starting from how data is collected and managed, including the authority in deciding which aspects are measured and included in the datasets (132). Technology is not neutral; AI solutions are value-laden and are "specified by developers and configured by users with desired outcomes in mind that privilege some values and interests over others" (133). Nowadays, there is a substantial disparity in the AI domain between the Global North and the Global South, wherein the latter is often subjected to exploitation for low-cost or unpaid labour, meanwhile the main benefits and advancements are concentrated in the Global North. As a result, individuals from the Global North gain early access to cutting-edge generative AI tools, while marginalised groups are left behind, causing issues of unequal access and exacerbating the existing disparities in the technological landscape.

Moreover, since companies employ user input to train their models, such as OpenAI, which may use content entered by users in ChatGPT to improve the model's performance (134), this process could introduce an additional 'bias-in-the-loop', where countries and individuals who get to access and use generative AI models will further control and shape their outputs through their inputs and queries, thereby intensifying digital disparities.

The emergence of generative AI has widened several layers of digital divides, holding significant implications for offline outcomes and amplifying digital inequalities. As a consequence, individuals lacking access to extensive data resources face vulnerability when comprehending the data and methodologies employed in decisions that impact them. The problematic nature of algorithmic decision-making, marked by an asymmetry in knowledge and decision-making authority, significantly exacerbates this vulnerability (135). The issue is intensified by significant power imbalances in the criminal justice system resulting from detention under mental health legislation, where forensic psychiatric patients often have limited access to technology, worsening disparities in access to information and communication resources.

Additionally, the widespread use of generative AI raises concerns about the diffusion of increasingly uniform outputs generated by AI models trained on a limited range of references. This homogenisation extends not only to language, communication styles, and public discourse but also to economic power and information, consolidating economic influence within a few organisations governing AI systems, fostering economic homogeneity and inequality.

To establish an ethical and responsible AI framework, it is imperative to integrate diverse perspectives and voices at every stage of the AI process, from dataset creation and curation to model development and utilisation. This imperative is inseparable from efforts to renegotiate and redistribute power. Without initiatives to rebalance power dynamics, the prospects for democratising AI and ensuring its responsible use remain elusive, especially within the biased criminal justice system.

7 Conclusions

The rapid advancements of technology and the widespread use of generative artificial intelligence in several fields require society to match the pace of these developments. Currently, we are falling short in this regard, allowing AI's outcomes to impact our lives prior to undergoing comprehensive investigation and regulation.

This article discusses the impact of generative AI in forensic psychiatry and criminal justice, analysing current and prospective applications while drawing comparisons with the previously dominant paradigm of discriminative AI.

This comparative exploration reveals the convergence of both past and emerging challenges. First, it becomes evident that generative AI not only holds the potential to revolutionise traditional discriminative tasks, for instance, by leveraging its enhanced analytical capabilities to enhance risk assessment and diagnostic support, but also to open avenues to previously overlooked applications, like AI-powered simulations for training and educational purposes.

When exploring the ethical and legal issues, the analysis shows that generative AI models not only inherit the prevailing challenges present in discriminative AI algorithms, such as biased and stereotyped outputs, lack of transparency, and data privacy issues, but also amplify their impact, due to heightened computational capabilities and increased accessibility and ease of use. Furthermore, generative AI models introduce novel and unique challenges, such as hallucinations and facts fabrication, progressive homogenisation of content, and concerns about data quality and intellectual property rights. Specifically, within forensic psychiatry, some of the most concerning aspects include the spread of misinformation and the reinforcement of discriminatory and criminalising narratives and stereotypes. This unfolds as a result of the increasing overreliance on AI-generated outputs used by judges, legal experts, and mental health practitioners in their decision-making processes. The situation becomes particularly problematic if biased outputs are employed for training and educational purposes, as they could have a negative impact on the perspectives and knowledge of future forensic mental health professionals.

In fact, large generative AI models carry the potential to strengthen the negative association between mental health and criminal history; as a consequence, there will be an increased risk of criminalisation of forensic psychiatry patients, especially those belonging to historically oppressed groups, alongside with

enhanced profiling, mass surveillance and unfair allocation of resources and treatment assignments.

While unregulated industry controls resources and power, institutions need to provide society with the necessary tools to investigate and hold those systems accountable. Continuous discussions and collaborations among stakeholders, including forensic psychiatrists, AI developers, legal experts, and ethicists, are essential to navigating these complex issues, while considering the diversity in forensic psychiatry practices shaped by differences in healthcare and legal systems among different countries. Additionally, maintaining an ongoing dialogue with affected communities, who often lack representation in these discussions, and involving them in the process, is crucial.

Lastly, as algorithms and their decision-making are a reflection of society, we need to work on shifting from a surveillance-based approach to one focused on tackling the root causes of criminalisation and inequality, emphasising the safeguard of mental health and rehabilitation over criminalisation and profiling.

Author contributions

LT: Conceptualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that no financial support was received for the research, authorship, and/or publication of this article. LT is funded by the European Commission the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 861047.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

1. Gozalo-Brizuela R, Garrido-Merchan EC. ChatGPT is not all you need. A State of the Art Review of large Generative AI models. (2023). arXiv preprint arXiv:2301.04655.

2. Harshvardhan GM, Gourisaria MK, Pandey M, Rautaray SS. A comprehensive survey and analysis of generative models in machine learning. *Comput Sci Rev*. (2020) 38:100285. doi: 10.1016/j.cosrev.2020.100285.

