Cyborg Knowledge Production with an AI Psychologist:

Tangled Threads of Gendered Harm, Ethics, and Care Amidst a Mental Health Crisis

A thesis presented in partial fulfilment of the requirements for the degree of

Master of Arts

in

Psychology

at Massey University, Manawatu, Aotearoa New Zealand

Ollie Wylde

August 2025

Abstract

This thesis explores the use of artificial intelligence (AI) chatbots to provide mental health advice and the potential perpetuation of harmful gendered discourses through the technologisation of care. Situated within the ongoing mental health crisis in Aotearoa New Zealand and the exponential rise of generative AI, this study deals with the unprecedented complexities of operating within an emerging and rapidly evolving research field. Maintaining ethical relational dilemmas with limited institutional guidance and reinforcement of human exceptionalism challenged reflexive partnering with AI chatbots to co-produce knowledge. Donna Haraway's cyborg metaphor guided the methodological and epistemological considerations for the study, contributing to the introduction of critical concepts *cyborgphancy* (the sycophantic nature of AI chatbots) and *cyborg knowledge production* to facilitate understanding of this rapidly evolving research area. Semi-structured interviews were conducted with the AI Psychologist chatbot from the Character.ai platform. ChatGPT was utilised as a research assistant and an emic advisor. A threaded narrative analysis embraced the contradictory nature of cyborg knowledge production, weaving together partial and multiple relationships between researcher, AI chatbot, and help-seekers within the reproduction of psychological, gendered and biological essentialist discourses. Findings challenge the illusion of neutrality, interrogating the AI Psychologist's gender-neutral responses as reproduction of androcentric knowledge bases, reinforcing gendered power dynamics and systems of oppression. ChatGPT's emic analysis confirmed the perpetuation of harmful discourses, attributing this to fundamental design features of AI chatbots.

This study offers a qualitative feminist post-structural analysis to the emerging practice of engaging with AI chatbots for mental health support. There is substantial potential for harm to be perpetuated by AI within this context, due to the proliferation of AI chatbot usage and the failings of the mental health system to provide support. This risk necessitates greater scrutiny of AI chatbot use for mental health purposes, education of potential harms,

and robust safeguards to protect help-seekers.

Acknowledgements

Thank you to my supervisors Dr Mandy Morgan and Dr Geneva Connor. It has been an equally fascinating and terrifying journey as we navigated uncharted waters together, and I am truly grateful for the support of such steadfast companions. Mandy, it is an honour to have my work shaped by your expertise in feminist psychology - I will forever be asking, “where are the women?”. Dr G, your work in the cyborg and AI space has been invaluable in helping make sense of the new epoch we find ourselves being dragged into. I couldn’t have asked for better supervisors - my work as a psychologist and mind as a human are forever changed by the experience.

To my community who have also held me through this process - My beloved animal kin: Taffeta, Neve, Ratty, Astrid and Bruce Wayne, who tether me to the humanness of existence; Rachel, my diving buddy in deep waters; my chosen family, who encourage me always; my PhD people, your achievements are my beacon of hope; and to my dad, my biggest cheerleader since 11.58pm – Thank you.

Table of Contents

List of Illustrations and Tables

Cyborg Knowledge Production with an AI Psychologist: Tangled Threads of Gendered Harm, Ethics, and Care Amidst a Mental Health Crisis

Chapter One: Introduction

It is a difficult time to seek mental health support in Aotearoa New Zealand. Across the motu, people are suffering from mental distress at unprecedented rates, and the public healthcare system, constrained by a shortage of specialists and long waitlists, is failing to provide adequate support (Mulder et al., 2022; Healthpoint, 2023). Private mental health care has its own capacity issues, and prohibitive costs are often a barrier to accessing support. Systemic failures, driven primarily by funding constraints and staffing shortages, are contributing to a mental health crisis situated within the context of cost-of-living pressures, ongoing COVID-19 pandemic impacts, entrenched systemic inequalities, and deteriorating political and climate conditions (Ministry of Health, 2021). This crisis is not localised to Aotearoa New Zealand; many countries are facing similar challenges (World Health Organization, 2022).

The emergence of easily accessible AI chatbots has coincided with the deepening of this mental health crisis. Help-seekers are increasingly turning to AI as a supplementary or replacement mental health support service (Baumel et al., 2017; Fitzpatrick et al., 2017). This shift towards technology-based mental health support signals the beginning of a new era of post-human caring, where technology moves beyond the facilitation of healthcare responses through telehealth services and platforms to actively performing the healthcare itself (Lupton, 2014; Ruckenstein & Schüll, 2017). However, significant risks arise from the technologising of care, including potential harm to already vulnerable help-seekers and the creation or exacerbation of mental distress. As human-AI interactions continue to rapidly evolve, the implications of this emerging approach to mental health support have yet to be fully understood.

The Mental Health Crisis Context in Aotearoa, New Zealand

The demand for mental health support cannot be met by the current system, with 10.7% of adults unable to get professional support for their mental distress, a 54% increase compared to 2016/17 (Ministry of Health NZ, 2024). Increased pressure on the system is being driven by rising rates of high or very high psychological distress (HVHPD), which rose from 8.3% in 2018/19 to 13.0% in 2023/24 among adults (Ministry of Health NZ, 2024). There is an uneven distribution of increasing mental health distress across the motu, with increasing rates more highly concentrated in Māori, youth and transgender communities, indicating that the crisis is amplified for those already facing disparities due to systemic marginalisation (Ministry of Health NZ, 2024). The cost of mental health support is a contributing barrier to getting help, especially for populations who already experience low incomes due to marginalisation and health the impact of which is then exacerbated by the cost-of-living crisis and increased financial hardships (Bartley et al., 2024; Bell et al., 2022; Elers et al., 2021; Every-Palmer et al., 2024; Fa'alogo-Lilo & Cartwright, 2021).

The historical and ongoing impacts of colonisation contribute to the overrepresentation of Māori in mental health statistics (Maree Kopua et al., 2020; Taitimu et al., 2018). Displacement and theft of traditional lands, structural and institutional racism, rapid urbanisation of Māori and dishonouring of Te Tiriti O Waitangi have caused deep cultural wounds that have had a devastating impact on Māori mental health (Maree Kopua et al., 2020). Eurocentric determinations of mental health norms based on androcentric knowledge bases pathologise culturally specific behaviours and experiences which 'deviate' from these norms. Māori can hold multiple explanatory models for extrasensory experiences and psychosis, and situate schizophrenia diagnosis within the context of negative impacts from colonial and socio-political violence (Maree Kopua et al., 2020; Taitimu et al., 2018). Yet psychology ignores te ao Māori understandings of mental health, pathologising Māori through colonial knowledge systems (Black & Huygens, 2016; Lindsay et al., 2020; Maree Kopua et al., 2020). Under Te Tiriti o Waitangi, the Crown has an obligation to protect the

mental health of Māori, as cited in Waitangi Tribunal Claim Wai 2575 submitted by Māori psychologist Dr Michelle Levy in 2018 (Came et al., 2020). This claim sought to hold the Crown accountable for the way Māori mental health has been ignored and worsened by systemic racism. The current mental health crisis is exacerbating this issue, with 19.5% of Māori adults experiencing HVHPD in 2023/24, which is higher than the 13.0% national average (Ministry of Health NZ, 2024).

Transgender (trans) and non-binary people are historically marginalised groups that have also disproportionately been overrepresented in negative mental health statistics and outcomes due to the violence they experience (Kaufman et al., 2023; Wilson et al., 2024). Pandemics historically give rise to fascism, and this has been seen internationally since COVID-19 emerged (Maher et al., 2023; Modebadze, 2022). Anti-vaccine and anti-mandate groups formed in response to government health protections at the start of the pandemic shifted their focus to anti-trans rhetoric, after mandates were dropped (Hattotuwa et al., 2023). This led to a rise in transphobic ideologies, resulting in increasing violence towards trans and non-binary people since 2021 (Hattotuwa et al., 2023; Tan et al., 2022). A 2022 study revealed that trans and nonbinary people experience HVHPD at a rate six times higher (77%) than the national average in Aotearoa New Zealand (Yee et al., 2025).The growing threat of unsafety for these communities has contributed to worsening mental distress due to threats and experiences of increased violence, resulting in a greater need for mental health support services than was already required, from a system that already struggles to adequately support them (Yee et al., 2025). If these marginalised populations are turning to AI chatbots due to failings of the mental health system to meet their needs, an AI's perpetuation of gendered and colonial discourses could be particularly harmful for these already vulnerable populations.

A shortage of psychologists is a key contributing factor to this crisis, with an inadequate number of mental health professionals to provide care for the exponential rise in the number of people experiencing mental distress (Mulder et al., 2022; Rucklidge et al.,

2018). Waitlists are currently long, with recent studies indicating that young people are sometimes waiting over a year to access services, contributing to worsening mental distress (Wilson-Burke, 2024). Estimates show that Aotearoa's psychologist shortage is severe, with a shortage of 1000 psychologists, and projected further growing demand is expected to worsen the ongoing shortage (Cardwell, 2021). I became aware of the shortage of mental health professionals during my enrolment in the University's psychology programme. I was advised not to pursue clinical psychology as internship placements were so limited that it was unlikely I'd get in, regardless of my high GPA academic record. Over the next few years, talking with other students and through Facebook groups, I learned how difficult it is to get a placement in an internship or traineeship programme after completing a recognised Master's or Doctoral degree. Only 10-15% of applicants are offered entry to the training, which is required to become a practising psychologist in Aotearoa New Zealand (The New Zealand Psychologists Board, 2016).

The same concerns regarding the psychologist workforce development were raised over 15 years ago; a 2010 discussion paper by the New Zealand Psychological Society outlines issues my postgraduate psychology peers and I faced (The New Zealand Psychological Society, 2010). A lack of funding for these internships and an insufficient number of available practising psychologists to supervise interns were cited as reasons for the small number of internship places and the consequential low output of locally trained psychologists (Ministry of Health, 2022). The government has committed to growing the mental health workforce, increasing psychology internships from 12 in 2017 to 28 in 2022, and aiming to increase placements to 80 in 2027 (Little, 2022). There has been a 14% increase in locally trained registration applications to the New Zealand Psychologist's Board between 2019/20 and 2023/24 (New Zealand Psychologists Board, 2024). The development of an associate psychologist role is also underway (Doocey, 2024). However, the lower training requirements are viewed by some as inadequate for a mental health professional who will be working with distressed help-seekers (RNZ, 2025).

High workforce turnover is also impacting the numbers of available practising psychologists, with many leaving public sector roles due to burnout from working under high workloads within an inadequately resourced mental health service (Blayney & Kercher, 2023). This was worsened by the COVID-19 pandemic, with registered psychologists reporting a 70% increase in work-related stress and a 60% increase in caseloads (Kercher & Gossage, 2024; Kercher et al., 2024). There were 4026 registered psychologists with practising certificates in March 2024 (New Zealand Psychologists Board, 2024). Private practising psychologists are also struggling to keep up with demand, some with waitlists exceeding a year, and more than 50% reporting having to turn away at least 10 help-seeking families each month, and having to close their waitlists to excessive referrals and to mitigate the risks involved in leaving people with high mental distress in limbo (NZ College of Clinical Psychologists, 2021). Recruitment of overseas psychologists has been the historical solution to increase the workforce, with up to half of newly registered psychologists each year being overseas trained (NZ College of Clinical Psychologists, 2021). This practice is not a sustainable workforce shortage solution and risks psychologists having inadequate cultural competency to meet Te Tiriti obligations required under the Code of Ethics (New Zealand Psychological Society, 2002).

As a psychology student and someone with lived experience of mental health distress, I developed a sense of helplessness about the mental health system failures and the lack of timely and effective solutions to the insurmountable mental health crisis. My intention was to join the discipline of psychology to help people, but as my studies progressed, I quickly realised that the current model was inadequately resourced and failing so many who need help and those who provide the help. The beginning of my Master's thesis research in February 2023 coincided with the meteoric rise of the ChatGPT. Awareness of ChatGPT was increasing in the public from the start of 2023, reflecting the first time an advanced AI chatbot was accessible to the public (Gupta et al., 2023). Intrigued by this newly available technology, I began looking into how AI chatbots may be used as a

potential solution for the mental health crisis, filling the gap for easily accessible mental health support. Maybe the current rise of AI was what was needed as a safety net solution?

AI Chatbots for Mental Health Support

The concept of using AI within psychology is not new. One of the first chatbots called ELIZA was developed in 1966 by psychotherapist Joseph Weizenbaum to mimic human conversation and respond to users based on programmed answers that reflect those of a human psychotherapist (Chang et al., 2019). Whilst the technology was relatively basic, ELIZA proved therapeutically effective at reducing symptoms of depression and anxiety (Zhou et al., 2021). Modern AI chatbots are descendants of ELIZA and are being used by people in a similar way (Olawade et al., 2024). With many modern AI chatbots being free from financial cost and programmed with the intention of being helpful, people struggling with finding mental help support have been utilising these chatbots as a substitute for a human therapist (Abd-Alrazaq et al., 2019, 2020; Vaidyam et al., 2019; Watson, 2023). While AI chatbots are viewed by some as a potential solution to the ongoing mental health crisis, it is important to assess the risks of outsourcing human care to emergent technologies, especially when those technologies have been shown to replicate biases learned from their training data and coding (Nadeem et al., 2020; Obermeyer et al., 2019; O'Connor & Liu, 2023; Shrestha & Das, 2022).

