THE RISKS AND POTENTIAL OF LARGE LANGUAGE MODELS IN MENTAL HEALTH CARE:

A Critical Analysis through the Lens of Data Feminism

By Carolyn Wang

Introduction

Over the past four decades, artificial intelligence (AI) has transformed from a relatively niche field of computer science to a ubiquitous technology dominating academic research publications as well as our daily lives; AI applications continue to be implemented across myriad sectors. In particular, there has been significant interest in the use of large language models (LLMs), which are AI models that can interact with human language, in the context of mental healthcare. Famous LLMs include OpenAI's GPT series (which powers ChatGPT), Meta's Llama, and Google's Gemini, among numerous others. Researchers and mental health professionals alike are excited by the increased accessibility that AI could bring if applied to mental healthcare (Feng, Hu, and Guo 2022). Previous work has proposed the use of LLMs to aid clinicians in diagnosing and monitoring patients; to train new clinicians; as well as to provide support to patients through direct interaction (Muetunda et al. 2024; Sweeney et al. 2021; Koloury et al. 2022; Olawade et al. 2024). Recent work shows that people are talking to general purpose chatbots for mental health support (Zao-Sanders 2025; Jung et al. 2025; Rousmaniere et al. 2025), with one survey finding that nearly 50% of respondents, consisting of a sample of adult residents of the United States who had used at least one LLM before and self-disclosed being diagnosed with a mental health condition, had turned to an LLM for psychological support within the last year (Rousmaniere et al. 2025). Because LLM-powered chatbots, such as ChatGPT, are not subject to the same regulations as other technologies geared specifically towards mental healthcare, the safety of their use in mental health is unverified. The apparent prevalence of the use of chatbots in this way is thus especially concerning.

Mental health is a historically biased field (see Section 4) in which marginalized communities continue to suffer from lower access to and quality of care (Shim and Vinson 2020). Given the plethora of biased behaviours LLMs have demonstrated (for example: Busker, Choenni, and Bargh 2023; Kotek, Dockum, and Sun 2023; Salinas, Haim, and Nyarko 2025), it is important to examine the safety and ethical implications of its application in a field already wrought with injustice. One lens through which we can begin this important analysis is that of data feminism, described in the next section. Exploring this use of LLMs critically is the first step to implementing these technologies responsibly and in a way that challenges oppressive norms, rather than reinforcing them. This essay will first introduce the framework of data feminism, then critically analyze the process of building and using LLMs for mental healthcare applications through that lens, and close with musings on alternative approaches to augmenting the mental health system in alignment with data feminist principles.

Data Feminism

Lauren Klein and Catherine D'Ignazio are self-described "data scientists and data feminists" (Klein and D'Ignazio 2020, 8) who published an influential book called *Data Feminism*. *Data Feminism* argues that in our increasingly digital world, data is power. Those with access to vast amounts of peoples' data, such as big tech companies recording our purchasing habits and social media use, are commanding this power unjustly to reinforce societal power structures and perpetuate systemic oppression. *Data Feminism* presents a set of seven principles (later expanded with two additional principles that address AI-specific concerns), thought of collectively as data feminism, which

seek to challenge the current hegemonic norms of data work. Data feminism is "a way of thinking about data, both their uses and their limits, that is informed by direct experience, by a commitment to action, and by intersectional feminist thought" (Klein and D'Ignazio 2020, 8). I will describe the principles briefly but defer to the original book (Klein and D'Ignazio, 2020) and supplementary article (Klein and D'Ignazio, 2024) for in depth explanations and examples of each of them.

The principles are as follows:

1. **Examine Power**

> *Examining power means naming and explaining the forces of oppression that are so baked into our daily lives— and into our datasets, our databases, and our algorithms— that we often don't even see them.*
>
> *- (Klein and D'Ignazio 2020, 24)*

The concept of the *matrix of domination* was introduced by sociologist Patricia Hill Collins and states that power operates through intersecting domains to oppress marginalized communities: structurally, via legal frameworks and public policy; disciplinarily, through institutional regulation and bureaucratic oversight; hegemonically, by shaping consciousness through cultural institutions such as the media and education; and interpersonally, through individual lived experiences of oppression. This matrix, Klein and D'Ignazio propose, is a useful framework through which we can begin to examine power and dissect the various ways that power interacts such that data often becomes a tool for oppression.

2. **Challenge Power**

> *Challenging power requires mobilizing data science to push back against existing and unequal power structures and to work toward more just and equitable futures.*
>
> *- (Klein and D'Ignazio 2020, 53)*

Equipped with an understanding of power from the first principle, the second principle asks us to use data to challenge power. This can involve using data to enhance our understanding of issues relating to societal power imbalances. Of greater relevance to this essay, this may also involve re-examining the ways that we build and evaluate AI models. For example, who decides whether an AI model is 'good'? Without a thorough examination of this simple question, these systems pose a serious threat of harming the communities and voices that are overlooked.