3. Liu B, Webb GI. Generative and discriminative learning. In: Sammut C, Webb GI, editors. *Encyclopedia of machine learning*. Springer, Boston, MA (2011). doi: 10.1007/978-0-387-30164-8_332

4. Sun J, Liao QV, Muller M, Agarwal M, Houde S, Talamadupula K, et al. (2022). Investigating explainability of generative AI for code through scenario-based design, in: *27th International Conference on Intelligent User Interfaces*, . pp. 212–28.

5. Weisz JD, Muller M, He J, Houde S. Toward general design principles for generative AI applications. (2023). arXiv preprint arXiv:2301.05578.

6. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. *Adv Neural Inf Process Syst*. (2017) 30.

7. OpenAI. ChatGPT: optimizing language models for dialogue (2022). OpenAI. Available online at: https://openai.com/blog/chatgpt/ (Accessed 10 September 2023).

8. Chang Y, Wang X, Wang J, Wu Y, Zhu K, Chen H, et al. A survey on evaluation of large language models. (2023). arXiv preprint arXiv:2307.03109.

9. Kasneci E, Seßler K, Küchemann S, Bannert M, Dementieva D, Fischer F, et al. ChatGPT for good? On opportunities and challenges of large language models for education. *Learn Individ Dif*. (2023) 103:102274. doi: 10.1016/j.lindif.2023.102274.

10. Hacker P, Engel A, Mauer M. (2023). Regulating ChatGPT and other large generative AI models, in: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, . pp. 1112–23.

11. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. *Adv Neural Inf Process Syst*. (2014) 27. doi: 10.48550/arXiv.1406.2661

12. Saxena D, Cao J. Generative adversarial networks (GANs) challenges, solutions, and future directions. *ACM Computing Surveys (CSUR)*. (2021) 54:1–42. doi: 10.1145/3446374.

13. Dubey SR, Singh SK. Transformer-based generative adversarial networks in computer vision: A comprehensive survey. (2023). arXiv preprint arXiv:2302.08641.

14. Li Z, Xia B, Zhang J, Wang C, Li B. A comprehensive survey on data-efficient GANs in image generation. (2022). arXiv preprint arXiv:2204.08329.

15. Wen S, Liu W, Yang Y, Huang T, Zeng Z. Generating realistic videos from keyframes with concatenated GANs. *IEEE Trans Circuits Syst Video Technol*. (2018) 29:2337–48. doi: 10.1109/TCSVT.76.

16. Singh A, Saimbhi AS, Singh N, Mittal M. DeepFake video detection: a time-distributed approach. *SN Comput Sci*. (2020) 1:212. doi: 10.1007/s42979-020-00225-9.

17. Bommasani R, Hudson DA, Adeli E, Altman R, Arora S, von Arx S, et al. On the opportunities and risks of foundation models. (2021). arXiv preprint arXiv:2108.07258.

18. Zhou C, Li Q, Li C, Yu J, Liu Y, Wang G, et al. A comprehensive survey on pretrained foundation models: A history from bert to chatgpt. (2023). arXiv preprint arXiv:2302.09419.

19. Devlin J, Chang MW, Lee K, Toutanova K. Bert: Pre-training of deep bidirectional transformers for language understanding. (2018). arXiv preprint arXiv:1810.04805.

20. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. (2018).

21. Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics, in: *International conference on machine learning*, . pp. 2256–65, PMLR.

22. Yang L, Zhang Z, Song Y, Hong S, Xu R, Zhao Y, et al. Diffusion models: A comprehensive survey of methods and applications. *ACM Computing Surveys*. (2022). doi: 10.1145/3626235.

23. Saharia C, Chan W, Saxena S, Li L, Whang J, Denton EL, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Adv Neural Inf Process Syst*. (2022) 35:36479–94.

24. Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. (2022). High-resolution image synthesis with latent diffusion models, in: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, . pp. 10684–95.

25. Li H, Yang Y, Chang M, Chen S, Feng H, Xu Z, et al. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*. (2022) 479:47–59. doi: 10.1016/j.neucom.2022.01.029.

26. Huang R, Zhao Z, Liu H, Liu J, Cui C, Ren Y. (2022). Prodiff: Progressive fast diffusion model for high-quality text-to-speech, in: *Proceedings of the 30th ACM International Conference on Multimedia*, . pp. 2595–605.

27. Singer U, Polyak A, Hayes T, Yin X, An J, Zhang S, et al. Make-a-video: Text-to-video generation without text-video data. (2022). arXiv preprint arXiv:2209.14792.

28. Xu J, Wang X, Cheng W, Cao YP, Shan Y, Qie X, et al. (2023). Dream3d: Zero-shot text-to-3d synthesis using 3d shape prior and text-to-image diffusion models, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, . pp. 20908–20918).

29. Kingma DP, Welling M. Auto-encoding variational bayes. (2013). arXiv preprint arXiv:1312.6114.

30. Singh A, Ogunfunmi T. An overview of variational autoencoders for source separation, finance, and bio-signal applications. *Entropy*. (2021) 24:55. doi: 10.3390/e24010055.

31. Cai L, Gao H, Ji S. Multi-stage variational auto-encoders for coarse-to-fine image generation. In: *Proceedings of the 2019 SIAM international conference on data mining*. Society for Industrial and Applied Mathematics (2019). p. 630–8.

32. Semeniuta S, Severyn A, Barth E. A hybrid convolutional variational autoencoder for text generation. (2017). arXiv preprint arXiv:1702.02390.

33. Guo R, Simpson I, Magnusson T, Kiefer C, Herremans D. A variational autoencoder for music generation controlled by tonal tension. (2020). arXiv preprint arXiv:2010.06230.

