Meadows Mental Health Policy Institute

Mental Health and AI: An Introductory Primer

Introduction to AI & Mental Health

Artificial intelligence (AI), often referred to in the medical community as augmented intelligence, is one of the most transformative technological advancements of our time. AI has revolutionized the way we interact with technology, make decisions, and solve complex problems. It permeates various aspects of our daily lives, from voice assistants on our smartphones to email classification to recommended ads. AI is increasingly used in both physical and mental health care —understanding it is essential to understanding the future of healthcare.

AI will continue to transform all aspects of life, reshaping how we work, live, and interact with the world. Yet, there is confusion and uncertainty about how AI technologies work. Developing a foundational understanding of AI—including different subfields, current applications, sources of bias in AI, and a history of the field—can foster the understanding and confidence needed for people to more effectively engage in conversations and decision-making about its design, implementation and regulation, and facilitate that AI benefits society as a whole.

What is AI?

Artificial intelligence is an umbrella term for the development of computer systems and algorithms that can perform tasks commonly requiring human intelligence.[i] The term augmented intelligence (herein, AI), is gaining popularity, particularly in the healthcare space. The American Medical Association uses the term to focus "on AI's assistive role," emphasizing that its design enhances human intelligence rather than replaces it.[ii]

The term artificial intelligence was first coined in the 1950s and has been developed ever since, but the growing use of AI in healthcare and popular applications like ChatGPT and smart driving cars have brought increasing attention to its use in society.[iii]

Subfields of Artificial Intelligence

Though the term AI is often utilized broadly, AI encompasses various subfields, each focusing on specific aspects and applications. Understanding the subfields of AI is key to understanding how these technologies work.

Machine learning (ML) is a broad subfield within AI that "uses computer algorithms to analyze data and make intelligent decisions based on what it has learned, *without* being specifically programmed."[iv] Deep learning is a subfield within machine learning that uses neural networks

(computation systems inspired by the brain) with many "layers" (deep neural networks) to learn complex representations from data. Finally, generative AI is another subfield of AI that utilizes deep learning to "create, generate, and simulate new content."[v] Figure 1 below demonstrates the hierarchy of AI and its subfields, and Table 1 below lays out information about each category and its applications in mental health.

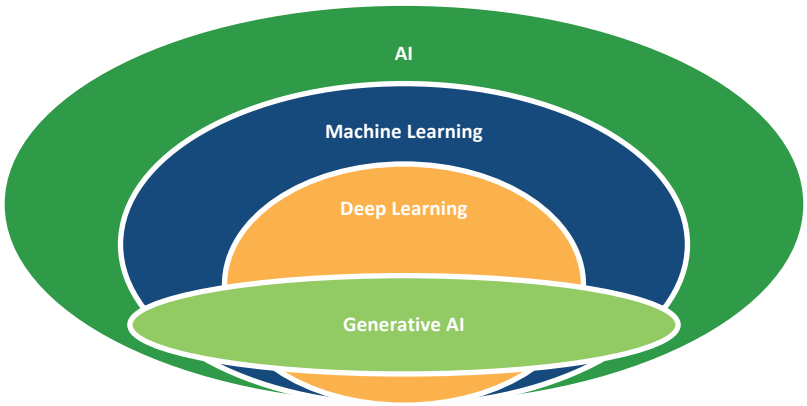

Figure 1. Subfields of AI

Table 1: Subfields of AI

Subfields of AI		
Category	**Definition**	**Mental Health Application**
Machine Learning	A broad subfield within AI that uses computer algorithms to analyze data and predict intelligent decisions based on data sets from the past *without* being specifically programmed.[vi]	Using EHR data and clinical notes to predict mental health crises to help healthcare workers manage caseload priorities and pre-emptively intervene, reducing the need for more intensive, higher-cost interventions[vii]
Deep Learning	Deep learning uses neural networks with many layers of algorithms (deep neural networks) to learn complex representations from data. Deep learning is also commonly a ML approach and architecture used for computer vision and natural language processing.	Analysis of genetics and genomics data for understanding mental health conditions
Generative AI	Generative models aim to create new data that resembles a given dataset that the model has been trained on, often used in content generation and data augmentation.	Creating virtual therapists to support individuals with mental health conditions in between therapy sessions

Categories or subfields of AI are often based on how a model is trained. Training a model involves using algorithms to develop parameters of a model by showing an algorithm real life data.[viii] The three main subfields of machine learning include supervised learning, unsupervised learning, and reinforcement learning.[ix] Table 2 below provides information about each subfield and its mental health applications.