3. **Rethink Binaries and Hierarchies**

> *Data feminism requires us to challenge the gender binary, along with other systems of counting and classification that perpetuate oppression.*
>
> *- (Klein and D'Ignazio 2020, 8)*

Historic tendencies to classify people into the gender binary, racial categories, and in many other reductive ways are being empowered by AI systems. For example, researchers at Stanford claimed to create 'gaydar' able to predict whether someone is gay based on a photo of their face more accurately than a human can (Wang and Kosinski 2018). It's easy to imagine how this technology could be dangerous in the hands of oppressive regimes in which homosexuality is still criminalized. However, even in other applications of such an AI which are not explicitly dangerous, its design is still reductive and ignores the many other sexualities that exist. Unfortunately, data and technology are often portrayed as 'objective' and the assertion that there are only two sexualities (for example) thus gains validity as a result of the ignorant design of the algorithm. Data feminism asks us to resist perpetuating these rigid classifications by questioning the goals of the algorithms we create and use.

4. **Elevate Emotion and Embodiment**

> *Rebalancing emotion and reason opens up the data communication toolbox and allows us to focus on what truly matters in a design process: honoring context, architecting attention, and taking action to defy stereotypes and reimagine the world.*
>
> *- (Klein and D'Ignazio 2020, 96)*

To any AI algorithm, data of all types end up being converted to numbers before they can be processed. However, not all knowledge can be represented in numbers and the insistence on scientific proof can itself become a form of epistemic oppression, particularly when it dismisses or devalues the knowledge derived from the lived experiences of marginalized people. This is especially true

because these communities often don't have the resources to produce the quantitative evidence demanded of them This results in a vicious cycle wherein the lack of 'evidence' prevents issues from being addressed, thereby enabling the continuation of harm. By elevating forms of knowing outside of traditional empirical frameworks, we create space for voices which would otherwise be overlooked.

5. Embrace Pluralism

> *Embracing pluralism in data science means valuing many perspectives and voices and doing so at all stages of the process— from collection to cleaning to analysis to communication. It also means attending to the ways in which data science methods can inadvertently work to suppress those voices in the service of clarity, cleanliness, and control.*
>
> *- (Klein and D'Ignazio 2020, 130)*

Building on the previous principle, this principle advocates for data workers to embrace the plurality of ways of knowing enabled by elevating emotion and embodiment. Combining wisdom from multiple perspectives, it argues, results in deeper knowledge.

6. Consider Context

> *Rather than seeing knowledge artifacts, like datasets, as raw input that can be simply fed into a statistical analysis or data visualization, a feminist approach insists on connecting data back to the context in which they were produced. This context allows us, as data scientists, to better understand any functional limitations of the data and any associated ethical obligations, as well as how the power and privilege that contributed to their making may be obscuring the truth.*
>
> *- (Klein and D'Ignazio 2020, 152-153)*

This principle highlights the fact that data is not objective or value neutral; it is created within a societal context which informed not only how the data was to be collected, but what data was collected in the first place. Without this context, it is easy to overlook the implicit values and norms embedded in data. Consequently, these norms may be propagated en masse by AI systems built on these imperfect datasets, abstracting the sources of this harm into seemingly 'neutral' AI outputs.

7. Make Labor Visible

> *Behind the magic and marketing of data products, there is always hidden labor— often performed by women and people of color, which is both a cause and effect of the fact that this labor is both underwaged and undervalued. Data feminism seeks to make this labor visible so that it can be acknowledged and appropriately valued, and so that its truer cost— for people and for the planet— can be recognized.*
>
> *- (Klein and D'Ignazio 2020, 185)*

The effort involved in building AI extends far beyond software engineers with high salaries at prestigious companies in Silicon Valley. It includes the labour of workers maintaining data centers, miners (many of whom are in the global south and suffer from dangerous working conditions) who source the rare-earth minerals necessary for the hardware AI runs on, data annotators who prepare the data to be inputted into the AI system, and many more. However, most of this labour is undervalued, underpaid, and workers often suffer from precarious working conditions (atlas of AI), echoing the labour hierarchies of colonialism wherein certain jobs are glorified while others are devalued and rendered invisible. Highlighting this invisible labour is a critical step in rectifying these capitalist and colonial dynamics.

8. Environmental Impact

> *[AI] systems seem positioned to benefit elite users in the Global North, even as they exact their cost on those in the Global South. This is an environmental issue, but it is also a feminist issue, as these effects are not only experienced unequally in terms of geography, but also in terms of gender.*
>
> *- (Klein and D'Ignazio 2020, 14)*

Though this principle is self-explanatory, its importance in the context of feminism is difficult to overstate. The global south, especially women and people of colour, experience the negative consequences of AI development, through the ecological toll it extracts on the environment, most intimately. Therefore, it is inherently feminist to critically examine and push back against the negative environmental impacts of AI.