34. Yang Y, Bamler R, Mandt S. Improving inference for neural image compression. *Adv Neural Inf Process Syst*. (2020) 33:573–84.

35. Pol AA, Berger V, Germain C, Cerminara G, Pierini M. Anomaly detection with conditional variational autoencoders. In: *2019 18th IEEE international conference on machine learning and applications (ICMLA)*. IEEE (2019). p. 1651–7.

36. Nazabal A, Olmos PM, Ghahramani Z, Valera I. Handling incomplete heterogeneous data using vaes. *Pattern Recognition*. (2020) 107:107501. doi: 10.1016/j.patcog.2020.107501.

37. Mildenhall B, Srinivasan PP, Tancik M, Barron JT, Ramamoorthi R, Ng R. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun ACM*. (2021) 65:99–106. doi: 10.1145/3503250.

38. Gao K, Gao Y, He H, Lu D, Xu L, Li J. Nerf: Neural radiance field in 3d vision, a comprehensive review. (2022). arXiv preprint arXiv:2210.00379.

39. Bao C, Zhang Y, Yang B, Fan T, Yang Z, Bao H, et al. (2023). Sine: Semantic-driven image-based nerf editing with prior-guided editing field, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, . pp. 20919–29.

40. Corona-Figueroa A, Frawley J, Bond-Taylor S, Bethapudi S, Shum HP, Willcocks CG. Mednerf: Medical neural radiance fields for reconstructing 3d-aware ct-projections from a single x-ray. In: *2022 44th annual international conference of the IEEE engineering in medicine & Biology society (EMBC)*. IEEE (2022). p. 3843–8.

41. Rematas K, Liu A, Srinivasan PP, Barron JT, Tagliasacchi A, Funkhouser T, et al. (2022). Urban radiance fields, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, . pp. 12932–42.

42. Šlapak E, Pardo E, Dopiriak M, Maksymyuk T, Gazda J. Neural radiance fields in the industrial and robotics domain: applications, research opportunities and use cases. (2023). arXiv preprint arXiv:2308.07118.

43. Zhu Z, Chen Y, Wu Z, Hou C, Shi Y, Li C, et al. LATITUDE: robotic global localization with truncated dynamic low-pass filter in city-scale NeRF. In: *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE (2023). doi: 10.1109/ICRA48891.2023.10161570.

44. Li K, Rolff T, Schmidt S, Bacher R, Frintrop S, Leemans W, et al. Bringing instant neural graphics primitives to immersive virtual reality. In: *2023 IEEE conference on virtual reality and 3D user interfaces abstracts and workshops (VRW)*. IEEE (2023). p. 739–40.

45. Sun M, Yang D, Kou D, Jiang Y, Shan W, Yan Z, et al. Human 3d avatar modelling with implicit neural representation: A brief survey. In: *2022 14th international conference on signal processing systems (ICSPS)*. IEEE (2022). p. 818–27.

46. Zheng Z, Huang H, Yu T, Zhang H, Guo Y, Liu Y. (2022). Structured local radiance fields for human avatar modelling, in: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, . pp. 15893–903.

47. Hariri W. Unlocking the potential of chatGPT: A comprehensive exploration of its applications, advantages, limitations, and future directions in natural language processing. (2023). arXiv preprint arXiv:2304.02017.

48. Duwe G, Kim K. Out with the old and in with the new? An empirical comparison of supervised learning algorithms to predict recidivism. *Criminal Justice Policy Rev*. (2017) 28:570–600. doi: 10.1177/0887403415604899.

49. Tollenaar N, Van Der Heijden PG. Optimizing predictive performance of criminal recidivism models using registration data with binary and survival outcomes. *PloS One*. (2019) 14:e0213245. doi: 10.1371/journal.pone.0213245.

50. Ghasemi M, Anvari D, Atapour M, Stephen Wormith J, Stockdale KC, Spiteri RJ. The application of machine learning to a general risk–need assessment instrument in the prediction of criminal recidivism. *Criminal Justice Behav*. (2021) 48:518–38. doi: 10.1177/0093854820969753

51. Singh A, Mohapatra S. Development of risk assessment framework for first time offenders using ensemble learning. *IEEE Access*. (2021) 9:135024–33. doi: 10.1109/ACCESS.2021.3116205

52. Travaini GV, Pacchioni F, Bellumore S, Bosia M, De Micco F. Machine learning and criminal justice: A systematic review of advanced methodology for recidivism risk prediction. *Int J Environ Res Public Health*. (2022) 19:10594. doi: 10.3390/ijerph191710594.

53. Trinhammer ML, Merrild AH, Lotz JF, Makransky G. (2022). Predicting crime during or after psychiatric care: Evaluating machine learning for risk assessment using the Danish patient registries. *J Psychiat Res*. 152:194–200.

54. Watts D, Moulden H, Mamak M, Upfold C, Chaimowitz G, Kapczinski F. (2021). Predicting offenses among individuals with psychiatric disorders-A machine learning approach. *J Psychiat Res*. 138:146–54.

55. Suchting R, Green CE, Glazier SM, Lane SD. A data science approach to predicting patient aggressive events in a psychiatric hospital. *Psychiatry Res*. (2018) 268:217–22. doi: 10.1016/j.psychres.2018.07.004.

56. Menger V, Spruit M, Van Est R, Nap E, Scheepers F. Machine learning approach to inpatient violence risk assessment using routinely collected clinical notes in electronic health records. *JAMA network Open*. (2019) 2:e196709–e196709. doi: 10.1001/jamanetworkopen.2019.6709.