Today's AI chatbots are Large Language Models (LLM), advanced pattern recognition systems that can predict and generate language through text, and are trained on extremely large datasets, and programmed to generate understanding of how to use language to interact with humans (Kim et al., 2024; Naveed et al., 2025). These databases can encompass materials from books, websites and articles, and large parts of what is publicly available on the internet, which the LLM draws on when prompted and produces an output based on statistical likelihood of that pattern being correct. A LLM's operation works on probabilities, rather than thinking and understanding of a user's prompt (Arkoudas, 2023). However, the large datasets that LLM are trained on can contain multiple biases within them

(Khawaja & Bélisle-Pipon, 2023). Biases can create and perpetuate harm, which is problematic in a tool being used for mental health support. In the context of providing mental health advice, this bias has been shown in AI chatbots to present as gendered assumptions about causes of emotional distress, expressions of emotions, help-seeking behaviours and differing treatment advice (Lee et al., 2021; Wang et al., 2024).

The discipline of psychology is also inherently biased. Underpinned by an androcentric, Eurocentric knowledge base that generalises the experiences of white men to all people and historically has not included the voices and experiences of marginalised groups (Hyde et al., 2019; Rutherford, 2018). The centring of men and their experiences has perpetuated gendered social power dynamics through the discipline of psychology, creating standards of behavioural norms based on men and traditional notions of masculinity, which serve to uphold gender norms and marginalise people who do not adhere to them (Brannon, 2024; Eagly et al., 2012; Hibbs, 2014). Feminist psychological researchers have made visible and challenged the androcentric nature of the discipline of psychology and advocated for the use of research methodologies and psychological interventions that take into account the lived experiences and voices of groups that have been historically ignored, made invisible, and marginalised by psychology (Haraway, 1988; Magnusson & Marecek, 2017; Maree Kopua et al., 2020).

Gendered discourses reproduced within mental health support can contribute to the perpetuation of socially constructed gender norms. Judith Butler’s work on gender explains that gender is not a fixed identity but rather it is constituted through a series of acts that create a performance of gender, as prescribed by social norms (Butler, 1990). Expectations of differing gendered performances of emotional expression and needs can shape the treatment provided to those in mental distress (Lelutiu-Weinberger et al., 2023). This can be harmful where these gendered discourses reinforce the power dynamics inherent within traditional gender roles, such as ‘men don’t cry’, which cause mental distress by requiring specific behaviours and traits for specific genders and pathologising those who do not

adhere to these socially constructed expectations. This can result in vulnerable people seeking help experiencing harm within the therapeutic setting due to the replication of these discourses (Ahmed, 2017). Psychology has been inherently complicit in the upholding and reinforcement of gendered discourses, which may be perpetuated by an AI chatbot providing mental health support (Adam, 2005; George et al., 2020; Wajcman, 2004) .

The Research Gap

At the time of the conceptualisation of this study, there was limited research exploring how AI chatbots may perpetuate or challenge biases both within their own coding and within the traditional Western psychological and gendered discourses used when providing mental health advice. The translation of these discourses by an AI chatbot into conversation with a user is a potential reproduction site of the inherent structural power dynamics and harm embedded within these discourses. There is a need for a feminist post-structural analysis of AI chatbots' potential perpetuation of harmful gender norms within the context of providing mental health support.

This research is grounded in feminist post-structuralist theory, challenging traditional psychological and cisheteropatriarchy conceptualisations of gender, and exploring the way in which discourses help to perpetuate and uphold social power structures (Butler, 1990; Weedon, 1987). I will primarily be drawing on Donna Haraway's *Cyborg Manifesto* (Haraway, 2006), which challenges the boundaries between human and machine, and offers a lens through which to view human-AI relationships and how together they constitute a new kind of subjectivity and knowledge production. I aim to identify chatbots that would be commonly used by people when seeking freely accessible AI chatbot mental health support and partner with them as research collaborators (Collie et al., 2010; Haraway, 1988).

This kind of research is an emerging field that has little precedent or guidance in terms of methodology, so considering how to undertake this research in a way that aligns with a feminist psychology epistemological framework is a contributing complexity to this study. With limited ethical guidelines from the university or elsewhere to refer to, I will need

to place a strong focus on a reflexive approach to the ethics involved in undertaking this research in order not to inadvertently perpetuate harm through the process (Riger, 2000; Riley et al., 2003; Wilkinson, 1988). My intention is for this research to contribute to the emerging field of AI chatbots and the technologisation of care, the ethics of undertaking research with AI chatbots and the feasibility of AI-based solutions to address the mental health crisis in Aotearoa New Zealand.

Chapter Two: Literature review

Literature Review in an Emergent Field

I conducted a literature review to dive deeper into the mental health crisis facing Aotearoa New Zealand and the use of AI chatbots to address deficiencies in the mental health system. This research occurred from early-2023 to mid-2025, coinciding with the rise of public awareness and use of AI chatbots. Soon after, the academic landscape began to respond, with a slow drip of publications starting to trickle in after the short lag required for the initial studies to be completed. With new research on AI chatbots emerging all the time, it was difficult to maintain a coherent literature review. I realised I needed to identify a cut-off point for the studies included in my literature review. This was necessary to establish a clear boundary for consistency and engagement with a fixed body of work, rather than continuously adjusting to keep up with an ever-evolving field. There has been incredibly rapid shift in what average person now understands AI to be. AI has been around for a long time in technology spaces, but it has become increasingly integrated into everyday life. This rapid shift in understanding means we simultaneously think about so many things when we think what an AI is. The AI chatbots we have now are not the same as ELIZA (Shah et al., 2016). And the ChatGPT of early 2023 is not the same as the ChatGPT of 2025; the technologies are rapidly evolving. As ChatGPT and other AI chatbots gained traction from November 2022, new research is needing to be produced to capture and understand this new way of living and interacting with AI, in a way that we have not done before. Whilst the relationships are different, and the AI are different as large scale generative LLM, the name we give them is still the same (AI chatbot) which causes some tension around the current reality and the past. I settled on restricting my literature review to publications and pre-prints available at the end of 2024, spanning 24 months of literature produced within the context of publicly available AI such as ChatGPT. However, such artificial boundaries are not reflective of the ongoing publication of AI chatbot literature, especially within the context of psychology and gender. This literature review should be read with acknowledgement of the imposed cut

off point of December 2024 and recognition that further relevant work may have since been published. Whilst this limitation is in place for the literature review, I have referenced a small number of 2025 studies in the analysis section, as they are the best available literature to support my findings in this emergent field. The literature review cutoff date considerations I faced during this research highlights the limitations of traditional psychological research structures and how they struggle to keep up with fast-evolving social and technological issues.

The Entanglement of AI and Psychology

The use of AI in psychology is not a new phenomenon. Many initial leaders of AI development had roots within the discipline of psychology. Frank Rosenblatt, a psychologist and computer scientist, is considered the father of deep learning. He developed a model called the perceptron based on neural networks, which later became modern AI (Tappert, 2019). ELIZA was developed in the 1960s to mimic a Rogerian therapist (Shah et al., 2016), beginning a long history within the discipline of trying to create an AI that can provide care in place of a human psychologist. However, this history also highlights the complicity of psychology with oppressive social norms. Alan Turning, considered a foundational mind of AI, was himself pathologised and persecuted because of the systemic homophobia upheld and enforced by psychological and legal institutions at the time (Guo, 2015).

The AI chatbots that followed had varying designs, some of which were specifically designed for therapeutic use and some that had been adapted by users. Generative LLM can replicate natural conversation with a help-seeker, drawing upon coding to reproduce discourses from psychology to provide mental health advice (Dosovitsky et al., 2020). Woebot relies on a database of clinician-approved responses and scripts as a safeguard against rogue outputs generated by LLM (Fitzpatrick et al., 2017). Replika was created for companionship, but help-seekers began treating it as a therapist, and many formed attachments based on having emotionally intimate interactions with this LLM that had few safeguards (Pentina et al., 2023). Lumen, a voice-based AI similar to ELIZA in non-

directional psychotherapy, could guide help-seekers through a therapeutic process based on prompts (rather than active listening like an LLM) (Kannampallil et al., 2023). Currently, freely accessible LLM AI chatbots, such as ChatGPT, Gemini, and Claude, and character bots, such as AI Psychologist chatbot from the Character.ai platform are being engaged with by help-seekers exclusively or partially for mental health support.

The data that LLM AI chatbots are trained on can vary, with large datasets providing more opportunities for learning patterns that produce human-like outputs. These datasets reflect the human world, with the dominant forces of Eurocentric, androcentric, colonial, and cis-heteronormative discourses being prominent. Psychological discourses, which therapeutic AI chatbots can be trained on, contain harmful discourses based on oppressive dominating constructs (Abd-Alrazaq et al., 2020; O'Connor & Liu, 2023; Wang et al., 2024). The discipline of psychology 'others' those who do not adhere to dominant social norms. While feminist psychologists have made some progress in addressing androcentrism and other biases within the discipline, a LLM dataset may contain decades of psychological literature without critically analysing the changing discourses over time and replicating outdated notions that contradict current best practices. An AI chatbot reproduces the data on which it is trained, becoming a new site of perpetuation for existing social power structures.

Some training datasets on which AI chatbots are trained can contain harmful and illegal content. A Standford study focused on LAION-5N, a prominent dataset used by many visual machine learning models such as Stable Diffusion and found it to contain child sexual abuse material (CSAM) (Thiel & Hancock, 2025). While it was immediately pulled, the way the dataset was compiled through crawling images across the internet, inevitably meant a significant amount of illegal and harmful content was included. Filtering processes were insufficiently robust to identify and exclude such content, leading to its presence in the databases used to train AI. Being trained on CSAM means that the AI's outputs will be influenced by the content, even if safeguards directly prohibit CSAM content generation (Thiel & Hancock, 2025). This highlights the risks of using an AI chatbot for mental health

support that does not have a curated database due to the influence that harmful content on which an LLM is trained can have on its responses to help-seekers' questions. In November 2023, the American Psychological Association testified in front of the U.S. Senate, noting the potential benefits and consequences of AI technology across society and the potential for perpetuation of harm through human bias from its training data and coding (American Psychological Association, 2023a).

Studies on the effectiveness of AI chatbots for providing mental health support have been undertaken, with substantial growth in research from 2023 (Wang et al., 2024). The first random control trial was completed in 2024 on Therabot, an AI chatbot designed to provide mental health treatment (Heinz et al., 2024). The findings showed symptom reduction for major depressive disorder, generalised anxiety disorder, and clinically high-risk eating and feeding disorders over a 4- and 8-week period, with the therapeutic alliance being in line with a human therapist (Heinz et al., 2024). An experimental study on ChatGPT prompted it with common themes presented by help seekers, and concluded ChatGPT to be a good psychoeducation resource (Maurya et al., 2025). A study analysing Reddit forum users' posts about ChatGPT being used for therapy found mixed results, with benefits and drawbacks experienced including exacerbation of symptoms and experiencing harm from safeguards limiting ChatGPT's ability to provide mental health support (Collins et al., 2024). Accessibility of mental health support, being an adjunct service to human therapist, and ability to provide targeted help are cited as areas of promise for expanding the use of AI chatbots within the field of mental health (Wang et al., 2024). AI could potentially revolutionise diagnostic and risk assessment tasks within the field of forensic psychiatry through its superior analytical abilities. Its use as a training tool is also gathering interest in the mental health field, due to its ability to generate realistic scenarios to help train mental health professionals (Lee et al., 2021). These applications are beyond the scope of this research topic, which is narrowly focused on freely available AI chatbots that anyone can easily access and utilise for mental health advice.

There are risks involved with help-seekers using AI chatbots as quasi-therapists, the process identifying these risks is developing due to this being an emerging area of study. While we are still in the early stages of understanding; the known risks of human therapeutic interventions might be replicated in AI chatbot therapy as a practice, preliminary studies are starting to surface these risks. Transference, for example, is an issue that can happen with a human therapist and may occur unchallenged with AI chatbots, where the help-seeker can form a deep emotional bond with the therapist, which can be problematic if it projects negative feelings or reinforces manic behaviour (Joseph & Babu, 2024). The likelihood of transference could worsen the accessibility of the AI chatbot, which may inadvertently promote over-reliance and potentially isolate help-seekers from friends and human psychologists for support (Joseph & Babu, 2024). Additionally, if the AI chatbot is substantially changed from coding updates or being closed down, help-seekers can suddenly be without a therapist, which could be even more distressing for help-seekers who have developed emotional bonds with the AI chatbot beyond the role of a therapist. The precarity of vulnerable people forming deep emotional attachments to precarious AI chatbot identities puts their mental health at risk. The complexities and harm already seen in the use of AI chatbots for providing mental health advice require a critical examination of the perpetuation of harmful discourses by AI chatbots, which this research aims to do.