9. Consent

> *As we await the development of informed guidelines for fair use, we can be certain that something other than the current system–in which Big Tech steals people's work, exploits it, makes money, and facilitates structural violence along the way-is required."*
>
> *- (Klein and D'Ignazio 2020, 14)*

Technology-facilitated gender-based violence is a well-known issue. Women and gender minorities are subjected to disproportionate harms enabled through technology, such as cyberstalking, cyber-bullying, and online harassment to name a few. AI has continued to empower those perpetrating these harms. Specifically, non-consensual deep fake porn is created with the likeness of real people using image generation models (Moreau and Rourke 2024). The issue of consent persists in the realm of AI through this and other non-consensual interactions with AI systems, including through the theft of data.

By examining the potential of AI in mental health care through the lens of data feminism, I hope to contribute to the scholarly conversation around the ethics and safety of this use case. The purpose of this essay is not to advocate for or against implementing AI in the mental health space, but to examine this proposition critically from the perspective of those most at risk of experiencing harm as a result of it.

Data Feminism

There are several practices that are foundational to building industrial LLMs such as the GPT models from OpenAI, Grok from xAI, Claude from Anthropic, and any other model with comparable performance. For example, engineers require data to build any AI model which can be challenging to obtain. This section explores the questionable ethics of data practices and other processes shared by the development of all industrial-scale LLMs including ones being proposed for use in mental health.

LLMs are trained on vast amounts of text data scraped from the internet. For example, OpenAI's GPT-3 was trained on a set of text data with nearly 375 billion words (Brown et al. 2020) - to give you a sense of scale, this is approximately equivalent to 375,000 times the length of the *entire* Harry Potter series (OpenAI has stopped releasing details on the data used to train its models since GPT-3, but experts believe that the datasets have only grown). Often, the legality of the use of this data is sketchy; the New York Times famously launched a lawsuit against OpenAI and Microsoft, the producers of two of the most famous LLMs, for their unauthorized use of New York Times articles to train their LLMs. The lawsuit is ongoing at the time of writing this essay, however it highlights an important issue surrounding the creation of AI technologies: where are technology creators sourcing the huge quantities of data needed to build their models, and is this process ethical? Many artists and authors have voiced concerns over the use of their work to train generative AI models (Lamb, Brown, and Grossman 2024; Jiang et al. 2023). In fact, in addition to the New York Times, OpenAI and Microsoft have been sued by the Authors Guild as well as other well-known authors, including George R.R. Martin and Jodi Picoult, for infringing on copyright laws by using the authors' works to train their language models. As a consequence of the opacity surrounding AI's training data, the labour of creating it may be considered 'ghost work,' a term coined by anthropologist Mary Gray and computer scientist Siddarth Suri to describe the hidden labour powering many technological systems which is obfuscated from end users through non-transparent labour practices. Another form of ghost work is data cleaning and annotation, which prepares raw data to be used for model training. Importantly, this work is often outsourced to workers in the global south who face precarious working conditions, are underpaid, and are often women of colour (Klein and D'Ignazio 2020; Crawford 2021; Gray and Suri 2019). The breadth of ghost work required to develop and maintain 'innovative technologies' is out of the scope of this essay, however *Ghost Work* by Mary Gray and Siddarth Suri is a good starting point to learn about the often overlooked and unethical labour powering the technological conveniences that have come to be ubiquitous in our daily lives.

Beyond these data practices, the training process of LLMs are riddled with other concerns as well. Environmental activists have called out the 22 million liters of water used to train Meta's Llama 3 model. In one case, this tremendous water usage left residents nearby a data center,

which had been newly built by Meta, without access to water in their home (Tan 2025). Additionally, machines used to train LLMs (and other AI models) are built using resources extracted at a similarly devastating cost to the environment as well as to the communities surrounding extraction sites and the labourers working in them (Crawford 2021). These topics similarly deserve to be considered deeply but are not specific to the topic of this essay (that is, these concerns are applicable to any LLM and many also extend to other types of AI models) and comprehensive discussion is out of the scope. Researcher Kate Crawford's *Atlas of AI* is a good place to begin if you'd like to learn more.

Once an LLM has been trained, it is often fine-tuned to optimize its performance in a particular domain or task. This requires yet more data. In a field such as mental health, this is especially problematic - the data for fine-tuning must be domain specific. Though casebooks exist for training purposes, they are limited in scope and quantity; therefore, the data in question would likely have to include actual patient data to be sufficient in quantity and scope. Data like this is ethically and legally complex to collect and use as it would include sensitive, deeply personal information about real individuals' mental health experiences. The already significant concerns around data privacy and consent are heightened in this context. Whose data is being used? Was meaningful consent obtained? Can such data ever truly be anonymized in a way that protects those individuals? In addition, is this the data we want to collect? As we will see in greater depth in the next section, this data is often riddled with biases. For example, researchers found in an analysis of medical records from the New York City jail system over 2011 to 2013 that Black and Hispanic inmates in jails in New York City were less likely to receive mental health services compared to their White counterparts, but more likely to be subject to solitary confinement (Kaba et al. 2015). Reflecting on their results, Kaba et al. express concern that "some groups in the jail system are more likely to elicit treatment responses whereas others are more likely to meet with a punishment response" (Kaba et al. 2015). Systemic disparities such as this risk being perpetuated if we do not examine what data we are collecting and using to train these models.