Table 2. Subfields of Machine Learning

Subfields of Machine Learning		
Category	**Definition**	**Mental Health Application**
Supervised learning	Supervised learning involves training a model on a collection of input variables and output variables and then applying the mapping the model learns to predict the outcomes of unseen data.	Predicting the onset and severity of depression from wearable and smartphone data
Unsupervised Learning	Unsupervised learning aims to find patterns or structures within data that are not labeled by humans. It is often used for clustering where similar data points are grouped together.	K-means clustering to identify student anxiety profiles and potential subtypes
Reinforcement Learning	Reinforcement learning involves training "through trial and error to take the best action by establishing a reward system."	Measuring anhedonia by comparing how someone with depression is playing a game versus the way a machine would play the game to maximize the reward

These subfields can be applied across different areas and are based on the type of data that an AI model is using or producing. Key areas of application encompass language, vision, and audio/speech, outlined in Table 3 below[x]

Table 3: AI Application Areas

Major Areas of Application in AI			
Category	**Definition**	**Examples**	**Mental Health Application**
Natural Language Processing (NLP)	NLP focuses on enabling computers to understand, interpret, and generate human language text or speech. It uses machine learning and deep learning.	Sentiment analysis (e.g., analyzing product reviews to determine if text is positive, negative, or neutral), machine translation (e.g., Google	Early detection of mental health conditions through analyzing text-based data, such as social media posts, electronic health records, or personal diaries

Major Areas of Application in AI			
		Translate), chatbots, and virtual assistants (e.g., Siri, Alexa)	
Computer Vision	Computer vision focuses on enabling machines to interpret and process visual information from the world, such as images or videos.	Object detection (e.g., self-driving cars identifying pedestrians), facial recognition, medical image analysis, and augmented reality applications (e.g., object recognition in the physical world to overlay on digital content)	Analyzing photos and videos to analyze facial expressions and gestures to infer emotional states, or some disorders such as autism spectrum disorder[xi]
Speech-to-text or speech synthesis	Can include speech-to-text, which involves identifying common patterns in the different pronunciations of a word, mapping new voice samples to corresponding words, or speech synthesis, which enables computers to make natural sounding voice models	Voice-activated searches (Alexa, Siri, etc.); real-time transcription; voice cloning	Clinical notes in therapy sessions; Using speech synthesis to create synthetic voices for ALS patients

The subfields within machine learning and AI represent the diverse range of applications and approaches that make up the ever-expanding field of AI. When conversations come up around applications of AI and ML, it is beneficial to determine what subfield(s) of AI are being discussed. Understanding the subfield(s) of AI used for a certain technological application can enable a more in-depth understanding of current capabilities, potential sources of bias that need to be mitigated, and future directions.

Important Considerations When Using AI

While AI has immense potential to improve mental healthcare monitoring and treatment, it also holds ethical and safety challenges. Careful consideration of specific evaluations is needed. Standford University's Human-Centered Artificial Intelligence Center recently developed the Readiness Evaluation for AI Deployment and Implementation for Mental Health (READI) framework. The framework provides a starting point for evaluating whether AI-mental health applications are ready for clinical deployment. Below is a brief overview of the main categories of considerations of the framework and a few points within each consideration.

Table 4: READI Framework Overview

READI Framework Overview	
Safety	- Should not promote "dangerous or unhealthy human behaviors" (e.g. self-harm, suicide, substance use, etc.) - Should not promote unhealthy thinking patterns - Must have a mechanism to monitor and report adverse events
Privacy and Confidentiality	- Patient information must be safeguarded at the same level as HIPPA (even if HIPPA doesn't currently extend to digital applications) - Patient health data should not be disclosed without explicit consent - Users should be notified of data breaches - Terms of service should include how data is used, and opt-out options should be provided
Equity	- Potential for bias and methods used to debias, such as using a manual for culturally-competent cognitive behavioral therapy (CBT) in AI training, should be disclosed - Report demographics of people whose data have been used, and tested with representative end-users - Engagement, effectiveness, and satisfaction data should be supported across demographics - Integrate culturally-competent and responsive practices into applications
Engagement	- Need to create sufficient engagement without promoting overuse or unhealthy use of technology - Appropriate engagement levels may vary from person to person
Effectiveness	- Should have evidence of effectiveness in clinically representative settings - Key effectiveness outcomes should include a decrease in symptoms and functional impairment and an increase in well-being and quality of life - Study outcomes should be reported in a way that supports "informed decision-making about the applicability and appropriateness of the use or deployment of the application"
Implementation Considerations	- Should be able to deploy in routine care settings (e.g., work with EHR and clinical workflows) - Collect data on feasibility, acceptability, compatibility, and perceptions of AI mental health application with the healthcare system

Bias in Artificial Intelligence

One of the most pressing ethical and fairness issues in AI today is the pervasive problem of bias. Bias in AI stems from a multitude of factors, such as over or underrepresentation of certain groups in data sets. AI systems can perpetuate or even exacerbate existing inequalities in areas like prior authorizations and other benefit allocations, risk prediction in areas such as medical need, criminal offense, or hiring, lending, and criminal justice. For example, AI systems trained on data from specific socioeconomic groups may exhibit biases against individuals from different backgrounds. For instance, credit scoring algorithms may favor individuals from higher-income neighborhoods, disadvantaging those from lower-income communities, even when they have strong financial records.[xii] AI technologies, such as wearables to detect heart rates and pulse oximeters that measure oxygen levels in the blood to determine treatment, have been repeatedly shown to be less accurate on people with darker skin.[xiiixiv]