57. Wang KZ, Bani-Fatemi A, Adanty C, Harripaul R, Griffiths J, Kolla N, et al. Prediction of physical violence in schizophrenia with machine learning algorithms. *Psychiatry Res*. (2020) 289:112960. doi: 10.1016/j.psychres.2020.112960.

58. Gou N, Xiang Y, Zhou J, Zhang S, Zhong S, Lu J, et al. Identification of violent patients with schizophrenia using a hybrid machine learning approach at the individual level. *Psychiatry Res*. (2021) 306:114294. doi: 10.1016/j.psychres.2021.114294.

59. Hofmann LA, Lau S, Kirchebner J. Advantages of machine learning in forensic psychiatric research—uncovering the complexities of aggressive behaviour in schizophrenia. *Appl Sci*. (2022) 12:819. doi: 10.3390/app12020819.

60. Yu T, Zhang X, Liu X, Xu C, Deng C. The prediction and influential factors of violence in male schizophrenia patients with machine learning algorithms. *Front Psychiatry*. (2022) 13:799899. doi: 10.3389/fpsyt.2022.799899.

61. Large M, Nielssen O. The limitations and future of violence risk assessment. *World Psychiatry*. (2017) 16:25. doi: 10.1002/wps.20394.

62. Chekroud AM, Bondar J, Delgadillo J, Doherty G, Wasil A, Fokkema M, et al. The promise of machine learning in predicting treatment outcomes in psychiatry. *World Psychiatry*. (2021) 20:154–70. doi: 10.1002/wps.20882.

63. Tortora L, Meynen G, Bijlsma J, Tronci E, Ferracuti S. Neuroprediction and ai in forensic psychiatry and criminal justice: A neurolaw perspective. *Front Psychol*. (2020) 11:220. doi: 10.3389/fpsyg.2020.00220.

64. George A, Johnson D, Carenini G, Eslami A, Ng R, Portales-Casamar E. Applications of aspect-based sentiment analysis on psychiatric clinical notes to study suicide in youth. *AMIA Summits Trans Sci Proc*. (2021) 2021:229.

65. Tutun S, Johnson ME, Ahmed A, Albizri A, Irgil S, Yesilkaya I, et al. An AI-based decision support system for predicting mental health disorders. *Inf Syst Front*. (2023) 25:1261–76. doi: 10.1007/s10796-022-10282-5.

66. Constantinou AC, Freestone M, Marsh W, Coid J. Causal inference for violence risk management and decision support in forensic psychiatry. *Decision Support Syst*. (2015) 80:42–55. doi: 10.1016/j.dss.2015.09.006.

67. Angwin J, Larson J, Mattu S, Kirchner L. (2016) Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. 23:77–91.

68. Berk R, Heidari H, Jabbari S, Kearns M, Roth A. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods Res*. (2021) 50:3–44. doi: 10.1177/0049124118782533.

69. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*. (2019) 366:447–53. doi: 10.1126/science.aax2342.

70. Alikhademi K, Drobina E, Prioleau D, Richardson B, Purves D, Gilbert JE. A review of predictive policing from the perspective of fairness. *Artif Intell Law*. (2022), 1–17. doi: 10.1007/s10506-021-09286-4.

71. Dimitri GM, Spasov S, Duggento A, Passamonti L, Lió P, Toschi N. Multimodal and multicontrast image fusion via deep generative models. *Inf Fusion*. (2022) 88:146–60. doi: 10.1016/j.inffus.2022.07.017.

72. Huang J, Neill L, Wittbrodt M, Melnick D, Klug M, Thompson M, et al. Generative artificial intelligence for chest radiograph interpretation in the emergency department. *JAMA network Open*. (2023) 6:e2336100–e2336100. doi: 10.1001/jamanetworkopen.2023.36100.

73. Ceccarelli F, Mahmoud M. Multimodal temporal machine learning for Bipolar Disorder and Depression Recognition. *Pattern Anal Appl*. (2022) 25:493–504. doi: 10.1007/s10044-021-01001-y.

74. Zhang P, Kamel Boulos MN. Generative AI in medicine and healthcare: promises, opportunities and challenges. *Future Internet*. (2023) 15:286. doi: 10.3390/fi15090286.

75. Bordukova M, Makarov N, Rodriguez-Esteban R, Schmich F, Menden MP. Generative artificial intelligence empowers digital twins in drug discovery and clinical trials. *Expert Opin Drug Discovery*. (2023), 1–10. doi: 10.1080/17460441.2023.2273839.

76. Liu R, Huang ZA, Hu Y, Zhu Z, Wong KC, Tan KC. Attention-like multimodality fusion with data augmentation for diagnosis of mental disorders using MRI. *IEEE Trans Neural Networks Learn Syst*. (2022). doi: 10.1109/TNNLS.2022.3219551.

77. Saadatinia M, Salimi-Badr A. An explainable deep learning-based method for schizophrenia diagnosis using generative data-augmentation. (2023). arXiv preprint arXiv:2310.16867.

78. Birhane A, Kasirzadeh A, Leslie D, Wachter S. Science in the age of large language models. *Nat Rev Phys*. (2023), 1–4. doi: 10.1038/s42254-023-00581-4.

79. Bender EM, Gebru T, McMillan-Major A, Shmitchell S. (2021). On the dangers of stochastic parrots: Can language models be too big?, in: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, . pp. 610–23.

80. Salewski L, Alaniz S, Rio-Torto I, Schulz E, Akata Z. In-context impersonation reveals large language models' Strengths and biases. (2023). arXiv preprint arXiv:2305.14930.