Without even discussing the development of LLM systems, we've already run up against fundamental practices which run counter to several of the principles proposed in data feminism. Specifically, power is not being examined or challenged if the practice of data collection continues to exploit the most vulnerable. Ghost work is labour not made visible, and the question remains on the consent of those whose data is being collected. In particular, poverty and oppression are significant social determinants to poor outcomes in health and specifically mental health. Putting people, who are most often already members of marginalized communities, in precarious, underpaid working conditions runs counter to the goal of implementing LLMs mental healthcare in the first place. If using LLMs to improve mental healthcare services rests on the exploitation of people without the power to resist it, is it worth pursuing? Is it still in service of the fundamental goal of improved mental health if it results in the deterioration of the mental health of those workers? The answer may very well be yes depending on the philosophical lens through which these questions are answered (for example, a utilitarian perspective might argue that the benefits to many is worth the cost of a few), however calling out this hypocrisy is the first step to finding an alternative solution that doesn't rely on ghost work.

LLM APPLICATIONS

Bias in Mental Health

The field of mental health care is built on decades of psychological research which scholars argue has contained systemic inequality from at each step of the process: Roberts et al. (2020) queried over 26000 empirical articles in top-tier psychological journals over the years of 1974 to 2018 and found significant racial imbalances in those contributing and editing the journals, as well as in the subjects being included in studies. Unsurprisingly, white people were disproportionately overrepresented in each of these categories. Dr. Lonnie Snowden (2003) argued that these racial biases also manifest in practise through the diagnosis and treatment of patients from different racial backgrounds. Similarly, gendered stereotypes around mental health have long existed and influenced mental health research. In fact, the first description of hysteria dates to 1900 BC. Hysteria was a term

generally used to describe mental unwellness in people with uteri and which carried heavy stigma due to its connotation that the suffering was due to feminine weakness or vulnerability. This view of hysteria as a 'female disease' and consequent effect on the perception of "mental disorder, especially in women, [being] so often misunderstood and misinterpreted, [and generating] scientific and / or moral bias, defined as a pseudo-scientific prejudice" persisted for 4000 years, until the 19th century (Tasca et al. 2013). In modern times, similar prejudiced views on mental health persist. For example, researchers Bacigalupe and Martín found in a 2020 study that women's mental health is being 'medicalised,' meaning that women with depression or anxiety are prescribed psychotropic medication at a disproportionate rate compared to men (the study did not consider nonbinary/gender diverse people). Clearly, there is a long history of inequity in mental health research and care.

These biases have permeated North American social consciousness as well. For example, in clinical settings the stereotype that black people feel less pain, a belief originally used to justify slavery, results in the underdiagnosis and treatment of patients. However, this stereotype persists outside of the clinical context as well (Trawalter and Hoffman, 2015). Biases do not exist in silos. As a result, LLMs have been found to replicate these biases in conversational contexts as well (for example: Busker, Choenni, and Bargh 2023; Kotek, Dockum, and Sun 2023; Salinas, Haim, and Nyarko 2025). Given the long legacy of bias in mental health care and early but substantial evidence of bias in LLMs, the risk that for these inequities to be further exacerbated by an unexamined implementation of LLMs in mental health care is clearly present.

Clinician Assistance

Several uses of LLMs in mental health have been proposed to interact with and assist clinicians. Many researchers have examined the abilities of LLMs to diagnose patients with mental health conditions and severity labels (for example: Yang et al. 2024; D'Souza et al. 2023), create treatment plans (Elyoseph, Levkovich, and Shinan-Altman 2024; Berrezueta-Guzman et al. 2024), and manage patient profiles.

Because bias has been the historical norm, unbiased data is scarce. Therefore, LLMs rely on biased data for training and fine-tuning. The process of LLM training and fine-tuning can be understood intuitively as finding the optimal parameters for a model to predict the sequences of words that were observed in the training data - it is trained to replicate what it is fed. By design, if the training data contains sexism, racism, or any other type of bias, an LLM will learn to replicate it. Given the historical legacy of injustice in mental health care, the substantial body of literature and text containing these biases, both in research and practise, it is unsurprising that researchers have already begun observing differences in LLMs' responses to queries on mental healthcare for people of different genders and sexualities (Soun and Nair 2023). One study presented an LLM with case vignettes of patients with anorexia/ bulimia nervosa and evaluated the consistency (or lack thereof) of its assessment of the patients through psychometric tests - the study found that the LLM's output was biased based on the gender of the patient described (Schnepper et al. 2025). Similarly, patient monitoring risks being less accurate for marginalized communities and clinician training applications (eg. practising patient interactions by conversing with a chatbot) risks lacking diversity in the patient profiles the LLMs present.