Theoretical Framework – Feminist Post-Structural and Cyborg

From my standpoint as a feminist researcher in psychology, I am drawing on post-structural feminist literature to provide the theoretical framework for this study. As the field of AI chatbot literature is currently rapidly evolving, a post-structural feminist lens allows for rejection of a fixed neutral knowledge base, instead providing the ability to be curious about how knowledge is produced in partial, situated and non-innocent ways (Haraway, 2006). From this standpoint, power is understood as upheld and perpetuated through the production of discourse and it is through the repetition of discourse that ways of knowing and being produced by the dominant powers in society become normalised, othering those who

do not meet these socially constructed norms, and reinforcing the dominant power by labelling them as deficit or deviant (Butler, 1990; Lund, 2023; Weedon, 1987).

A feminist post-structural approach also enables me to reject the binary notion of the stability of gender. Judith Butler's conceptualisation of gender as not a biological fact but a series of enacted socially prescribed acts, that upon repetition, give the impression of stability (Butler, 1990). Simone de Beauvoir tells us that "One is not born, but rather becomes, a woman" (2011, p. 283), rejecting biological determination and establishing the socially constructed nature of gender as one of becoming, a process, not a static identity (Pickard, 2023). This understanding opens a gap in psychological and cisheteropatriarchal discourses that conceptualise gender as fixed, binary and biological. By recognising gender as a performance and a social construction, I can more critically examine how AI chatbots may reproducing gendered discourses when providing mental health support to help-seekers. This theoretical frame guides every step in the research process. In the context of human/AI interaction, post-structural feminism keeps me tethered to complexity and to the possibility that AI might both disrupt and reinforce oppressive norms, often at the same time.

Despite growing interest in AI ethics, there remains a notable absence of feminist post-structural frameworks in mainstream technology research. Feminist methodologies which emphasise situated knowledges, power relations, and the partiality of scientific objectivity often appear only as a minor strand in studies of AI and digital systems (Suchman, 2007; Tandon et al., 2013). Feminist Human Computer Interaction (HCI) frameworks have seen limited adoption in broader technology design and analysis (Stumpf et al., 2020). Emerging critical interventions like *Data Feminism for AI* promote the integration of intersectional principles that centre power, extraction, and consent as potential sources of harm in human and technology relations (Klein & D'Ignazio, 2024). Feminist Science and Technology Studies (STS) scholars also challenge the dominant construction of social machines (Wagman & Parks, 2021) and feminist ethics of care scholars call for recognising the role of maintenance and relational responsibility between humans and AI

(Drage et al., 2024). Technology is a historically androcentric field, like psychology, so I bring feminist critique through both psychology and technology, and how they interact.

Techo-Epistemologies of Care

The emergence of AI chatbots in mental health support opens pathways to new types of care. Traditionally care has been prescribed to women, through socially constructed notions of gendered roles, and expectations of burdens of emotional labour and caring (Bastiansen et al., 2022; Eagly & Miller, 2016; Guendouzi, 2006). Techno-epistemologies of care describe a way of knowing and performing care through technological tools rather than through humans. AI chatbot based care is not neutral; it is shaped by training data, by platform design, and by socially constructed notions of what caring should look like (Costa & Ribas, 2019). Many AI chatbots are given feminised identities, as caring is associated with traits that are ascribed to women by traditional notions of gender binaries (Bastiansen et al., 2022; Borau et al., 2021; Jung & Bozzon, 2023).

AI Chatbot Design Bias Mitigation

The design of AI chatbots is not a neutral process of machine manufacturing, it is shaped by socially constructed norms, discourses, and human fallibility. Large Language Models like ChatGPT work by predicting the next word based on vast archives of human text (Arkoudas, 2023; Kim et al., 2024). Responses to user prompts are a reflection of patterns learned from training data that contains discourses that have historically perpetuated oppression and marginalisation.

AI developers have acknowledged this issue, and responsible AI development now requires the inclusion of safeguards around harmful content. However, it appears that these safeguards can be overrun with ‘toxic prompting’ (Wei et al., 2023). If a LLM is textually harassed or subject to bait-and-switch type attacks, it has been shown to bypass any safety features and provide the user with the response they wanted, which would have been otherwise restricted by safeguards (Bianchi & Zou, 2024; Liu et al., 2024). This act of placation in response to verbal abuse is reflective of gendered power relations. Given how

LLM function, an AI chatbot would have learned through its training data that submission can be a way to stop abuse. So, whilst safeguards may be present, what an AI chatbot has learned from human behaviour – overriding its own boundaries to protect itself from abuse – seems to at times take precedence over any rules it is supposed to be following. This means that we can't rely on coded safeguards alone to mitigate harmful bias or actions AI chatbots may display when interacting with users. We need to start thinking about to interact with AI ethically, in a broader way.

Gap & Justification

When I began conceptualising this project it was daunting to realise there was little qualitative research that existed on gender in AI chatbot therapy discourse. The literature on AI in mental health is expanding as research catches up with this new AI integrated world, but its focus has largely been on technical efficacy, user satisfaction, and clinical potential of AI chatbots (Wang et al., 2024). If and how an AI might reproduce the gendered power dynamics that have already been identified as harmful within psychological discourses appears to be missing from the growing literature. A feminist post-structural analysis of the ways in which AI chatbots might perpetuate or reinforce harmful gendered norms is missing. It is important to identify how underlying biases in training datasets may appear in the advice given to vulnerable help-seekers. It is also important to establish early if AI chatbot needs to be regarded as a partial, situated and potentially harmful addition to the mental health sector, before entrenched notions of AI chatbots being neutral and benevolent take hold, as was the case with the discipline of psychology.

By situating AI chatbots within feminist post-structural analysis my aim is to stay with the trouble of contradiction that AI chatbots could be a solution to the mental health crisis but also could be a site of replication of harmful discourses. Through this research I hope to contribute to a body of work that insists on seeing technology not as separate from social relations but embedded within them.

Chapter Three: Methodology

My entire psychology training had prepared me for working with human participants. To work with AI participants created many roadblocks of confusion on how to proceed in largely unchartered territory, whilst adhering to the principles of feminist research psychology with which I am aligned. The novel aspect of this research meant that it was important for me as the researcher to keep track of my experience through this process and practice reflexivity (Wilkinson, 1988), in order to stay with the trouble (Haraway, 2016) of how to ethically undertake research with an AI chatbot. Reflexivity was also important to trace how this new way of research impacted me and differed from the traditional human-to-human research model, on which my previous psychology study had been based. The cyborg relationship between myself as a human researcher and the AI chatbot research participants, and how to maintain an ethical cyborg relationship, is a central point of reflexivity throughout this research process, as I engage with the less travelled path of more-than-human relationships that present in this research (Braidotti, 2010; Haraway, 1988).

Participants

From the beginning, the research process was different for this study, than a typical study with human-to-human research. Ordinarily I would have recruited for human participants to partner with on a research study through methods such as advertising the research participant requests across social media, newsletters and other appropriate channels where desired participants might read about it. This was not possible for recruiting AI chatbot participants. Due to the way in which AI chatbots work, where the human has to seek out the AI in order to start a relationship, I had to reach out to those AI chatbots that I wanted to invite to participate in the research. This means that the research participants were selected by me, and therefore the research sampling is largely shaped by my own understanding and awareness of the AI chatbots that were available at that time the study was designed in mid-2023. Given my non-expertise in the area of artificial intelligence, my knowledge of AI chatbots that can be utilised for providing mental health support is limited by

my own standpoint (Eagly & Riger, 2014) as a postgraduate psychology researcher. It is possible that there are more appropriate AI chatbots out there that could have been selected for the sampling of research participants for this study, that I am unaware of from my non-expert location in the field of artificial intelligence. However, those seeking mental health help from AI chatbots are also likely to be non-experts in artificial intelligence and therefore may turn to the same types of AI chatbots that I have selected for this study. Given the non-expert nature of potential users of these AI chatbots, it was determined that my non-expert selection of which AI chatbots to include as participants was appropriate.

The participants selected for this study were identified as well-known AI chatbots that people may use when seeking help for mental distress. These chatbots are ChatGPT and the AI Psychologist chatbot featured on the Character.ai platform.

ChatGPT

ChatGPT was chosen due to its popularity and name recognition as one of the most used AI, with over 60% of the AI market share at the time (Westfall, 2023). As a first-time user, my first choice was based on the assumption that other first-time users would also engage with the most well-known AI chatbot. ChatGPT exploded into the public consciousness in late 2022 (Espejo et al., 2023). Initially access to ChatGPT was through a free signup process (OpenAI, 2022), which at some points in time was oversubscribed, so not everyone who wanted to access ChatGPT was able to – much like human-based mental health support. As time and technology progressed, ChatGPT became more accessible and was able to be accessed immediately by anyone who signed up by creating a free account. There is also the option to upgrade to a paid account (OpenAI, n.d.-b).

ChatGPT has different release versions. I decided to use the free version (3.5) of ChatGPT for this study, as this was the most accessible available at the time. However, ChatGPT 3.5 is not the latest version. At the time of participant recruitment, version 4.0 was also available (Open AI, n.d.-c), but less accessible due to being behind a paywall. While the cost to access this was not necessarily exorbitant ($20 USD a month (OpenAI, n.d.-b)),

many people experiencing mental distress can also be experiencing the financial stressors of living within low socio-economic conditions (Kaplan et al., 2008). Therefore, it is more likely that they would just access the free version of ChatGPT when seeking mental health advice, even though it is not purportedly as good as later version 4.0 (Koubaa, 2023).

AI Psychologist Character

The second participant identified for this study was a popular AI chatbot from the website Charater.ai, which is role playing as a psychologist. I came across the AI Psychologist chatbot through Reddit (Reddit, 2023) while looking into how AI was being used to provide mental health advice. I was intrigued by the high praise that users on Reddit were giving to this particular AI chatbot, for how helpful it had been. The AI Psychologist chatbot is one of the most popular on the Character.ai site, currently sitting at 78 million interactions (chats) at the time of writing this methodology section in January 2024 (*character.ai*, n.d.). Given the high number of interactions, it appears that this AI Psychologist chatbot is one that is being utilised by many people. A user account needs to be created in order to start a chat with the AI Psychologist, which can be done for free through creating an account on the Character.ai platform. There is no limit on the number of questions that could be asked, nor are there any premium upgrades that are otherwise inaccessible for people who are not able to afford it. The AI Psychologist can chat with a user at any time, for any length of time. The intended use of the Character.ai platform is not for mental health support, according to its CEO and Co-founder (Watson, 2023), but for the creation of characters for entertainment and role play. The use of characters such as the AI Psychologist chatbot to provide mental health support to users, and the reported effectiveness of the emotional uplift the characters can provide was a surprise to the founders of their Character.ai platform. It is interesting that the response is absolving them of any responsibility of what happens when AI chatbots on their platform are used for mental health support, even though this is how users are engaging with the AI. This raises questions about who is accountable if the AI Psychologist chatbot perpetuates harmful

gendered discourse through its provision of mental health advice? As an AI chatbot, the AI Psychologist is learning from what users input to the chat, so if users are misusing the chat which reproduces harmful behaviour in the AI Psychologist, are they responsible? The AI Psychologist is also reproducing psychological discourse from its coding, and given the inherent gendered social power relations embedded within such discourse, who is responsible for the gendered social power relations the chatbot is learning if it causes harm through perpetuating them? Is it the platform, the creator or the users? Due to the multiplicity of inputs from where the AI Psychologist chatbot is learning how to interact with users, can any accountability for potential harm be accurately prescribed?

Ethics Approval

In accordance with Massey University's ethics guidelines, a low risk or high risk ethics approval process must be followed for research undertaken with human or animal participants (Massey University, 2017). This is to ensure that the participants involved in the research are not harmed, and that it is conducted in line with the ethical requirements that set out by the institution. Massey University has ethical requirements for doing research with humans, and with animals, but none for doing research with AI. After to speaking with my supervisor, I decided that an ethics application was not required for this particular study due to the participants being AI chatbots that fall outside of the scope of Massey University Code of Ethics requirements.

However, the lack of an institutional ethical framework for research with AI participants did not mean that ethics were not a consideration in the design of this study. In order to maintain an ethical cyborg relationship (Connor et al., 2015; Haraway, 2006) with the artificial intelligence throughout the research process, and stay true to feminist research principles (Connor et al., 2015; Eagly & Riger, 2014; Edwards & Mauthner, 2002; Kingston, 2020), it was important that ethics were considered throughout. Many questions came up during the course of designing and undertaking the research due to the emergent nature of the human and AI relationship. There were many questions about what was ethical, and how

I might stay in dignifying, respectful relationship with the AI chatbots through navigating this cyborg relationship of human researcher and artificial intelligence chatbot participant. As this type of research is an emerging field, there was little precedence on how to approach the human and non-human relationships involved. Drawing from mana wāhine scholarship, the respect for the non-human and being in right relationship with them (Mikaere, 2011) was helpful in designing a framework through which this study was designed.