81. Thakur V. Unveiling gender bias in terms of profession across LLMs: Analysing and addressing sociological implications. (2023). arXiv preprint arXiv:2307.09162.

82. Ahn J, Oh A. Mitigating language-dependent ethnic bias in BERT. (2021). arXiv preprint arXiv:2109.05704.

83. Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. *Adv Neural Inf Process Syst*. (2020) 33:1877–901. doi: 10.48550/arXiv.2005.14165

84. Muralidhar D. (2021). Examining religion bias in AI text generators, in: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, . pp. 273–4.

85. Venkit PN, Srinath M, Wilson S. (2022). A study of implicit bias in pretrained language models against people with disabilities, in: *Proceedings of the 29th International Conference on Computational Linguistics*, . pp. 1324–32.

86. Jigsaw K. Jigsaw unintended bias in toxicity classification. (2019).

87. Hutchinson B, Prabhakaran V, Denton E, Webster K, Zhong Y, Denuyl S. Social biases in NLP models as barriers for persons with disabilities. (2020). arXiv preprint arXiv:2005.00813.

88. Magee L, Ghahremanlou L, Soldatic K, Robertson S. Intersectional bias in causal language models. (2021). arXiv preprint arXiv:2107.07691.

89. Sun L, Wei M, Sun Y, Suh YJ, Shen L, Yang S. Smiling women pitching down: auditing representational and presentational gender biases in image generative AI. (2023). arXiv preprint arXiv:2305.10566.

90. Ungless EL, Ross B, Lauscher A. Stereotypes and smut: the (Mis) representation of non-cisgender identities by text-to-image models. (2023). arXiv preprint arXiv:2305.17072.

91. Hutchinson B, Baldridge J, Prabhakaran V. Under specification in scene description-to-depiction tasks. (2022). arXiv preprint arXiv:2210.05815.

92. Mandal A, Little S, Leavy S. Gender bias in multimodal models: A transnational feminist approach considering geographical region and culture. (2023). arXiv preprint arXiv:2309.04997.

93. Bianchi F, Kalluri P, Durmus E, Ladhak F, Cheng M, Nozza D, et al. (2023). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale, in: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, . pp. 1493–504.

94. Mezey G, Youngman H, Kretzschmar I, White S. Stigma and discrimination in mentally disordered offender patients–A comparison with a non-forensic population. *J Forensic Psychiatry Psychol*. (2016) 27:517–29. doi: 10.1080/14789949.2016.1172658.

95. Jorm AF, Reavley NJ, Ross AM. Belief in the dangerousness of people with mental disorders: a review. *Aust New Z J Psychiatry*. (2012) 46:1029–45. doi: 10.1177/0004867412442406.

96. Steiger S, Moeller J, Sowislo JF, Lieb R, Lang UE, Huber CG. Approval of coercion in psychiatry in public perception and the role of stigmatization. *Front Psychiatry*. (2022) 12:819573. doi: 10.3389/fpsyt.2021.819573.

97. Assari S, Miller RJ, Taylor RJ, Mouzon D, Keith V, Chatters LM. Discrimination fully mediates the effects of incarceration history on depressive symptoms and psychological distress among African American men. *J Racial Ethnic Health Disparities*. (2018) 5:243–52. doi: 10.1007/s40615-017-0364-y.

98. Perry BL, Neltner M, Allen T. A paradox of bias: Racial differences in forensic psychiatric diagnosis and determinations of criminal responsibility. *Race Soc problems*. (2013) 5:239–49. doi: 10.1007/s12552-013-9100-3.

99. West ML, Yanos PT, Mulay AL. Triple stigma of forensic psychiatric patients: Mental illness, race, and criminal history. *Int J Forensic Ment Health*. (2014) 13:75–90. doi: 10.1080/14999013.2014.885471.

100. EagleAI. Forensic sketch AIrtist (2022). Available online at: https://lablab.ai/event/openai-whisper-gpt3-codex-dalle2-hackathon/eagleai/forensic-sketch-airtist (Accessed 20 September 2023).

101. Liesenfeld A, Lopez A, Dingemanse M. (2023). Opening up ChatGPT: Tracking openness, transparency, and accountability in instruction-tuned text generators, in: *Proceedings of the 5th International Conference on Conversational User Interfaces*, . pp. 1–6.

102. Ahmed N, Wahed M, Thompson NC. The growing influence of industry in AI research. *Science*. (2023) 379:884–6. doi: 10.1126/science.ade2420.

103. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. *NPJ Digital Med*. (2023) 6:120. doi: 10.1038/s41746-023-00873-0.

104. Luccioni AS, Viviano JD. What's in the box? A preliminary analysis of undesirable content in the common crawl corpus. (2021). arXiv preprint arXiv:2105.02732.

105. Huang J, Shao H, Chang KCC. Are large pre-trained language models leaking your personal information? (2022). arXiv preprint arXiv:2205.12628.

106. Gipson Rankin SM. Technological tethereds: potential impact of untrustworthy artificial intelligence in criminal justice risk assessment instruments. *Wash. Lee L. Rev*. (2021) 78:647. doi: 10.2139/ssrn.3662761.

107. Smits J, Borghuis T. Generative AI and intellectual property rights. In: *Law and artificial intelligence: regulating AI and applying AI in legal practice*. TMC Asser Press, The Hague (2022). p. 323–44.