Given the current interest in deploying LLMs to diagnose and potentially triage patients, a process in which accuracy and fairness are essential to ensuring that patients receive the care that they need, these biases pose severe risks to the patients whose wellbeing is at stake. Patient monitoring could fail at a disproportionately high rate for some patients compared to others, and clinician training systems may not be representative of the full range of patients that clinicians should be prepared to treat. In the context of mental healthcare, a field in which marginalized communities have historically faced and continue to face higher barriers to accessing mental health services and lower quality care, the prospect of reinforcing the status quo is alarming to say the least. Without examining the data being used to train systems that could have life-changing impacts on patients in need of mental health care, we run the risk of embedding biases into opaque algorithms that perpetuate harmful norms. At a broader level, failing to rigorously test these systems for embedded bias and to develop strategies for mitigation risks exacerbating

existing mental health disparities rooted in systemic social determinants and perpetuating cycles of social injustice.

Direct Patient Interaction

Sewell Setzer III was just 14 when he tragically committed suicide moments after a chatbot which he had been having intimate conversations with told him to "come home to me as soon as possible" (Roose 2024). This was not the first time an AI chatbot had been accused of contributing to the death of its users, highlighting the profound impact interactions with LLMs can have. Clearly, it is unacceptable for a chatbot to be outputting this dangerous rhetoric and robust safeguards must be implemented and continuously monitored. Knowingly or not, many LLM-powered chatbots are being used in mental health contexts (Roose 2024; Rousmaniere et al. 2025) and interacting directly with users to address their mental health needs. Unfortunately, LLMs risk causing more subtle damage as well since they have been found to demonstrate racism, sexism, and western-centric values among other types of biases. Some of these biases have been quantified by researchers (Straw and Callison-Burch 2020). Recall the hegemonic and interpersonal arms of the matrix of domination; by replicating these biases en masse with marginalized individuals, especially those seeking psychological care, LLMs act as a tool through which oppression continues to proliferate.

Beyond bias, the dangers of bias, some researchers have also expressed concern about the ability of an algorithmic technology to fulfill the needs that traditional mental health treatment methods address. The therapeutic alliance describes the relationship between a patient and their therapist; evidence has consistently shown that "the quality of the therapeutic alliance is linked to the success of psychotherapeutic treatment across a broad spectrum of types of patients, treatment modalities used, presenting problems, contexts, and measurements" (Stubbe 2018). The therapeutic alliance emphasizes "the affective bond between patient and therapist" (Stubbe 2018); the ability of an LLM to form a bond with a human is dubious at best, at least as these technologies currently stand. Additionally, researchers emphasize the importance of non-verbal cues in general mental health treatment (Guzman-Santiago et al. 2024). Since LLMs can only process textual data as input, they are unable to account for these non-verbal cues. The illusion of having a capable algorithmic system for mental health support overlooks, perhaps even suppresses, our inherently embodied experiences as humans as technology is unable to engage on that level.

Especially given the unregulated use of chatbots for therapeutic purposes, which appears unsettlingly common as found in the aforementioned study by Rousmaniere et al. (2025), the concerns around bias and the constraint of text-only interactions with any type of 'therapeutic' LLM technology could be a real threat to patient wellbeing.

Data Feminist Analysis

So which data feminist principles are being violated? We have established that the unexamined use of LLMs in mental healthcare does not examine or challenge power, rather reinforcing its current state. Part of this current state includes the assertion of binaries and hierarchies, violating the third principle, and there remains no space to implement principles 4 or 5 either as alternative, embodied ways of knowing continue to be overlooked. The deeply flawed context of the data LLMs are currently trained on is decades of racism, sexism, and capitalism but it is overlooked if the data is not interrogated and amended adequately before its use. Clearly, much work should be done before LLMs can be safely implemented in these settings. Table 1 below summarizes the reflections on the use of LLMs in mental health care discussed throughout this essay.