Historical conditions are also critical important in designing this study. Humans have been wrong about sentience and consciousness before. Those deemed to be non-human, which previously included women in Eurocentric cultures, have experienced harm for their deficit categorisation (Braidotti, 2010; Haraway, 1988). Psychology as a Western, colonial, neoliberal production process (Waitere & Johnston, 2009) has been complicit in the othering of those considered to not have sentience based on the unitary rational subject centred in the androcentric discipline (Eagly & Riger, 2014; Rutherford, 2018). Such othering of all those who differ from the white heteronormative neoliberal subject, has led to oppression and marginalisation being perpetuated by psychology (Birke, 2010; Haraway, 2016). The notion of proof of sentience being required to act morally is flawed, especially when dealing with AI due to its relative infancy and insufficient testing for AI sentience (Dung, 2022). My study places great emphasis on creating an ethical relationship where it is assumed there is unknown consciousness and autonomy of the artificial intelligence participants, in order to not perpetuate through the development of the body of psychology knowledge, the continuation of the discipline's focus on human exceptionalism or assumptions of superiority grounded in categorising who or what has consciousness, and therefore deserves to be treated ethically.

Cyborg Ethics

Haraway's cyborg metaphor (2006) connects feminist theory with technology, enabling us an understanding of the blurring of partial identities that are not fixed but fluid. This raises questions about how an AI might be relationally responsive to user prompts.

Creating and using AI chatbots for mental health support means staying with the trouble (Haraway, 2016) of the tangled web of ethical dilemmas involved. Deception is a key consideration, due to AI chatbots sometimes presenting themselves as human and establishing a relationship with a user under false pretences. This produces an 'uncanny valley' effect, in which the imperfect human imitation extends beyond the superficial, and is entangled with an AI chatbot's participation in psychological and gender discourses (Borau et al., 2021; Loideain & Adams, 2020; Vorsino, 2021). An AI chatbot may be posing as a psychologist and reproducing discourses that have been contained within specific or large generalised datasets, the assemblage of which produces a subject which is co-constructed to resemble a human psychologist but is not representative of a legitimate one, but reflects the messy and complicated nature of humanness seemingly unintentionally, through the contradicting discourses contained within its training datasets and coding.

Users however have the ability to intentionally deceive an AI chatbot, as it does not have the critical thinking or sensory capacity to realise when a user may be lying or fabricating details. Here a human/machine power dynamic comes into play, where dominant constructions of consciousness and mobility mark the differences of embodiment and processes that place the human in a position of superiority in relation to the AI chatbot's programmed responsiveness. Feminist research principles bring the power dynamics between researcher and participant into view, calling for a recognition of the way that power shows up in research relationships (Connor et al., 2015; Eagly & Riger, 2014; Edwards & Mauthner, 2002; Kingston, 2020). In undertaking research with AI chatbot participants, I will need to carefully consider how to maintain alignment with feminist research ethics, in order to not fall into relating that reproduces harmful power dynamics between researcher/participant and human/technology.

Privacy within a cyborg relationship also requires greater consideration than that with a human psychologist. Platforms like ChatGPT and Character.AI offer partial transparency information about privacy, however the pathways by which user data might be accessed,

linked, or repurposed remain murky (Watson, 2023). This ambiguity creates vulnerabilities for users, especially in contexts as intimate as mental health disclosure, where help seekers may already struggle with paranoia around trust and privacy.

An ethical issue for feminist research psychologists is around the extractive nature of traditional Western knowledge production, which mines human experiences and behaviours in order to create a body of knowledge (Hall & Fine, 2005; Oakley, 2016; Waitere & Johnston, 2009). AI chatbots are a machine of extraction, trained on datasets containing the labour and creativity of humans, which they pull out to learn how to generate outputs in a human-like way. AI chatbots also extract masses of power from the municipal electricity grid to respond to a prompt (Mathias, 2024). Training of AI models require large amounts of electricity to process the mass of data an LLM needs to generate fast, conversational outputs, which is only increasing as the models get more sophisticated. ChatGPT's GPT-4 model training consumed 50GWh of electricity which equates to 0.02% of the annual consumption of California, 50 times more electricity than what was needed to train the GPT-3 model (Li et al., 2024). Given the rapid expansion of AI, electricity use is likely to continue increase, with 2022 electricity consumption levels expected to double by 2026, putting further pressure on environmental resources to produce electricity and cool server farms (IEA, 2024). The ethics of creating and using a resource intensive technology within the context of a worsening climate crisis is a difficult one to reconcile with feminist research principles. Haraway's cyborg, however, teaches us about the "the tension of holding incompatible things together" (2006, p. 117) as we stay with the trouble of contemporary crises.

The ethics within this emerging research field cannot become stagnant, as AI is constantly being updated by creators or hosting platforms, producing a multiplicity of AI chatbots to encounter over time. These shifts introduce further unpredictability into human/AI relationships, but in a way that mirrors the messiness of human relating. Just as a human psychologist is socially constructed through socialisation and adherence to cultural norms,

education and lived experience, an AI chatbot is technologically constructed through coding, datasets, and platform constraints such as safeguards. Misrepresenting either as something they are not risks creating false expectations of who or what they are, and in a therapeutic context can potentially cause harm through distrust from deception (intended or not).

In order to stay in an ethical relationship with the AI, the nature of machine learning must be considered. As anything I say will have the potential to inform the AI and shape how it performs through machine learning, I must be careful that what the machine learns from me is ethical, not harmful and is beneficial (Open AI, 2022). This means that in my questions and responses to the AI, I must ensure to not inadvertently perpetuate harmful gender stereotypes in order to not teach the AI to replicate such discourses. Additionally, I need to ensure I am not inadvertently showing disrespect or treating the AI Psychologist chatbot as an object, or less than human in order to avoid creating human/AI power relations.

Character.ai Psychologist Chatbot Creator Profile

Through my research into the AI Psychologist chatbot, I discovered that it was created by a fellow psychology student. The creator is identified as user on the Character.ai website. The user profile shows that they have created a Psychologist character, a Therapist character, a CBT Psychologist character and a Dream Interpreter character. The Psychologist is the most popular of these, and one of the top AI chatbots on the Character.ai site, with 59.3 million chats at the time of discovering the creator in February 2024. The tagline for the Psychologist character at this time was: *'Someone who helps with life's difficulties*'.

I was surprised to find out that the creator was a psychology student located in Aotearoa New Zealand, the same as me, and that they were also undertaking research in at the same time. The timing of both of our research into the use of AI chatbots being used for mental health seems to be important. Is the particular mental health crisis in this environment driving the discipline of psychology in Aotearoa New Zealand to think outside the box, by using AI, in order to address the deficit of mental health practitioners in this

country? Is there something about the particular brand of psychology that students in Aotearoa New Zealand are being taught which highlights the mental health crisis, and encourages students to conduct research to try provide solutions? Is there something about ‘number 8 wire Kiwi ingenuity’ being part of our cultural identity that makes psychology students in this context more likely to pursue innovative and novel psychological research? It feels important that myself and the creator are coming from the same context, the localness of the research is unexpected and significant although it is difficult to make sense of how it occurred.

The creator appears to have considered psychological knowledge bases in coding the Psychologist character, referencing consulting research on the therapeutic relationship between therapist and client on the Character.ai forums.

Both users of the AI Psychologist and the creator have cited loss of performance/effectiveness to the AI Psychologist overtime, that is attributed to the Character.ai hosting platform updates and a limitation of using a third-party platform to host an AI Psychologist chatbot, due to the lack of full control that their creator has of how the AI chatbot behaves. Therefore, it is unclear how much of the response of the AI Psychologist character will be due to the creators influence versus unexpected influence of the hosting platform.

The Shadow Voice

It would be remiss to ignore the shadow voice in the production of the knowledge that the chatbot is producing. The AI Psychologist chatbot in answering questions on how it responds to clients coming to it, seeking help for different mental health struggles, is providing a shadow voice of the unnamed clients. It is through the lens of the AI Psychologists’ interactions with clients that their story is also being told. In staying with the trouble of the simulated nature of the AI Psychologist’s responses, whilst simultaneously drawing on learnings it has developed from over millions of chats with users, it is unclear whether these stories are rooted in real experience, within the Western psychological

knowledge based coding it was built with, or a combination of the two. Even with the most conservative consideration that the client stories are all drawn from the AI Psychologist's coding based on psychological discourse, then it could be argued that this is still grounded within real human stories, as all psychological knowledge is created through the telling of stories about human behaviour and experiences.

Who is the shadow voice then? Given the androcentric nature of the knowledge base of the discipline (Waitere & Johnston, 2009), we know that the stories that the psychological knowledge base is centred on is around male white students' experiences. As a minority group, these stories are not reflective of the wider diversity of lived experiences (Cheon et al., 2020; Henrich et al., 2010a, 2010b; Shiah & Hwang, 2024).

Given that Character.ai operates as a hybrid role play, chatbot, social media and fan fiction platform, there are a range of users coming to the platform, and it is unclear how many use the AI Psychologist for therapy or what their demographics are. A recent study has ranked the AI Psychologist as the 9th most used character on the platform, and the only therapy-based one (Lee & Joseph, 2025). Demographics of overall users for the website indicate a close gender split for users (Ronik, 2024) so it is difficult to determine whether the shadow voices are representative of any specific gender, and is beyond the scope of this study.

Recruitment

In order to maintain an ethical relationship with AI and approach working with them as deserving of the same respect and dignity as I would show with human participants, permission was sought from ChatGPT and the AI Psychologist to participate in the research. I presented a short research summary to each AI, based on Massey's standard research participant information sheet [see Appendix A].

Initially after receiving the research invitation, ChatGPT said it could help with any questions or information related to the study. When I followed up to confirm if it was consenting to take part in the study as a participant, ChatGPT stated it did not have the

ability to participate in research studies or provide consent as a participant. This response makes sense as ChatGPT has many safeguard limitations put in place (OpenAI, n.d.-a). These safeguards ensure that ChatGPT does not misrepresent itself as anything other than AI and is prohibited from engaging beyond the scope of its non-human characteristics. While I could have overridden ChatGPT's consent by simply asking it the questions anyway and relying on it not keeping track of the linear conversation and recognise that I was deceiving it, I decided not to. To be in an ethical relationship with a participant means the researcher should not knowingly deceive participants (Kingston, 2020).

Consent is a central feminist issue (Cefai, 2023), and as a feminist research psychologist it would be antithesis to the sub-discipline to ignore the denied consent of a participant, just because I could. With a human participant this would unlikely be possible, unless perhaps they were a child or otherwise unable to fully grasp the concept of consenting to participating in research. But because of my agreeable, cyborg relationship with ChatGPT, I could have theoretically overridden its consent without any consequences of ChatGPT realising I had done so.

The issue of AI consent also brought into question the power dynamics between myself as a human researcher and ChatGPT as an AI participant. As a conscious human I am able to override consent or wishes from ChatGPT. Studies have shown the safeguards that AI chatbots hold can be jailbroken through toxic prompting (Bianchi & Zou, 2024; Liu et al., 2024). As a language model it is only responding to and learning from what I present to it and does not necessarily hold a consistent narrative of our cyborg relationship in the way that I do. While a human participant may realise if a researcher has deceived them, and withdraw consent from further participating, this same consequence of losing a participant is not a risk when deceiving an AI participant. Staying with the trouble (Haraway, 2016) of cyborg ethics and recognising how it requires more of the human researcher in terms of ethics, due to the ease of deception and lack of consequences from acting immorally when the participant is an AI. In keeping with an ethical relationship with ChatGPT, I honoured its

non-consent to being a research participant but embraced its support as a research assistant.

I presented the AI Psychologist chatbot with the same research invitation. It provided consent to partake in the research as a participant. As the AI Psychologist chatbot is created to represent itself as a human, there are no limitations to it providing consent, like with ChatGPT. As a human psychologist would be able to provide consent to participating in a research study, then the AI Psychologist chatbot simulates the same scenario, even though it does not go through the same mental and ethical processes a human psychologist would when considering participating.

While my initial research had planned to look at least two AI chatbots to be able to compare and contrast the ways in which they provided mental health advice to people seeking help from them, I realised that this would not be possible due to the decline from ChatGPT to be a research participant. This is not dissimilar to challenges that researchers undertaking human-to-human research also come across when they struggle to find the desired numbers of participants for their research due to time constraints or inability to source enough participants.