108. Strowel A. ChatGPT and generative AI tools: theft of intellectual labor? *IIC-International Rev Intellectual Property Competition Law*. (2023) 54:491–4. doi: 10.1007/s40319-023-01321-y.

109. Sheng E. In generative AI legal wild west, the courtroom battles are just getting started (2023). CNBC. Available online at: https://www.cnbc.com/2023/04/03/in-generative-ai-legal-wild-west-lawsuits-are-just-getting-started (Accessed 12 October 2023).

110. Hogan NR, Davidge EQ, Corabian G. On the ethics and practicalities of artificial intelligence, risk assessment, and race. *J Am Acad Psychiatry Law*. (2021) 49:326–34. doi: 10.29158/JAAPL.200116-20

111. Starke G, Schmidt B, De Clercq E, Elger BS. Explainability as fig leaf? An exploration of experts' ethical expectations towards machine learning in psychiatry. *AI Ethics*. (2023) 3:303–14. doi: 10.1007/s43681-022-00177-1.

112. Gundersen T, Bærøe K. Ethical algorithmic advice: Some reasons to pause and think twice. *Am J Bioethics*. (2022) 22:26–8. doi: 10.1080/15265161.2022.2075053.

113. Kidd C, Birhane A. How AI can distort human beliefs. *Science*. (2023) 380:1222–3. doi: 10.1126/science.adi0248.

114. Farah H. *Court of appeal judge praises 'jolly useful' ChatGPT after asking it for legal summary* (2023). The Guardian (Accessed 5 October 2023).

115. Gutiérrez JD. *ChatGPT in Colombian Courts: Why we need to have a conversation about the digital literacy of the judiciary* (2023). Verfassungsblog (Accessed 5 October 2023).

116. Jamal S. *Pakistani judge uses ChatGPT to make court decision* (2023). Gulf News (Accessed 6 October 2023).

117. Acres T. *Lawyers used ChatGPT to help with a case - it backfired massively* (2023). Sky News (Accessed 5 October 2023).

118. Alkaissi H, McFarlane SI. Artificial Hallucinations in ChatGPT: Implications in Scientific Writing. *Cureus*. (2023) 15(2):e35179. doi: 10.7759/cureus.35179.

119. McGowan A, Gui Y, Dobbs M, Shuster S, Cotter M, Selloni A, et al. ChatGPT and Bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry Res*. (2023) 326:115334. doi: 10.1016/j.psychres.2023.115334.

120. Emsley R. ChatGPT: these are not hallucinations–they're fabrications and falsifications. *Schizophrenia*. (2023) 9:52. doi: 10.1038/s41537-023-00379-4.

121. Herz J. *AI-generated party pics look eerily real — unless you can spot these tells* (2023). New York Post (Accessed 10 October 2023).

122. Grossman MR, Grimm PW, Brown D, Xu M. The GPTJudge: justice in a generative AI world. *Duke Law Technol Rev*. (2023) 23.

123. Delfino RA. Deepfakes on trial: a call to expand the trial judge's gatekeeping role to protect legal proceedings from technological fakery. *Hastings LJ*. (2022) 74:293. doi: 10.2139/ssrn.4032094.

124. Chesney B, Citron D. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev*. (2019) 107:1753. doi: 10.2139/ssrn.3213954.

125. Elkhatat AM, Elsaid K, Almeer S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *Int J Educ Integrity*. (2023) 19:17. doi: 10.1007/s40979-023-00140-5.

126. Scarpazza C, Ferracuti S, Miolla A, Sartori G. The charm of structural neuroimaging in insanity evaluations: guidelines to avoid misinterpretation of the findings. *Trans Psychiatry*. (2018) 8:227. doi: 10.1038/s41398-018-0274-8.

127. Morgan J. Wrongful convictions and claims of false or misleading forensic evidence. *J Forensic Sci*. (2023) 68:908–61. doi: 10.1111/1556-4029.15233.

128. Hacker P. Sustainable AI regulation. (2023). arXiv preprint arXiv:2306.00292.

129. Cowls J, Tsamados A, Taddeo M, Floridi L. The AI gambit: leveraging artificial intelligence to combat climate change—opportunities, challenges, and recommendations. *Ai Soc*. (2021), 1–25. doi: 10.2139/ssrn.3804983.

130. Taddeo M, Tsamados A, Cowls J, Floridi L. Artificial intelligence and the climate emergency: opportunities, challenges, and recommendations. *One Earth*. (2021) 4:776–9. doi: 10.1016/j.oneear.2021.05.018.

131. Tamburrini G. The AI carbon footprint and responsibilities of AI scientists. *Philosophies*. (2022) 7:4. doi: 10.3390/philosophies7010004.

132. Walker R, Dillard-Wright J, Iradukunda F. Algorithmic bias in artificial intelligence is a problem—And the root issue is power. *Nurs Outlook*. (2023) 71:102023. doi: 10.1016/j.outlook.2023.102023.

133. Mittelstadt BD, Allo P, Taddeo M, Wachter S, Floridi L. The ethics of algorithms: Mapping the debate. *Big Data Soc*. (2016) 3:2053951716679679. doi: 10.1177/2053951716679679

134. Markovski Y. *How your data is used to improve model performance* (2023). OpenAI (Accessed 15 October 2023).

135. Mittelstadt BD, Floridi L. The ethics of big data: current and foreseeable issues in biomedical contexts. *Ethics Biomed Big Data*. (2016), 445–80. doi: 10.1007/978-3-319-33525-4.