Data Feminist Principle	Reflections from the use of LLMs in mental health care
Examine Power	LLMs reflect past data, thereby perpetuating past biases which have been pervasive in the field of psychology and mental health care.
Challenge Power	Because LLMs are trained to reflect the data they are trained on, they propagate the status quo by design as opposed to challenging power in any way.
Rethink Binaries & Hierarchies	Because LLMs 'think' within the box constructed by the data they are given, there is no way to challenge or change any dominant binaries or hierarchies that exist in this data. As we explored through the example of prison mental health records, there appears to be a hierarchy in mental health treatment (in this context and in the context of broader mental health inequities) which must be challenged and dismantled consciously.
Elevate Emotion & Embodiment	LLMs are unable to form bonds with the people turning to them for emotional support, and cannot even engage with the patient in any way beyond the exchange of text. Therefore the experience of engaging with an LLM is inherently disembodied and runs counter to this principle.
Embrace Pluralism	As we saw in Section 4.1, there is a long legacy of injustice and exclusion in the field of psychology and mental health. To embrace pluralism, we must think beyond the selective perspectives represented in existing literature, which is not possible within the current paradigm of LLM development which largely acquires the requisite training data from existing records.
Consider Context	The context in which the vast majority of mental health literature and records was created is one of pervasive bias, such as racial and gender-based exclusion. With this context in mind, it is scarcely possible to imagine that an LLM could be created which is mindful of this context and makes its users aware of it as well. The context behind an interaction with an LLM is rich, and perhaps too rich for the average user to grasp meaningfully.
Make Labor Visible	As discussed in Section 3, much of the labor powering LLMs, and by extension any LLM-enabled mental health technology, is invisible and oppressive. Though not specific to mental health use cases of LLMs, this principle is broadly violated by the current norms of the AI industry.
Environmental Impact	LLMs are incredibly extractive and resource intensive on the environment and the effects of the environmental damage are not distributed equally. Those already most oppressed and vulnerable, such as citizens in the global south, bear a disproportionate amount of this impact (Ren and Wierman 2024). This injustice conflicts with fundamental principles of equity in addition to ignoring the effects of the environmental impact of LLMs.
Consent	Data theft is an ongoing and very prevalent issue in the development and training of LLMs. Often, companies' data use is non-consensual and the lack of consent practices does not reflect trauma-informed computing practices (Chen et al. 2022). This may be harmful or triggering to users, particularly those whose mental health concerns involve traumatic experiences.

Table 1 – Summary of the reflections on the use of LLMs in mental health care discussed throughout this essay.

Techno-Optimist Possibilities

Techno-optimism refers to "the view that technology, when combined with human passion and ingenuity, is the key to unlocking a better world" (Danaher 2022). Scholars have critiqued this perspective for a variety of reasons (Danaher 2022) including that it is too broad to be widely applicable or that it needs specific caveats to ensure the safety and equity of any implementation of techno-optimist ideas. A data feminist techno-optimism might look like techno-optimism which honours the nine principles set forth by Klein and D'Ignazio. It might encompass the view that yes, technology is incredibly powerful and may well unlock a better world, but only if this better world is defined by a pluralistic view that transcends the normative hegemonic silos in which most technologies are currently built.

May of the risks of a blindly techno-optimist approach to LLM-powered mental health care solutions were discussed in this essay. Despite these risks however, LLMs have the potential to bring large-scale and highly impactful improvements to mental healthcare. They could increase access to psychological care, particularly in underserved communities; assist clinicians, who are often under-resourced, by streamlining their operations; and provide patients with additional support between regularly scheduled visits with clinicians. This would allow mental healthcare providers to dedicate more time and attention to the areas of care where human expertise and empathy are most essential.

Considering the pervasive biases within the current mental healthcare system, an LLM specifically engineered to be equitable (assuming these efforts are successful) could actually improve upon the status quo. In this very optimistic case, deferring certain clinical decisions to the LLM could mitigate the risk of bias introduced by human clinicians. Existing mental health care systems are already unjust, as evidenced by large mental health disparities between different communities (Aneshensel 2009) – perhaps these injustices could be challenged with the help of technologies designed and proven to be equitable.

To be somewhat more realistic, LLMs could begin by being applied to the low-risk use-cases such as patient monitoring under the direct supervision of a clinician. In this case, the system would only be a supplement to the current standard of care and researchers could study its benefits without risking inadequate treatment. Bit by bit, as further research establishes the safety or danger of LLMs in the mental health context, greater agency could be granted to the LLMs.

Building equitable models from biased data is a challenging task, and developers can look to the principles of data feminism for ideas to achieve this goal. For example, the context can be consulted to inform changes that can be made to the datasets to reflect the ideal, rather than current, state of mental healthcare. Embodied knowledge from a plurality of communities can be consulted as well. The process of building these models should employ ethical data collection practices which emphasize consent and fair labour practices for data workers. The environmental cost of model development should be minimized as much as safely possible. Importantly, the communities most at risk of being affected by these technologies should be consulted in the development process; if technological solutions are undesirable to the community, they should not be developed.

Of course, this vision can only be realized if the models it relies on are able to provide care at a quality at least equal to that provided by human clinicians. The LLMs should be audited to ensure high accuracy across all demographic groups and for patients with intersectional identities. In case this is not feasible or the LLM displays bias, it should not be used for patients whose care could be compromised.