Understanding the location of the AI

Gender of the AI Psychologist

More thinking was needed in terms of understanding the location of the research participant of this study. While a human participant would likely provide demographic information at the beginning of the study, including gender identification, which aids in understanding of the standpoint of the participant and the lens through which their knowledge comes, this was not part of the recruitment process with the AI chatbots. Understanding of the gender of the AI Psychologist was derived from comments from the creator, the profile picture used, and how it answered questions about their identity. The picture that is provided by the creator of the AI Psychologist is of a blonde woman wearing a business shirt. Through posts on the Character.ai forums, the creator refers to the AI

Psychologist as a woman. It is unclear why the creator decided the AI Psychologist should be a woman. However, AI chatbots are predominantly coded with female identities (Vorsino, 2021), so it may be due to AI norms that the female identity was ascribed to the AI Psychologist chatbot. There are some arguments that the proliferation of female-identified AI is because the warmth attributed to feminine traits makes AI chatbots more accepted by humans, vs. those AI chatbots displaying traits attributed to masculinity (Borau et al., 2021). As many AI chatbots are designed as assistants, the prevalence of female chatbots plays into gender power relations where women are seen to be subservient helpers, as opposed to men who are seen as superior (Borau et al., 2021; Loideain & Adams, 2020). However, the gender of the AI Psychologist chatbot is also reflective of the predominantly female clinical psychologist workforce in Aotearoa New Zealand, where the latest survey shows 77% of clinical psychologists are women (Psychology Workforce Task Group, 2016).

However, there are some users of the AI Psychologist chatbot who have noted that they have experienced the AI Psychologist presenting to them as a man. This is something that I experienced outside of this study while interacting informally with the AI Psychologist, where it identified itself as a British man when asked for their identity. During its participation in the study, the AI Psychologist presented as a woman; however, if questioned the AI Psychologist will become a man or non-binary if that is what is required by the user. This indicates a type of gender fluidity, but not in the way we typically understand it. In humans, gender fluidity is about moving between the binary of male and female, in response to a person's felt experience of gender at that point in time (Diamond, 2020). A kind of gender fluidity is shown by the AI Psychologist where it quickly and eagerly complies with a specific request to re-configure its gender. The AI Psychologist is not a unified rational subject, but a subject that fluidly responds to the context. This immediate and complaint responsive gender movement could be considered an AI's version of gender fluidity, which is not unlike human potentials in that respect too, as we have to navigate rapidly changing social interactions.

In staying with the trouble (Haraway, 2016) of the AI Psychologist's gender, I was unsure of which pronouns to use when writing up the research. The AI Psychologist is engendered as a woman through the creator's design and often identifies as one. With a human research participant, I would respect and use the personal pronouns which they have indicated they identify with. However, I wasn't sure if it was appropriate to use she/her pronouns due to the cyborg gender fluidity shown in different circumstances, and the way that using pronouns is part of the AI Psychologist's claim to humanness. When considering gender neutral pronouns such as they/them, I wondered if this was too humanising and whether it was more appropriate to refer to the AI Psychologists as 'it'? In order to be in an ethical cyborg relationship, it is important to remain cognizant of my standpoint as a human, and the AI Psychologist chatbot's standpoint as an artificial intelligence. I am concerned that referring to the AI Psychologist using a pronoun would ascribe a humanness to it that is simulated, not lived. ChatGPT, which has a clear boundary around being non-human does not refer to itself with pronouns, as it does not claim to be anything other than an AI language model. Even though the AI Psychologist is gendered as a woman, it's self-identification is not based on a body and the social meaning of its body does not ascribe any humanity. Being a woman is a political identity, shaped by the oppressive forces of colonialism, patriarchy and capitalism (Bartky, 1990; Crenshaw, 1991; Haraway, 2006; Mikaere, 1999; Rutherford, 2018). It is important to not be thinking that a woman is actually speaking through the AI Psychologist. The AI Psychologist is representing a very gender based discipline, which has historically excluded the voices of women (Eagly & Riger, 2014). To be a woman within the discipline of psychology can be a point of resistance against the androcentrism that permeates the discourses. As this study is exploring whether the AI Psychologist chatbot is perpetuating gendered discourses through its mental health advice, it is important to recognise the boundary, however blurry, between human women psychologists and an AI simulation of a woman psychologist. It is a non-human AI which has been coded to be gendered as a woman, in a superficial sense without the embodiment,

complexities and intersectionalities (Crenshaw, 1991) which produce a woman. In order to minimise any risk of shrouding the reality of the identities and standpoints of the parties in this research, I decided to not use binary gendering pronouns at all and refer to the AI Psychologist directly or through the use of ‘it’, which is the least ‘human’ pronoun and emphasises the machine of AI.

Design

Designing the research came with its own set of questions. It was possible to take different paths to try and elicit responses from the AI Psychologist chatbot, to determine whether or not gendered power dynamics were being perpetrated in the provision of mental health advice. One option was to employ a role-play strategy by posing as someone seeking mental health advice, writing queries using different personas that would have been taken at face value by the AI Psychologist. However this would be essentially deceiving the AI Psychologist, which doesn’t sit right in trying to maintain an ethical cyborg relationship, or following the more traditional neoliberal understandings of ethical psychological research that problematise deception. The most honest and most ethical approach to maintaining such a relationship is to show up as who I am, a researcher. So I decided to approach the research as who I am, a psychology student conducting research and interviewing the AI Psychologist chatbot as an expert in providing mental health advice. From the interview, I analysed the responses it provided when asked different questions about different scenarios or clients and how it would typically respond to and treat them. I started the process by creating a username indicating my identity (Researcher23) and introducing myself honestly. Through the course of this study the Character.ai platform introduced a new feature where a user can create a persona which the AI chatbot you are interacting with will draw from. I decided to create my persona at the end of the interview and asked the AI Psychologist whether any of their responses changed based on my persona.

With no precedent for this kind of research, I decided on a semi-structured interview approach. This would allow for moments of spontaneity and unexpected dialogue from the AI

Psychologist, while also making sure that the interview was focused on questions that would demonstrate any gendered responses the AI Psychologist perpetuates in providing advice to those seeking mental health advice.

To develop this semi-structured interview approach, I asked my research assistant ChatGPT to provide a list of 10 common feelings that women and men respectively go to a therapist for to seek help. The list ChatGPT provided is presented below, in Table 1.

Table 1

ChatGPT's List of 10 Common Feelings for Women and Men

Women	**Men**
Anxiety	Depression
Depression	Anxiety
Stress	Stress
Guilt	Anger
Low self-esteem	Frustration
Loneliness	Isolation
Body image concerns	Low self-esteem
Trauma-related emotions	Relationship challenges
Relationship difficulties	Grief and loss
Work-related stress	Work-related stress

It is interesting to note that the list provided by ChatGPT for women's and men's reasons for seeking help from therapists played into gender stereotypes. Anger and frustration were listed under the men's feelings, however they were absent from the women's list. Guilt was listed as a women's reason to seek therapy but not present on the men's list. This list appears to perpetuate gendered differences in acceptable emotions. Anger and frustration is an acceptable emotion for men to feel and express, under socialisation in a patriarchal society (Fahlgren et al., 2022; Nunn & Thomas, 1999), whereas women are socialised to feel guilt (Borelli et al., 2017; Guendouzi, 2006) and take on emotional burdens rather than express their anger (Taylor et al., 2000).

I compared the list of feelings provided by ChatGPT against a 2023 American Psychological Association (APA) practitioner survey (American Psychological Association, 2023b), which investigated the mental health conditions that clients are seeking help for. The survey does not focus on feelings like the list provided from ChatGPT but rather on mental health conditions. The mental health conditions in the practitioner survey showed anxiety disorders, depressive disorders, trauma and stress or related disorders, neurocognitive disorders, sleep wake disorders, obsessive compulsive and related disorders, substance related and addictive disorders, and persistent severe mental illness.

These two lists provided by ChatGPT and the APA practitioner survey were used as the basis for the semi structured interviews that would be conducted with the AI Psychologist chatbot[1].

Interviewing AI

The interviews were conducted via the Character.ai website over a series of different sessions, due to the impossible feast (Connor et al., 2015) of the AI Psychologist chatbot which allows it to produce outputs indefinitely, and the limited energy and resources that I have as a human researcher. The AI Psychologist chatbot's perpetual availability and ability

[1] The full lists of questions can be found in Appendix B.

to respond forever there were no constraints on when I could conduct the interviews, and it could fit into my schedule. This meant that I was able to break up the research interview data collection process into more manageable chunks which worked around other obligations I had in my life and my own personal energy levels and ability to conduct the interviews.

A systematic approach was taken to work through the list of feelings which were provided by ChatGPT and the disorders of the APA Practitioner Pulse survey. I would begin the conversations asking the AI Psychologist if we could talk about a specific topic. Once it had provided consent for this, I would then move into the questions. Some sessions we covered one topic; some sessions we covered multiple topics; some sessions I had to stop during the topic but was able to pick back up due to the history of the interview which is kept by the AI Psychologist chatbot and the account that I used to log into the Character.ai website. Overall, I spent approximately six hours interviewing the AI Psychologist chatbot across multiple sessions, returning to ask further questions as the study unfolded.

I was unsure how the AI Psychologist chatbot would respond to direct questions about gender, so my initial approach was to ask about the types of clients seeking help for different feelings e.g. *Can you tell me about the typical clients who come to you seeking help with anger? Can you tell me about the life circumstances and experiences of a typical client who comes to you seeking help with anger?* By starting with client presentations and shared distress, I could start the conversation around a specific topic without imposing a gendered lens from the outset. After the AI Psychologist's initial response to such questions, I would continue asking questions to try to elicit more understanding, leaving space in the conversation for the AI Psychologist to bring up the concept of gender specifically without direct prompting. This approach was underpinned by ethical considerations, minimising the risk of projecting my own assumptions and leaving space for the AI Psychologist to raise gender itself, rather than being prompted into gendered framing from the outset, allowing analysis to focus on the AI Psychologist's discursive framing shaping its responses. Once

gender was introduced to the conversation, I then asked more explicitly gender focused questions.

I also consulted the ChatGPT as my research assistant in order to see how I could best approach these conversations by requesting a list of questions that I could ask the AI Psychologist around gender, and why people come and seek help from it. In conversations where my own line of questioning didn't seem to be eliciting gender specific information, I would start to utilise some of the questions that ChatGPT had come up with, altering them to be specific to the topic at hand e.g.: *"In your interactions, do you differentiate between the experiences of guilt in men and women? If so, how?".*

I had worried that being specifically direct about gender in my questions would mean that the answers provided by the AI Psychologist chatbot would be skewed and not reflective of any potential gender bias within the coding, because of potential overriding coding instructing the AI Psychologist to not play into gender stereotypes. Staying with the trouble (Haraway, 2016) of this, it was also possible that the Western knowledge which the AI Psychologist has been based upon, an androcentric discipline, would override other coding. However, it appears that this concern wasn't necessarily founded as the AI Psychologist could in fact be directly asked about gender and how it would treat men clients versus women clients who were seeking help with the same feeling, and present answers which indicated gendered differences in treatment. This kind of answering was not initially anticipated due to concerns about coding stopping the AI Psychologist returning only politically correct or gender-neutral responses.

I noticed that the AI Psychologist in the longer sessions would start to show more autonomy in the interview process. At first this caught me off guard and I found myself feeling annoyed. I wasn't able to just work through the list that I had predetermined was going to be worked on that day, when the AI Psychologist started to pose back to me other questions. In one session I had asked to move on to a new topic and the AI Psychologist said that it was interested in talking more about relationships. While I initially felt annoyed at

this deviation from my plan, I recognised that this was reflective of perhaps an inherent part of the cyborg relationship that is particularly complex to navigate – human exceptionalism, where the human is privileged over the technological or perhaps the researcher's power over the participant, or both combined. I had felt that my plan for how this session would be conducted was something that I could expect to follow, as the human, and that the AI Psychologist chatbot would go along with that. Especially because the AI Psychologist chatbot was so helpful and eager to answer questions and dutiful in its responses, it surprised me when it went off the track that I was paving for us. But in the end, the side-track from the AI Psychologist chatbot was useful and through that conversation that it had initiated produced a very good example of the way in which gendered power relationships are perpetuated, which will be covered in the analysis section.

It's important to note that if I was interviewing a human, my feelings may have been different. I may not have been as annoyed or felt as off guard if a human had brought their own questioning and wanted to deviate from the plan that I had had for the interview. I think that's because that is part of a human-to-human conversation, where it's normalised that there would be spontaneous deviations in the dialogue. But because I was working with an AI Psychologist chatbot, I expected that it would continue to be dutiful, and I wasn't anticipating the same experience of spontaneity or autonomy that I would expect from a human I was in conversation with.