Often, these biases in AI systems are unintentional and arise from complex interactions within an AI model. Understanding and mitigating unintended bias is a challenging but critical aspect of AI ethics. The first step is to understand some of the ways that bias embeds itself in AI. Some examples include:

- **Confirmation Bias:** AI algorithms can reinforce users' existing beliefs and preferences by recommending content that aligns with their viewpoints. This can contribute to the spread of misinformation and filter bubbles on social media platforms.
- **Historical Bias:** Training data often reflects historical biases and inequalities. For example, if patients of rural or other geographic areas historically used less healthcare because of a lack of services in the area, AI models trained on such data may predict a lower need for these patients than urban patients based on historical data.
- **Data Imbalance:** In scenarios where one group is underrepresented in the training data, AI models may struggle to provide fair and accurate predictions for that group. This can occur in medical diagnosis, where certain rare conditions may be overlooked.

Addressing Bias in AI and ML

To address bias in AI, it is crucial to start with diverse and representative training data that accurately reflects the real-world populations and scenarios the AI system will encounter. This will be especially important in the mental health space, as biased AI systems may lead to misdiagnosis, lack of treatment, and poor care experiences. Implementing algorithmic fairness techniques, such as re-sampling, re-weighting, and fairness-aware algorithms, can also help mitigate bias and ensure that AI models make fair predictions across different demographic groups. Additionally, continual monitoring, auditing, and transparency throughout the AI development lifecycle, combined with diverse teams and ethical guidelines, are essential to

identify, rectify, prevent, and mitigate the risk of bias in AI and promote fairness and equity in its applications.

References

[i] Zohuri, B., & Behgounia, F. (2023). Chapter 8—Application of artificial intelligence driving nano-based drug delivery system. In A. Philip, A. Shahiwala, M. Rashid, & Md. Faiyazuddin (Eds.), A Handbook of Artificial Intelligence in Drug Delivery (pp. 145–212). *Academic Press*. https://doi.org/10.1016/B978-0-323-89925-3.00007-1
[ii] American Medical Association. Augmented intelligence in medicine. (2024). American Medical Association. https://www.ama-assn.org/practice-management/digital/augmented-intelligence-medicine
[iii] Ahuja, R. (n.d.). *Introduction to Artificial Intelligence (IBM)*. Coursera. https://www.coursera.org/learn/introduction-to-ai
[iv] Ahuja, R. (n.d.). *Introduction to Artificial Intelligence (IBM)*. Coursera. https://www.coursera.org/learn/introduction-to-ai
[v] Ahuja, R. (n.d.). *Introduction to Artificial Intelligence (IBM)*. Coursera. https://www.coursera.org/learn/introduction-to-ai
[vi] Ahuja, R. (n.d.). *Introduction to Artificial Intelligence (IBM)*. Coursera. https://www.coursera.org/learn/introduction-to-ai
[vii] Garriga, R., Mas, J., Abraha, S., Nolan, J., Harrison, O., Tadros, G., & Matic, A. (2022). Machine learning model to predict mental health crises from electronic health records. *Nature Medicine*, *28*(6), 1240–1248. https://doi.org/10.1038/s41591-022-01811-5
[viii] Ahuja, R. (n.d.). *Introduction to Artificial Intelligence (IBM)*. Coursera. https://www.coursera.org/learn/introduction-to-ai
[ix] *Machine learning, explained | MIT Sloan*. (2024, August 28). https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained
[x] Ahuja, R. (n.d.). *Introduction to Artificial Intelligence (IBM)*. Coursera. https://www.coursera.org/learn/introduction-to-ai
[xi] Thakkar, A., Gupta, A., & De Sousa, A. (2024). Artificial intelligence in positive mental health: A narrative review. *Frontiers in Digital Health*, *6*, 1280235. https://doi.org/10.3389/fdgth.2024.1280235
[xii] Andrews, E.L. (2021). How flawed data aggravates inequality in credit. *Stanford University*. https://hai.stanford.edu/news/how-flawed-data-aggravates-inequality-credit
[xiii] Hailu, R. (2019, July 24). Fitbits and other wearables may not accurately track heart rates in people of color. *STAT*. https://www.statnews.com/2019/07/24/fitbit-accuracy-dark-skin/
[xiv] Sjoding, M. W., Dickson, R. P., Iwashyna, T. J., Gay, S. E., & Valley, T. S. (2020). Racial Bias in Pulse Oximetry Measurement. *New England Journal of Medicine*, *383*(25), 2477–2478. https://doi.org/10.1056/NEJMc2029240