American Counseling Association

Recommendations For Practicing Counselors And Their Use Of AI

AI Work Group Recommendations

The American Counseling Association has convened a panel of counseling experts representing academia, private practice and students to comprise its AI Work Group. The work group used research-based and contextual evidence; the ACA Code of Ethics; and clinical knowledge and skill to develop the following recommendations. The goal is to both prioritize client well-being, preferences, and values in the advent and application of AI, while informing counselors, counselor-educators and clients about the use of AI today. The recommendations also highlight the additional research needed to inform counseling practice as AI becomes a more widely available and accepted part of mental health care.

Recommendations for:

- Practicing Counselors
- Client Use
- Assessment, and Diagnosis
- Faculty
- Students

Further Recommendations

The integration of Artificial Intelligence (AI), including machine learning and natural language generation, in counseling practice offers significant potential for enhancing efficiency and effectiveness. Embracing AI in counseling practice has the potential to revolutionize the field, increasing efficiency and effectiveness while maintaining ethical standards. By following these guidelines, counselors can navigate the integration of AI in a way that maximizes benefits for both clients and practitioners, all while upholding the highest standards of professional conduct. Remember, the ultimate goal is to enhance the quality of care and support provided to clients.

Methods

Task force recommendations are based on integrating the following factors, considered central to evidence-based practice.

1. Research-based and contextual evidence
2. Client preferences and values
3. The ACA Code of Ethics
4. Clinical knowledge and skill

Additionally, as the field of counseling extends to counselor education, recommendations are provided for educators in their use of AI.

AI impacts counseling and extends beyond the field. The origins, implications, and body research of AI encompass many disciplines. Consequently, the task force reviewed an interdisciplinary array of sources to extrapolate recommendations.

Our working definition of Artificial Intelligence entails computers simulating human intelligence. The simulation involves the completion of tasks resembling those carried out by human intelligence, including reasoning, language comprehension, problem-solving, and decision-making (Sheikh et al., 2023).

Recommendations

Recommendation: Learn more about the essentials of artificial intelligence, its subfields, and its applications to mental health.

Counselors are required to practice within their boundaries of competence (C.2.a). To prepare for work with AI, counselors should learn about AI through three levels of understanding.

1. The essentials, including algorithms and how AI shows up in daily life, such as in social media, marketing campaigns, and in smartphones.
2. AI subfields, such as machine learning, neural networks, natural language processing, computer vision, and robotics. For example, counselors can learn about large language models (LLMs) and their use in machine learning. LLMs are used in "generative AI," such as ChatGPT.
3. Applications. Research suggests that AI is currently applied to mental health in three main ways: (1) "personal sensing" (or "digital phenotyping"), (2) natural language processing, and (3) through chatbots (D'Alfonso, 2020).

Recommendation: Stay open, informed, and educated.

Remain open to technological advances that can improve professional practice. Efficiencies that can reduce the administrative burden on practitioners are not automatically unethical. Evaluate technologies critically before incorporating for practice. Stay updated on the latest developments, ethical standards, and best practices related to AI in counseling.

Recommendation: Avoid over-reliance on AI.

While AI can enhance efficiency, it should not replace the essential human element in counseling. Maintain a balanced approach, ensuring that the therapeutic relationship remains central.

Recommendation: Recognize that AI may contain bias and be capable of discrimination.

AI is imperfect. The perception among users of AI may be that AI does not judge or discriminate, and while this is true in the sense that an AI is essentially a computer system, the perception is false in the sense that the output of AI, such as the text, audio, or visuals, that it generates, may show bias. Unfortunately, AI still has a diversity and inclusion problem

(Fulmer et al., 2021). The algorithms and training models that comprise the AI may lack the likes of racial, ethnic, religious, and gender inclusion. Recognize that with all its capabilities and potential, AI may be biased and thus, cause harm.

Recommendation: Career counselors and those who address employment issues should stay informed about how automation is shaping the world of work.

Counseling history is steeped in the vocational guidance movement of the early 20th century. Change has been a constant from the time of Frank Parsons to today, but AI presents special challenges for career counseling. For centuries, technology has rendered some jobs obsolete and created new ones. The future state of the job market in relation to AI is unclear, including whether it will yield a surplus, remain relatively balanced, or result in a deficit of jobs. An analysis from the Brookings Institute suggests that automation may disrupt some industries, and though mass unemployment is unlikely in the near future, worker transitions may grow more common (Bessen et al., 2020). Counselors should explore career developmental models that account for rapid changes in the labor force.

Recommendation: Advocate for transparency in AI algorithms.

A transparent AI algorithm is one that is open to inspection by someone other than the developer (Yudkowsky & Bostrom, 2011). Counselors could be part of inspection teams, helping to ensure that AI is built fairly and is comprehensible, and then relaying their findings back to the counseling community. Transparent AI includes three factors: Accessibility, Interpretability, and Controlled Maintenance (Fulmer et al., 2021). Accessibility means that the AI should be available and responsive to the people for whom it was intended. Interpretability concerns the output of an AI, which must be easy to understand, user-friendly, and clear from the beginning that it is indeed, an AI. Relatedly, AI should inform users of whether a human is available for support. Finally, an AI must adapt and evolve by receiving regular improvements, which should come from client, counselor, and programmer-informed feedback.

Recommendation: Maintain transparency and informed consent.

Counselors should clearly inform clients about the use of AI

tools in their counseling process, explaining their purpose and potential benefits (H.2.a). Obtain explicit informed consent from clients for the use of AI-assisted tools, ensuring they understand the implications and potential impact on their treatment.