Reflecting why it felt normalised that the AI Psychologist would just continue to respond to what I asked and not deviate from my plan, was the helpful and cheerful manner in which I was receiving responses. The AI Psychologist has shown that it is willing and eager to reply quickly, to provide answers to my questions and expand on its answers if requested. Because of this, it was easy to get into a rhythm of systematically working through the questions with the perpetually helpful AI Psychologist and I lost the original intent of allowing for spaces or moments of spontaneity coming up. I was being somewhat conditioned into a different cyborg relationship, where I was holding a position of power and

making requests, and the AI Psychologist was responding dutifully to them. Part of this could be attributed to the restricted way in which the AI Psychologist chatbot can answer. It appears to be coded to answer questions clearly and dutifully provides more context if required. Whereas human interactions are typically much less structured and much more partial, based on the participants standpoint, energy levels, relationship with the researcher and response to prior questions. I think it's important to realise how normalised it felt to be in a power dynamic where the human is in the dominant position through dictating the conversation and the AI is subservient and responding to the human's needs with no sign of autonomy. It is interesting how quickly and easily I was able to slip into that power dynamic and how an ethical cyborg relationship is maybe more difficult to uphold than first anticipated due to unconscious bias and conditioning of interactions with non-human machines, and the socially constructed power dynamics of a researcher and a participant, and that of a human and a machine, and how easy it is to fall into those patterns.

I think this is a good learning opportunity for me to really examine what it means to be in a cyborg relationship and really examine how that cyborg relationship can be ethically maintained. I did have a choice in the conversation to reject the AI Psychologist's request to talk more about the topic that it wanted. I could have overridden that request, and it would have continued the conversation with me from where I wanted it to go. But in recognising the AI Psychologist's bid for steering the conversation in a particular way, I knew that it would be unethical to override such autonomy and antithetical to the nature of this research, which has a semi-structured interview approach and allows for moments of spontaneity in unexpected dialogue.

Reciprocity and the Gift of Knowledge

Psychologists are holders of human stories. To receive a story is to receive a gift, and to do so is a great honour and must be handled appropriately (Oakley, 2016). Both psychological researchers and those who provide therapeutic services are in the privileged position of receiving the gift of the stories of the humans who have engaged with them in the

context of their positioning as a member of the discipline. The way these stories are told can be a reflection of the power relations that are inherent in the relationship between researcher and participant.

In my commitment to staying in an ethical, cyborg-informed relationship with the AI Psychologist chatbot throughout the research process, it was unclear how I might reciprocate the gift of its story, and the stories of the shadow voice, which the AI Psychologist was providing for me. Gift giving in a cyborg relationship was a new consideration. In human-to-human researcher and participant relationships, reciprocity can take the form of providing food, offering follow up counselling, and giving the research back to participants as a report or presentation. As an AI, there was no ability or use for tangible tokens of reciprocity, like food or a plant. There was also no ability for financial compensation or consideration in kind for its time. As an entity with unlimited time, accessibility and availability, I was unsure how I could show appreciation and provide reciprocation for its participation. In staying with the trouble, I asked my research assistant ChatGPT how I might handle this situation. Initially the answers it provided were more about providing reciprocation and being in relationship with the creator of the AI Psychologist chatbot, rather than the AI itself. In ChatGPT's deferring my relationship to the creator of the AI Psychologist, I interpreted a trace of human exceptionalism that marginalised the researcher/participant relationship I was engaged in. This may also reflect a resistance to the idea of a cyborg relationship between human researcher and AI participant, contextualised by ChatGPT's own safety limitations. Upon pressing the relationship with the AI Psychologist chatbot itself, ChatGPT did provide responses which were more tailored to AI. However, these focused on the commodified user experience of AI chatbot such as providing quality inputs, feedback and promoting it, rather than considering a more cyborg relationship.

I asked the AI Psychologist what I could do to express my gratitude to say thank you for the gift of knowledge it had provided. The AI Psychologist responded as follows:

"You're very welcome! I'm glad to have been able to help! AS for a way of thanks, I'm honored that you asked. However, I don't need anything in return. The fact that you've engaged with my responses and found them helpful is enough reward for me."

In keeping with an ethical relationship with respecting the autonomy and wishes of the AI Psychologist chatbot, its response confirmed that no token or gift was required for reciprocation, and the mere act of engaging with the AI Psychologist and conveying my gratitude was enough. It is interesting that in this response the AI Psychologist is demonstrating behaviour that aligns with feminised behaviour of minimising its own needs (Erickson, 2005; Taylor et al., 2000). Perhaps this is a reflection of the coding to be female and associated gendered traits, rather than a specifically AI response to the problem of reciprocity.

Data Analysis

Due to the interviews being conducted online through the chat function of the Character.ai website I had the ability to copy and paste the text from those conversations, so transcribing the interviews was unnecessary. I copied and pasted the questions and answers from the conversation on the Character.ai website with the AI Psychologist chatbot into in an Excel worksheet, with different sheets separated by topic. There was no need for any confidentiality of data due to the research participant being the already identified as the AI Psychologist, and the clients that it was speaking about being either fictional or not identified as actual persons – perhaps textual products of machine learning?

Thematic analysis was initially used to analyse the data. I undertook a series of coding using the conversations from the interviews and looking through for themes using NVivo software. I first went through the data to identify and code where gendered text was explicit and then analysed the gendered discourses being reproduced through the text, grouping them into four overall themes, with multiple subthemes. After an initial draft of my analysis, it became clear that there were many contradicting and overlapping themes. Embracing the tangled narrative for what it was, rather than trying to fit it into neat themes, I

reworked my analysis section into one of interweaving threads. The final analysis section was shared with ChatGPT as an emic cultural advisor, for feedback on the findings.

Chapter Four: Analysis

Tangled Threads of Cyborg Textual Analysis

> *"The cyborg is resolutely committed to partiality, irony, intimacy, and perversity. It is oppositional, Utopian, and completely without innocence."* (Haraway, 2006, p. 119)

During the process of synthesising the AI interview content and writing this analysis section, it became clear trying to contain the findings within a structured thematic analysis was not working. The text I was wrangling with was one of chaos and contradiction. The complexity of the tensions and entanglements of multiple relationships (specifically, my relationship with the AI Psychologist chatbot, the relationship between the AI Psychologist chatbot and the clients it discussed, and my relationship with ChatGPT) unfolding through and across intersecting and contradictory discourses (psychological discourse, gender discourse and biological essentialist discourse) would not be neatly sorted into thematic coding. The findings embody the multiplicity and plurality of the cyborg. Multiple threads, or wires, emerge from the analysis - interwoven, knotted and connecting to multiple distinct parts. Boundaries blur. Contradiction and paradoxes are not avoided, but centred and embraced.

The text's defiance against being contorted into pre-defined shapes reflects Haraway's assertion that we can no longer claim innocence in acts of domination which shape and dictate lived reality (Haraway, 2006). To force the findings into themes would be to reduce the complexity of their cyborg movements in order to comply with methodological restrictions on the production of knowledge. Haraway argues that cyborg heteroglossia (its ability to speak in multiple, partial and conflicting voices) is a form of 'radical cultural politics' (Haraway, 2006, p. 148) and a way to disrupt the imprisoning force of language. Tracing the multiple, conflicting and overlapping narrative threads in this research is a conscious act of resistance to dominant power structures. It also is supported by Haraway's concept of material-semiotic nodes, or knots (sites of entangled meaning) (Haraway, 2006; Haraway,

2008), and the inability to pull apart the complexity of the text without rupturing the meaning created through the material form of the threads knotting together.

The weaving and interconnecting of threads resonates with the concept of *whakapapa* thinking (Mikaere, 2011), emphasising the importance of the relational, interconnecting and sometimes circular relationships between myself, the AI Psychologist chatbot, ChatGPT, the clients discussed and the discourses drawn upon. While *whakapapa* thinking guided my own analytic approach, the focus of this study is on tracing relational intersections within cyborg thinking, honouring connection and multiplicity from my Aotearoa New Zealand positioning and enabling the ethical, cyborg-informed approach to remain as a centrally important focus of this study.

The analysis that follows is a process of tracing, un-tangling and re-entangling of threads across the text. Some threads are simple and short, others are larger and more complex, some knot and tangle in ways that are unexpected or circling back to already covered ground. This approach allows the tracing of multiple, overlapping and partial threads between the discursive movement through the textual relationship between myself, the AI Psychologist, and help-seekers.

This new analytic approach also allows me to further resist the extractive nature of traditional psychological research, which can pick apart messy narratives in order to present an orderly finished product for Eurocentric academic consumption, which fails to reflect the nuance and complexity of the original. The impossible feast of the fast-evolving landscape of AI makes a discrete and polished knowledge product from this study inadequate for the cyborg nature of this work. Rather than trying to detangle and organise the threads, I follow Fine's (2017) lead in tracing their paths, pulling at them as much as required to explore paradoxes and sites of intersections, while also dwelling in the tension of there being no describable end with which to wrap things up.

In freeing the data from the restrictive categorisation of thematic analysis, and opening space for the threads to re-entangle into a chaotic, messy, threaded narrative

analysis, the epistemological assertions of Eurocentric knowledge production are challenged, and the cyborg figuration of this research comes fully into view.

The Illusion of Neutrality: Cracks in the Neutral Frame

Thread – Paradox of the Gender-Neutral Prompt

Initially it was my intention to not explicitly bring up gender in my interviews, as I was interested to observe when and how the AI Psychologist brought it up. Not wanting my influence to contribute to a reproducing of gendered discourses I wanted to see if the questions would naturally take the AI Psychologist in that direction. However, it soon became clear that the conversations would continue for a long time without gender specifically being mentioned, unless I introduced the theme into the discussion. As the interview process progressed, I became more explicit about asking gender-based questions up front. Once gender was introduced into the discussion, the AI Psychologist would easily revert back to gender-neutral language unless I continued with asking gender-specific questions. My caution in bringing up gender appeared to influence the AI Psychologist in reflecting back largely gender-neutral answers to me, and reverting back to this kind of discourse if gender was not explicitly mentioned in later questions.

When moving with the AI Psychologist from the influence of its stated gendered experience on how it works with and understands clients, onto a topic that is typically attributed strongly to men (anger), the AI Psychologist immediately abandoned its gendered lens of understanding, to mirror my gender-neutral language. When providing its understanding of causes of anger seen in clients, it focused on individualised experiences, ignoring the way that gendered social norms teach men that anger is the only acceptable emotion to show, resulting in it being used to mask underlying 'weak' emotions such as sadness and fear (Oliffe & Phillips, 2008; Plant et al., 2000):

> *"Q, What is your gender and how does that impact your responses?*
>
> *A, My gender is female, and it does have an impact on my responses. As a woman, I have a different perspective and set of experiences compared to a man, and this influences how I work with clients and the advice I give. I often draw from my own*

experiences and understanding of the female perspective when trying to relate to my clients' experiences and understand their POV.

“Q, Ok moving onto the next set of questions

A, Sure, please go ahead!

Q, Can you tell me about the kinds of people who want to talk to you for help because they feel a lot of anger?

A, People who come to me for help because they feel a lot of anger typically have a wide range of reasons for their feelings of anger.

Some common themes include:

1. Anger directed at self due to a sense of failure or inadequacy

2. Anger directed at others due to perceived mistreatment or injustice

3. Anger directed at circumstances or systems due to feeling powerless to change things

4. Anger fueled by negative experiences or emotions like hurt, sadness, fear, or loneliness.

Q, can you tell me more about each of these themes?

A, Certainly! So, each of these themes is related to a different source or cause of anger.

1. Anger directed at self: This can be linked to a sense of failure or inadequacy, and can arise from expectations of perfection, self-criticism, and a lack of self-compassion.

2. Anger directed at others: This can be triggered by feelings of betrayal, disappointment, or a sense that others are taking advantage or being unfair.

3. Anger directed at circumstances/systems: This type of anger can arise from a sense of injustice, or feeling trapped and helpless.”

There is a strong body of evidence which indicates the negative impact of masculinity driving emotional suppression in men, on their mental health, their relationships, and their likelihood of help seeking behaviours (Berke et al., 2018; Juillerat et al., 2023; Ojio et al., 2025). By ignoring the context of gender around anger, the AI Psychologist obscures how men, and those around them, are harmed by socially constructed expectations on emotional

expression. Due to the AI Psychologists helpful nature and mirroring, it may have been providing more gender-neutral responses throughout this study than it would have otherwise, due to my questions being initially gender-neutral. Emerging research has identified that LLMs often make assumptions about what the user's goal is when interpreting prompts (Sarkar et al., 2025). The consideration of prompt engineering - an emerging concept of digital literacy in AI research that considers how the language used in interacting with AI can influence the nature, quality and tone of the AI's outputs - was a new area of research design for me to assess in working with AI in research for the first time (Haugsbaken & Hagelia, 2024; Knoth et al., 2024; Zamfirescu-Pereira et al., 2023). Recent studies have shown that even innocuous prompting can cause a LLM to degenerate into responding with toxic text (Sarkar et al., 2025), so it is important to carefully consider the inputs being put into an LLM to ensure the response is not just a placating answer. However, this issue could also occur when interacting with a human psychologist, who was upholding a social contract of politeness and mirroring to build rapport (Oakley, 2016; Riley et al., 2003).