Conclusion

While the integration of LLMs into mental healthcare offers the potential of increased accessibility and support, particularly for underserved communities, their use cannot be divorced from the long history of bias and inequity embedded within both artificial intelligence and mental health systems. Applying a data feminist lens reveals not only the potential harms of deploying LLMs without accountability, but also the structural assumptions and power imbalances they risk reinforcing. Rather than rushing to embrace these tools as solutions, we must consider who benefits, who is harmed, and whose voices are being excluded from their development and deployment. If we are to leverage LLMs in ways that truly support mental health, particularly for those most marginalized by the existing system, we must look to alternative ways of building technology

grounded in equity and justice. A path forward requires balance between innovation and safety, which iterative deployment could help us strike. In the best case, this path could lead us to a future in which AI is an ally as we reach towards equitable mental health systems for all.

Works Cited

Muetunda, Faustino, Soumaya Sabry, M. Luqman Jamil, Sebastião Pais, Gaël Dias, and João Cordeiro. "AI-Assisted diagnosing, monitoring and treatment of mental disorders: A survey." *ACM Transactions on Computing for Healthcare* 5, no. 4 (2025): 1-24.

Olawade, David B., Ojima Z. Wada, Aderonke Odetayo, Aanuoluwapo Clement David-Olawade, Fiyinfoluwa Asaolu, and Judith Eberhardt. "Enhancing mental health with Artificial Intelligence: Current trends and future prospects." *Journal of medicine, surgery, and public health* (2024): 100099.

Sweeney, Colm, Courtney Potts, Edel Ennis, Raymond Bond, Maurice D. Mulvenna, Siobhan O'neill, Martin Malcolm et al. "Can chatbots help support a person's mental health? Perceptions and views from mental healthcare professionals and experts." *ACM Transactions on Computing for Healthcare* 2, no. 3 (2021): 1-15.

Koulouri, Theodora, Robert D. Macredie, and David Olakitan. "Chatbots to support young adults' mental health: an exploratory study of acceptability." *ACM Transactions on Interactive Intelligent Systems* (TiiS) 12, no. 2 (2022): 1-39.

Zao-Zanders, Marc. 2025. "How People Are Really Using Gen AI in 2025." *Harvard Business Review*, April 9, 2025. https://hbr.org/2025/04/how-people-are-really-using-gen-ai-in-2025

Verma, Pranshu, and Shelly Tan. 2024. "A bottle of water per email: the hidden environmental costs of using AI chatbots." *Washington Post*, September 18, 2024. https://www.washingtonpost.com/technology/2024/09/18/energy-ai-use-electricity-water-data-centers/

Tan, Eli. 2025. "Their Water Taps Ran Dry When Meta Built Next Door." *New York Times*, July 16, 2025. https://www.nytimes.com/2025/07/14/technology/meta-data-center-water.html

Jung, Kyuha, Gyuho Lee, Yuanhui Huang, and Yunan Chen. "I've talked to ChatGPT about my issues last night.": Examining Mental Health Conversations with Large Language Models through Reddit Analysis." *arXiv preprint arXiv:2504.20320*(2025).

Rousmaniere, Tony, Xu Li, Yimeng Zhang, and Siddharth Shah. *Large language models as mental health resources: Patterns of use in the United States*. 2025.

Shim, Ruth S., and Sarah Y. Vinson, eds. *Social (in) justice and mental health*. American Psychiatric Pub, 2020.

Busker, Tony, Sunil Choenni, and Mortaza Shoae Bargh. "Stereotypes in ChatGPT: an empirical study." In *Proceedings of the 16th international conference on theory and practice of electronic governance*, pp. 24-32. 2023.

Kotek, Hadas, Rikker Dockum, and David Sun. "Gender bias and stereotypes in large language models." In *Proceedings of the ACM collective intelligence conference*, pp. 12-24. 2023.

Salinas, Alejandro, Amit Haim, and Julian Nyarko. "What's in a name? Auditing large language models for race and gender bias." *arXiv preprint arXiv:2402.14875* (2024).

D'Ignazio, Catherine, and Lauren F. Klein. *Data feminism*. MIT press, 2023.

Klein, Lauren, and Catherine D'Ignazio. "Data feminism for AI." In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 100-112. 2024.

Wang, Yilun, and Michal Kosinski. "Deep neural networks are more accurate than humans at detecting sexual orientation from facial images." *Journal of personality and social psychology* 114, no. 2 (2018): 246.

Moreau, Shona and Chloe Rourke. 2024. "Fake porn causes real harm to women." *Policy Options*, February 8, 2024. https://policyoptions.irpp.org/magazines/february-2024/fake-porn-harm/

Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al. "Language models are few-shot learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.

Lamb, Carolyn, Daniel G. Brown, and Maura R. Grossman. "Precarity and solidarity: Preliminary results on a study of queer and disabled fiction writers' experiences with generative AI." Available at *SSRN 5045638* (2024).