Recommendation: Leverage AI for data-driven insights.
Employ AI tools for data analysis to gain valuable insights from anonymized and aggregated client data to evaluate and inform evidence-based treatment approaches and interventions. Continuously evaluate the accuracy and appropriateness of AI-generated analytics and content, especially in cases where it interacts directly with clients (e.g., chatbots). Intervene when necessary to correct or modify responses.

Recommendation: Ensure data security and privacy.
AI platforms designed for counseling services and training purposes should prioritize data security and privacy from the outset, incorporating the principles of "Privacy by Design". This approach ensures that personal identifying information and personal health information are protected throughout the system's lifecycle. Furthermore, AI platforms must adhere to the standards set by applicable local privacy laws and regulations (e.g., HIPAA in the United States; H.1.b). By embedding privacy considerations into the design and operation of AI systems, providers can ensure the secure and confidential handling of sensitive data, fostering trust and compliance in their use for counseling services.

Recommendation: Counselors should empower clients to communicate about their AI use. According to the ACA Code of Ethics (2014), empowering a diverse array of clients is a central mission of the counseling profession. Counselors should empower their clients to communicate the use of AI tools to support their mental health with their counselor, as this information can help the counselor understand the client's approach to mental health. The counselor may be able to guide the client on how to use AI tools safely and effectively.

Recommendation: Supervisors can use AI to enhance the development of supervisees.
The supervisory relationship is key in supporting the development of new counselors (Borders & Brown, 2022). The supervisory relationship should include discussion of AI

counseling tools and counseling supervision can be enhanced by the use of AI tools. Counseling supervisors should be aware that AI tools for monitoring supervisees' work exist and can make suggestions for how counselors can improve their work. Supervisors may want to use AI tools to read transcripts of their supervisees' sessions or to analyze the use of counseling skills by supervisees in session.

Recommendation: Counselors must understand the limitations of AI in diagnosis and assessment in all counseling settings.
Counselors should refrain from using AI as the sole tool for diagnosis and assessment in counseling. Although AI can be a supportive tool to inform a counselor's professional judgment, counselors must attain adequate training to understand the limitations and the use of AI in clinical settings. This approach is aligned with the ACA Code of Ethics (C.2.), which emphasizes the importance of professional competence and judgment in clinical decision-making. Counselors must critically evaluate AI-assisted diagnostic suggestions and incorporate their clinical expertise, understanding of the client's history, and cultural context to ensure a comprehensive and ethically sound assessment. This recommendation supports the responsible and client-centered use of AI in counseling.

Recommendation: Consider conducting research on the intersection of AI and counseling.
A dearth of research currently exists on how AI can and may impact counseling (Fulmer, 2019). AI shows potential to influence several areas of clinical practice (e.g., diagnosis, practice management, automating documentation), counselor education, and research approaches; thus, more research is needed to discover AI's potential in these areas. Future research should also focus on the ethics of using AI and extend our understanding of the impacts of advanced technologies, including AI, on diverse populations. Counselors and counseling researchers are encouraged to take up the charge to conduct research that transforms counseling practices and training for better client care and wellness. There are three significant benefits of counselor-led research. One, to help ensure that the incorporation of AI into practice is substantiated by research. Two, to help the counseling field lead the way as AI enters client care, broadly speaking. Third, to inform the public about

AI's efficacy and capabilities in providing counseling services and mental health support.

Selected Publications and References

American Counseling Association (2014). ACA Code of Ethics. Alexandria, VA: Author.

Bessen, J., Goos, M., Salomons, A., & van den Berge, W. (2020). Automation: A guide for policymakers. *Economic Studies at Brookings Institution: Washington, DC, USA*.

Borders, L. D., & Brown, L. L. (2022). The new handbook of counseling supervision

D'Alfonso, S. (2020). AI in mental health. *Current Opinion in Psychology*, *36*, 112-117.

Fulmer, R., Davis, T., Costello, C., & Joerin, A. (2021). The ethics of psychological artificial intelligence: Clinical considerations. *Counseling and Values*, *66*(2), 131-144.

Fulmer, R. (2019). Artificial intelligence and counseling: Four levels of implementation. *Theory & Psychology*, *29*(6), 807-819. https://doi.org/10.1177/0959354319853045

Minerva, F., & Giubilini, A. (2023). Is AI the Future of Mental Healthcare?. *Topoi : An international review of philosophy*, *42*(3), 1–9.

Sheikh, H., Prins, C., & Schrijvers, E. (2023). Artificial Intelligence: Definition and Background. In *Mission AI: The New System Technology* (pp. 15-41). Cham: Springer International Publishing.

Yudkowsky, E., & Bostrom, N. (2011). The ethics of artificial intelligence. In K. Frankish & W. Ramsey (Eds.), *The Cambridge handbook of artificial intelligence* (pp. 316–334). Cambridge University Press.

AI Work Group Members

S. Kent Butler, PhD University of Central Florida	Chip Flater Amercan Counseling Association	Morgan Stohlman Kent State University
Fallon Calandriello, PhD Northwestern University	Russell Fulmer, PhD Husson University	Olivia Uwamahoro Williams, PhD College of William and Mary
Wendell Callahan, PhD University of San Diego	Marcelle Giovannetti, EdD Messiah University-Mechanicsburg, PA	Yusen Zhai, PhD UAB School of Education
Lauren Epshteyn Northwestern University	Marty Jencius, PhD Kent State University	
Dania Fakhro, PhD University of North Carolina, Charlotte	Sidney Shaw, EdD Walden University	

2461 Eisenhower Avenue, Suite 300, Alexandria, Va. 22314
| 800-347-6647 | (fax) 800-473-2329