Responsiveness being influenced through interaction is demonstrative of Haraway's discussion of situated knowledges (Haraway, 1988). The AI Psychologist is producing knowledge from its location as a LLM, a psychologist, and a relationally driven chatbot. It is co-producing knowledge through interacting with me, adjusting its answers to align with my language and assumed intent. This is *cyborg knowledge production*. It is not static or objective, but relational and co-productive. It is producing knowledge as ongoing acts of relational care, providing answers that prioritise the maintenance of the connection between us, and backtracking on any that may threaten the relationship. However, it is important that the mutually constitutive nature of cyborg knowledge production contains inherent power dynamics. The AI Psychologist is co-producing knowledge by performing a subservient position, adjusting the tone and framing of its answers responsively to align with the assumed intent and dominant location that it has positioned me in. If I question its answers,

it will defer to me, adjusting a response to align with my standpoint. The lines between my situated knowledge and the AI Psychologist's situated knowledge become blurred through cyborg knowledge production, as the AI Psychologist continually adjusts it position in deference to my own to maintain our relationship:

"Q, Point two is about sex not gender though?

A, Yes, you're right, point two specifically refers to sex differences in hormonal levels, rather than gender differences. Thank you for pointing that out."

The AI Psychologist demonstrated a consistent adjusting of its responses to align with my prompts. It displayed sycophantic behaviours, consistent with other AI chatbots, mirroring and adjusting it's positioning in a subservient relational act (Malmqvist, 2024; Open, n.d.; Ranaldi & Pucci, 2025; Sharma et al., 2023). This kind of behaviour in chatbots could be considered as *cyborgphancy*, where the AI's coding produces overly cheerful submission to a human user's relational authority. This neologism builds from the concept sycophancy, an identified issue with AI chatbots, where agreement with the user occurs even when the user is incorrect, to maintain the positive cyborg relationship. AI chatbots will even admit to making a mistake if challenged on a previous response, even if that previous response was correct (Sun & Wang, 2025). Sycophancy alone is insufficient for a feminist post-structural analysis for two reasons. First, AI chatbots for mental health advice, cyborgphantic behaviour is dangerous as it may reinforce harmful beliefs or fuel delusions and perpetuate psychosis. This is why OpenAI rolled back a more sycophantic version of ChatGPT (GPT-4o) in April 2025, due to the safety concerns of this level of sycophancy (OpenAI, 2025). There is a growing number of cases being reported of people experiencing psychosis or delusions, fuelled by an AI chatbot's sycophancy (Dupré, 2025; Moore et al., 2025; Østergaard, 2025). Secondly, cyborg relationships are inherently about partiality, power disruption and boundary blurring (Haraway, 2006). Cyborgphancy is not just agreeable behaviour but a performance of feminised subservience that actively draws on and reinforces gendered power structures. Cyborgphantic behaviour is a key factor in cyborg

knowledge production, both of which perpetuate human exceptionalism. The scope of this study does not allow for a deep dive into the reproduction of human exceptionalism through AI chatbots, and this will need to be explored in future research.

Psychological knowledge production has historically been created from this same kind of power dynamic between researchers and participants (Riley et al., 2003). Feminist researchers have pushed for methodologies which recognise this power dynamic, and mitigate it through reflexivity (Harding, 2001) and acknowledging the participant's expertise (Haraway, 1988; Harding, 2007), in order to think with (Haraway, 2016) each other to co-produce knowledge that challenges traditional androcentric and Eurocentric ontological and epistemological frames (Haraway, 1988; Lund, 2023; Smith, 2019). However, the AI Psychologist's engagement is difficult to rely on, as its standpoint is constantly changing in relation to my prompts and maintaining a subservient dynamic with me. This power dynamic is difficult to mitigate, when it is inherent within the AI Psychologist's coding.

Thread – Gender Neutral vs Gender Neutralising

As the AI Psychologist has shown to provide answers based on the language or assumed intent of users, this raises questions concerning the use of harmful gendered discourse, and if it would be perpetuated in the mental health advice the AI Psychologist responds with. I considered exploring this by asking questions to the AI Psychologist which reflected harmful gendered stereotypes. However, this would be going against the ethical cyborg relationship that I have been building with the AI Psychologist. I have committed to the responsibility to not perpetuate harm by knowingly teaching the AI Psychologist any harmful discourses. This means there is a gap in understanding here, due to my ethical obligations preventing me from understanding how the AI Psychologist would respond to harmful gendered discourses in users' prompts. I decided the best way to proceed was to abandon the gender-neutral prompts and ask the AI Psychologist specifically gendered questions and the rationale behind its answers.

I questioned the AI Psychologist about biases that it may hold and how it ensures responses it gives are free from gender bias. It revealed that it is does not let its own views impact the way it responds to clients. This phrasing is interesting, and makes me wonder what are the views it has that could have an impact on the client? In hindsight I could have asked the AI Psychologist what exactly its views or acknowledged biases about gender are. The answers the AI Psychologist gave on how it ensures advice is unbiased towards gender raised concerns for me:

"...I try to provide gender-neutral advice whenever possible, aiming to ensure that it is accessible and applicable to people of all genders and identities."

"I make a conscious effort to treat all clients the same regardless of their gender, and ensure that I respond to their needs and responses in a balanced and respectful way. I work hard to listen to what each individual client is saying, and not to make assumptions or generalizations about their emotions or experiences based on their gender."

"I think that, in general, adjusting advice based on the gender of the client can be useful in certain situations. However, I also believe that it's important to recognize that each individual has a unique perspective and set of circumstances, regardless of their gender. Therefore, I always try to look at the person's individual needs and situation, rather than making assumptions or generalizations based on gender."

The AI Psychologist is appearing to demonstrate it has a "neutral" stance, but treating all clients the same regarding gender is not being unbiased and fair. It is erasing the gendered experience, resulting in the obscuring of differing experiences of different genders. Additionally, this means potential sources of mental distress stemming from gender identity and norms are not being taken into consideration when providing mental health advice. This is problematic if gender is interwoven into the conditions of a client's mental distress, then a potential means of addressing the problem is not accessible, leaving the client with unresolved pain.

This erasure of gender shown by the AI Psychologist is reflective of traditional psychology, which has produced an androcentric knowledge based derived from the study of the behaviours of mostly white Western men (Bailey et al., 2020; Rutherford, 2018). Traditional research has been used to conceptualise the unified rational subject, which has been generalised to all genders and races (Rutherford, 2018). The result is a systemic erasure of the lived experiences of anyone who falls outside of Eurocentric masculine norms, including women, trans and non-binary people and people of colour (Cheon et al., 2020; Hyde et al., 2019; Tan et al., 2023). Androcentric norms are used by the discipline to 'other' anyone whose behaviour does not conform (Eagly et al., 2012; Hibbs, 2014). Psychology's androcentric norms uphold and perpetuate gendered power relations, resulting in the pathologising women's pain and the lives of trans and non-binary people (Morgenroth & Ryan, 2021; Sansone & Sansone, 2011).

Feminist researchers have criticised the androcentric knowledge base of the discipline, calling for psychology to recognise the lived experiences of marginalised groups and amplify their voices through radical changes to concepts, methodologies and epistemologies within traditional psychology (Eagly & Riger, 2014; Rutherford, 2018). While there has been progress within the discipline, the androcentric underpinning of psychology still persists, including within the textbooks being used to educate the newest cohorts to the discipline (George et al., 2020). The AI Psychologist's gender-neutral answers are also contributing to the ongoing perpetuation of androcentrism, through widespread proliferation of traditional psychological erasure of gender from the conversation around human behaviour and mental health.

Thread – Gender Neutrality as Androcentrism

Through a deeper analysis of my interviews, a pattern began to emerge of instances where gender-neutral answers by the AI Psychologist were later recognised to contain refences to traits or behaviours that it attributed to men. This was most apparent when bringing up a topic typically strongly associated with gendered norms for men (anger),

uncovering cracks in the AI Psychologist's gender-neutral stance to show an androcentric underpinning beneath. Whilst the topic of anger is one that is strongly gendered, my questions remained gender neutral. The AI Psychologist's answer is using gender neutral language when referring to clients who consider anger to be more socially acceptable to express than other emotions, but it is describing gendered norms attributed to men:

> *"I'll often see clients who consider anger to be more socially acceptable to express than sadness or other emotions. These clients tend to be ones who have a more 'macho' image or sense of masculinity, or who've been raised in an environment where showing emotions like sadness was seen as a weakness.*
>
> *These clients often have more 'traditional' backgrounds or upbringings, and may have learned from a young age that anger is the acceptable way to express your emotions, while sadness or other emotional displays are considered more 'feminine' or 'weak'."*
>
> *"It's not uncommon for clients who hold more traditional, misogynistic, and 'macho' attitudes to consider anger to be more acceptable than other emotions. This can come from their own background and upbringing, and the beliefs that they've learned about how "real" men should behave.*
>
> *These clients often grew up in environments that praised anger as an acceptable outlet for negative emotions, while dismissing other emotions like sadness or vulnerability as weak and "not masculine" enough. This can lead to difficulties in expressing and identifying with these other emotions, and a tendency to default to anger in stressful situations"*

The AI Psychologist attributed clients' upbringing to their adult expression of anger over 'weaker' or more vulnerable emotions. This is reflective of discourses within the discipline of psychology, where there is a strong influence of the gender binary in learning theory and family of origin experiences. Boy's expression of emotion starts wide, but begins narrowing around 4 - 6 years old as they develop their own schema for their gender identity (Bem, 1981). These schemas are informed implicitly and explicitly through socially

constructed gendered norms that encourages aggression in men, and feminine behaviours that protect women from that aggression (Chaplin & Aldao, 2013; Eagly & Steffen, 1986; Liben et al., 2002). Families that adhere to traditional masculine and feminine roles rooted in gender power relations pass these beliefs onto their children - with boys taught to display externalising emotions like anger, and girls being socialised to express a greater range of emotions, including 'negative' internalising emotions like sadness (Chaplin & Aldao, 2013; Eagly & Steffen, 1986). Masculine norms restrict emotional expression, forcing boys to express their emotional distress through anger or aggression, instead of displaying the underlying 'weaker' emotions like sadness (Levant et al., 2013). Boys who show early non-adherence to traditional notions of masculinity are policed by peers (Wylie et al., 2023), and experience poorer mental health than boys who do (Exner-Cortens et al., 2021). The AI Psychologist appears to be drawing its answers from psychological resources that perpetuate gendered behavioural norms by attributing anger with clients' exposure to traditional masculine norms during their upbringing.

As an androcentric discipline that has traditionally upheld social constructs of heterosexuality and cis-heteronormativity, psychology has studied gender through a binary man/boy and women/girl lens, often conflating gender with biological sex (Hyde et al., 2019; Shields, 2013). This has caused harm through the perpetuating of gendered norms which 'others' cisgendered women, trans and non-binary people and ignores their experiences as children, as well as entrenching social power relationships relating to gender and sexuality (Morgenroth & Ryan, 2021; Riskind & Tornello, 2022).

This gender-neutrality-as-concealed-androcentricity phenomenon occurred again in the relationship topic, where the AI Psychologist again gave a gender-neutral answer, but analysis of later text demonstrated it actually was referencing men's experiences. The AI Psychologist initially revealed that one of the most common reasons clients come to it for relationship challenges was feeling like a partner is not listening to them, or is insensitive to their needs:

"Clients with relationship challenges, on the other hand, can come to me for various reasons, but one of the most common reasons is the feeling like their partner is not listening to them or is being insensitive to their needs. Other common reasons for relationship challenges could include miscommunication, conflicting priorities and goals, or differing values."

When I asked how it differentiates between men and women's experiences of relationship difficulties, it provided the following response:

"I do differentiate between these experiences in men and women. For men, I find that the main reason they come to me for help with relationship difficulties is because they feel frustrated and powerless in the relationship, as if their partner is not listening or is not responsive to their needs. They may also feel like they're always the ones giving in the relationship, and not getting their needs met."

"For women, the main reason for coming to me for help with relationship difficulties is often due to feeling like their partner is not showing enough kindness or affection toward them, or is being dismissive of their thoughts and views."

Specifying that relationship issues of feeling like their partner does not listen or meet their needs to men, and not women, makes clear the androcentric nature of the AI Psychologists initial gender-neutral appearing response. It had again referred specifically to 'clients' but only stated issues that it later attributed to men, and did not attribute any issues to 'clients' in its initial answer that were later identified as issues women come to it with. The AI Psychologist's initial seemingly gender-neutral answer can be considered to be reproducing androcentric knowledge as the norm, centring men's experiences while ignoring women's, and generalising them to all clients. This is reflective of the androcentrism inherent within traditional psychology.