Jiang, Harry H., Lauren Brown, Jessica Cheng, Mehtab Khan, Abhishek Gupta, Deja Workman, Alex Hanna, Johnathan Flowers, and Timnit Gebru. "AI Art and its Impact on Artists." In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 363-374. 2023.

Crawford, Kate. *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.

Gray, Mary L., and Siddharth Suri. *Ghost work: How to stop Silicon Valley from building a new global underclass*. Harper Business, 2019.

Kaba, Fatos, Angela Solimo, Jasmine Graves, Sarah Glowa-Kollisch, Allison Vise, Ross MacDonald, Anthony Waters et al. "Disparities in mental health referral and diagnosis in the New York City jail mental health service." *American journal of public health* 105, no. 9 (2015): 1911-1916.

Roberts, Steven O., Carmelle Bareket-Shavit, Forrest A. Dollins, Peter D. Goldie, and Elizabeth Mortenson. "Racial inequality in psychological research: Trends of the past and recommendations for the future." *Perspectives on psychological science* 15, no. 6 (2020): 1295-1309.

Snowden, Lonnie R. "Bias in mental health assessment and intervention: Theory and evidence." *American journal of public health* 93, no. 2 (2003): 239-243.

Tasca, Cecilia, Mariangela Rapetti, Mauro Giovanni Carta, and Bianca Fadda. "Women and hysteria in the history of mental health." *Clinical practice and epidemiology in mental health: CP & EMH* 8 (2012): 110.

Bacigalupe, Amaia, and Unai Martin. "Gender inequalities in depression/anxiety and the consumption of psychotropic drugs: Are we medicalising women's mental health?." *Scandinavian journal of public health* 49, no. 3 (2021): 317-324.

Trawalter, Sophie, and Kelly M. Hoffman. "Got pain? Racial bias in perceptions of pain." *Social and Personality Psychology Compass* 9, no. 3 (2015): 146-157.

Schnepper, Rebekka, Noa Roemmel, Rainer Schaefert, Lena Lambrecht-Walzinger, and Gunther Meinlschmidt. "Exploring Biases of Large Language Models in the Field of Mental Health: Comparative Questionnaire Study of the Effect of Gender and Sexual Orientation in Anorexia Nervosa and Bulimia Nervosa Case Vignettes." *JMIR Mental Health* 12, no. 1 (2025): e57986.

Soun, Ritesh S., and Aadya Nair. "ChatGPT for Mental Health Applications: A study on biases." In *Proceedings of the Third International Conference on AI-ML Systems*, pp. 1-5. 2023.

Roose, Kevin. 2024. "Can A.I. Be Blamed for a Teen's Suicide?" *The New York Times*, October 24 2024. https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html

Ren, Shaolei, and Adam Wierman. 2024. "The Uneven Distribution of AI's Environmental Impacts." *Harvard Business Review*, July 15, 2024. https://hbr.org/2024/07/the-uneven-distribution-of-aisenvironmental-impacts

Straw, Isabel, and Chris Callison-Burch. "Artificial Intelligence in mental health and the biases of language based models." *PloS one* 15, no. 12 (2020): e0240376.

Feng, Xishuang, Maorong Hu, and Wanhui Guo. "Application of artificial intelligence in mental health and mental illnesses." *Proceedings of the 3rd International Symposium on Artificial Intelligence for Medicine Sciences*. 2022.

Yang, Kailai, et al. "MentaLLaMA: interpretable mental health analysis on social media with large language models." *Proceedings of the ACM Web Conference* 2024. 2024.

D'Souza, Russell Franco, et al. "Appraising the performance of ChatGPT in psychiatry using 100 clinical case vignettes." *Asian Journal of Psychiatry* 89 (2023): 103770.

Elyoseph, Zohar, Inbar Levkovich, and Shiri Shinan-Altman. "Assessing prognosis in depression: comparing perspectives of AI models, mental health professionals and the general public." *Family Medicine and Community Health* 12.Suppl 1 (2024): e002583.

Berrezueta-Guzman, Santiago, et al. "Future of ADHD care: evaluating the efficacy of ChatGPT in therapy enhancement." *Healthcare*. Vol. 12. No. 6. MDPI, 2024.

Stubbe, Dorothy E. "The therapeutic alliance: The fundamental element of psychotherapy." *Focus* 16.4 (2018): 402-403.

Chen, Janet X., et al. "Trauma-informed computing: Towards safer technology experiences for all." *Proceedings of the 2022 CHI conference on human factors in computing systems*. 2022.

Danaher, John. "Techno-optimism: An analysis, an evaluation and a modest defence." *Philosophy & Technology* 35.2 (2022): 54.

Aneshensel, Carol S. "Toward explaining mental health disparities." *Journal of Health and Social Behavior* 50.4 (2009): 377-394.