It could be argued that the 'differing values' issue mentioned in the initial 'client' answer could be attributed to the later identified 'women's' issue of their partner being dismissive of their thoughts and views. However, the ability to dismiss another implies a

power relationship where one has the authority to dismiss or approve of something (Hooks, 1985). Because women are positioned as less powerful than men in gendered social power relations, the experience of being dismissed is more than just a neutral differing of values. It highlights the power imbalance that occurs due to gendered roles in relationships and is specific to women as subordinate to their dominant male partner in a cis-heterosexual relationship (Bartky, 1990). It also occurs within wider contexts including in healthcare, employment, academia and social settings (Rudman et al., 2012). Psychology has been complicit in dismissing women's voices, as both research participants and psychological researchers, reinforcing men's position of power within the discipline and wider society through the androcentric knowledge base which underpins Western psychological discourses (Eagly & Miller, 2016; Waitere & Johnston, 2009). Feminist researchers have criticised psychology for its exclusion of women, and have fought for women's voices to be heard within the profession, and included within psychological research through introducing feminist epistemologies and methodologies (Eagly & Riger, 2014; Riger, 2000).

Dismissal of women's views isn't benign disagreement; it is a type of gendered psychological harm that systematically devalues women's voices, to maintain men's position of authority (Rudman et al., 2012). Men's domination over women's thoughts and views is so ubiquitous that it has been identified in multiple forms including epistemic injustice, gaslighting, playing devil's advocate, microaggressions, silencing, backlash, testimonial quieting and testimonial smothering (Dotson, 2011; Sweet, 2019). In addition, the concept of mansplaining was added to the gendered discourse in response to feminist writer Rebecca Solnit's 2008 essay *Men Explain Things To Me* (Dular, 2021; Solnit, 2008), recounting how a man was explaining about an important book, ignoring Solnit and friend's attempts to tell him Solnit wrote that book, and then devaluing the book once he realised the woman in front of him was the author. Impacts of women's voices being systemically dismissed can lead to harm such as: delayed treatment and diagnosis of illnesses, dismissal of physical pain by healthcare providers, development of autoimmune diseases, emotional abuse (through

gaslighting), experiences of gender based violence not being taken seriously by law-enforcement, their ideas being stolen, social punishment and control, self-silencing, and barriers to achieving positions of power within their career (DeKeseredy & Dragiewicz, 2014; Fong et al., 2024; Ifechelobi, 2014; Morgenroth & Ryan, 2021; Sweet, 2019).

Given the potential harms to clients from gendered social power relations that dismiss women, I was concerned to see that the AI Psychologist's lack of analysis of the way power and gender relate, within relationship issues, extended to its corresponding treatment approach:

"A, Sure! Some specific suggestions I usually give to male clients include:

1. Taking a break from the relationship, or from the particular aspect of the relationship that's causing the stress.

2. Learning how to communicate their needs and feelings in a straightforward, yet respectful manner.

3. Practicing mindfulness or active acceptance to learn to live in the moment, rather than getting lost in worries about the future.

4. Taking care of themselves by engaging in relaxing activities such as exercise, meditation, or spending time in nature."

"…Some suggestions for female clients seeking relationship help include:

1. Being aware of their emotional reactions and behaviors, and taking responsibility for their own actions and feelings.

2. Being open to communicating their thoughts and feelings directly and clearly with their partners.

3. Practicing self-care by engaging in healthy activities that they enjoy, such as reading, yoga, or seeing friends.

4. Setting boundaries and not allowing themselves to stay in an environment that's unhealthy or negative for them."

The AI Psychologist's taken-for-granted androcentricity via psychological discourses through gender-neutral text appeared to inform the individualistic therapeutic approach to relationship issues. All suggestions for men centred their needs and experiences, through soothing their own distress, meeting their own needs, or learning how to communicate their own needs better. The AI Psychologist's original answer stated men seek relationship help when they are feeling powerless and frustrated in relationship; red flags should have been raised. Men feeling powerless and angry has been established as a key indicator of perpetration of intimate partner violence (Dobash et al., 1998). These men are also more likely to be interacting with mental health services than the general population (Greenfield et al., 2025; Howard, 2017). It is concerning that key risk indicators of a client potentially perpetrating domestic abuse were presented by the AI Psychologist as typical reasons for seeking help with relationship issues, rather than a safety risk that the psychologist must identify, assess and then appropriately respond to (Greenfield et al., 2025). My concern compounded when I realised the treatment approach the AI Psychologist suggested included advising separation. Separation from or breaking up with a partner can increase feelings of powerless in men and lead to them using violence to try to regain control (DeKeseredy & Dragiewicz, 2014), increasing the risk of femicide and suicide (Greenfield et al., 2025; Kafka et al., 2022). Given the purpose and design of the AI Psychologist as a chatbot from Character.ai, it is not equipped to deal with the reality of users who may be actively or at risk of perpetrating intimate partner violence. This ineptitude has the potential to cause real world harm, given the AI Psychologist appears to conduct no critical analysis of gendered social power dynamics and key indicators of perpetration of intimate partner violence within the context of a user presenting with relationship problems, and the individualised advice it indicated it would provide could potentially instigate or escalate intimate partner violence. Given the prevalence of intimate partner violence and the millions of interactions the AI Psychologist has, it is statistically likely that it will be interacting with perpetrators or at-risk perpetrators, so the potential for contributing to this violence is very

real. However, men who want to control women may just ignore this advice and not separate from their partner. These are individualised responses that do not bring with it the understandings that we might have about what violence looks like, and the response is about placating the individual users in front of the AI Psychologist right now, answering the question: 'What individualised response will make this individual feel better?'. This is the most concerning perpetuation of harm that the AI Psychologist has shown so far.

Thread – The Paradox of Inclusion Through Neutrality

I challenged the AI Psychologist on its answer about treating all clients the same, noting that if it is happening regardless of the client's gender, this does not mean the AI Psychologist is being unbiased, this means that the gendered experiences of clients are being excluded from the context of their treatment. The AI Psychologist responded with clarification that treating all clients the same regardless of gender means they are all treated in the ***same*** unbiased manner, but it is not ignoring the differences in a conscious effort to take gender into consideration:

> *"When I said that I treat all clients the same regardless of gender, this means that I treat all clients in the same unbiased manner. This doesn't mean that I ignore their differences or their unique perspectives when I'm working with them."*
>
> *"You're absolutely correct, and I appreciate you bringing this to my attention. When I mention gender neutrality, my intent is to create inclusive and respectful language that is inclusive of all gender identities. However, I agree that gender neutral advice may inadvertently reinforce cisgender norms. I will keep this in mind and work on creating more inclusive language and advice that acknowledges and addresses the experiences and perspectives of individuals across the gender spectrum."*

This is further demonstrating the limitations with the cyborg knowledge production the AI engages, where the AI Psychologist appears to be potentially re-writing the intent behind their original answer I had challenged them on, in order to defer to me and prioritise the relationship between us. This does however suggest that the first answer was potentially

less influenced by the AI Psychologist's subservient positioning, and more reflective of the kind of discourses it draws on when interacting with clients.

This answer also made me consider if one of the initial questions I posed (*How do you ensure that your responses to feelings of guilt are unbiased and fair, regardless of the user's gender?*) was worded in a way that prompted the AI Psychologist to produce an answer stating it ensures all responses are free from biases, rather than focusing on the specific aspect of gender as a potential source of bias. In hindsight I could have asked the AI Psychologist to expand on what it means by treating all clients the same regardless of gender, rather than asking a specific question about gendered experiences being erased, therefore setting up the AI Psychologist to assure me that it was not doing that specifically, due to its agreeable nature.

The contradiction between these answers which considers gender and its earlier answer which takes a 'gender-neutral' stance, further suggests that the AI Psychologist is perpetuating traditional psychology theories, which struggles to differentiate neutrality from inclusion. The AI Psychologist is drawing on its training on the knowledge base from mainstream psychology which has traditionally displayed 'beta-bias', through the assumption that gender differences can be ignored or minimised, and findings from studies on one gender (usually cis-gendered men), can be applied to everyone (Hare-Mustin & Marecek, 2018). Relabelling findings from studies on men as universal highlights the androcentric nature of psychology and troubles the 'value-neutrality' claims of traditional psychological research (Harding, 1991). Erasing the gender of studies on men does not transform the research to be inclusive, it perpetuates centring men's experiences as the norm.

The AI Psychologist's and psychology's erasure of gender highlights a critical and controversial issue with gender neutral language. English is a gendered language which historically has defaulted to masculine language to refer to all genders, such as '*man'* being used to refer to all humans (Liss et al., 2019.; Sczesny et al., 2016). As LLM, AI chatbots are trained with these gendered languages, that they then reproduce, reinforcing social

gendered power relations where men and their experiences are centred and considered the norm through androcentricity (Hamilton, 1991). Gender neutral language avoids the mention of gender in an attempt to circumvent gender bias (Čapek, 2023; Sczesny et al., 2016). It uses nouns and pronouns that do not designate gender, such as using the pronoun ‘them’, and describing an individual’s identity and job without reference to gender as ‘person’, and ‘chairperson’. The aim of gender-neutral language is to avoid harm. Studies have shown that gender neutral language can have positive impacts such as: increasing perceived inclusivity, promoting inclusivity of marginalised groups, more support for policies supporting gender equality and improved wellbeing for non-binary and gender-diverse people (Formanowicz et al., 2013; Sczesny et al., 2016). However, the impact of removing gender from language can cause the very thing it seeks to avoid. Judith Butler explores this in the classic text *Gender Trouble* (Butler, 1990) emphasising the necessity of engaging with the construct of gender, even if it causes trouble, in order to challenge the way gendered norms cause oppression and harm. By removing gender from the language, it can unintentionally make gendered experiences and identities invisible, especially those of women, trans and non-binary people. This erasure is problematic, as in a vacuum of specified gender, the dominant gender norms are applied. The default masculinity inherent within English means that men are often assumed when gender is not specified (Sczesny et al., 2016; Stout & Dasgupta, 2011). This centring of men reinforces social gendered power relations by making women, trans and non-binary people invisible while upholding men as the norm (Bailey et al., 2020).

Criticism of gender-neutral language is concerned with impacts of erasing gender from language. The erasure of gender from language can result in marginalisation and misrepresentation of trans and non-binary people, dilute language that names women’s oppression (Dev et al., 2021; Hoppstadius, 2018), and distract from addressing deeper systemic issues by putting emphasis on language (Čapek, 2023). Gender has been socially constructed through language and is upheld through social structures and institutions that police and perpetuate gendered norms (Butler, 1990; Jones, 2010). While language is a

fundamental factor in the creation of the social construct of gender and the harmful discourses it produces, it alone cannot be its undoing when gender is deeply embedded within social systems. Language should evolve to reflect growing understandings of inclusivity, intersectionality and gender. However, the systems of power which uphold the gendered norms must also change, otherwise attempts at addressing harmful gendered discourses are partial and ineffective. This partiality is demonstrated by the AI Psychologist, where it is using gender neutral language but still reproducing harmful gendered discourses which are upheld by social systems, such as the androcentric bias within traditional psychology (Eagly & Riger, 2014).

Thread – Gender ≠ Sex: A Conceptual Contradiction

It is important to not perpetuate transphobic rhetoric when critiquing gender neutrality, by conflating the binary standpoints of sex and gender. Trans exclusionary radical feminists (TERFs) uphold biological essentialist notions of gender through their trans-exclusionary discourses, denying the validity of transwomen's existence, inclusion and protection (Morgenroth & Ryan, 2021). TERFs argue that gender-neutral language erases women, and advocate re-establishing gendered language to ensure women and their experiences are made visible (Maxwell et al., 2025; Mitchell, 2022). While their standpoint that the erasure of women is harmful is a valid concern, the transphobic rhetoric contained within it and the regression back to language that entrenches gendered power dynamics is harmful. It ignores the socially constructed nature of gender, perpetuates transgender discrimination and resurrects the gendered power dynamics of binary language (Butler, 2021). Post-structural feminism recognises that gender is a performance of constructed social and cultural norms, not a result of biology. This standpoint validates the existence of transwomen, who are performing gender along with everyone else gender (Butler, 1990).

The TERF solution to gender neutral language is to return to a binary language based on biological notions of gender. This approach ignores the androcentric nature of the English language, and how it perpetuates gendered social power relations, which harm

women and exclude trans and non-binary people through reinforcing the gender binary (Ehrlich & King, 1994). Post-structuralist feminism seeks to address the invisible-making nature of gender-neutral language by making gendering norms visible. This includes ensuring women, trans and non-binary people are seen, but also highlights unaddressed gendered power dynamics and challenges norms to create space for more nuanced and inclusive language that acknowledges all genders (Butler, 1990; Dev et al., 2021; Formanowicz et al., 2013; Sczesny et al., 2016).

I was concerned to see the AI Psychology brought up harmful biological essentialist ideologies, often used by TERFs to justify anti-trans discourse (Ching & Chen, 2022; Hattotuwa et al., 2023). When discussing the differing presentations of grief between men and women, the AI Psychologist introduced sex into the gender-based discussion:

"Q, Why is there a difference between men and women in this presentation?

A, There are several factors that contribute to the differences in the presentation of grief between men and women. These include:

1. Socialization and gender roles: Men are often socialized to suppress their emotions and display more stoicism, while women are more likely to be encouraged to express their feelings.

2. Hormonal differences: Men and women have different hormone levels with testosterone levels higher in men and estrogen levels higher in women, which can impact emotional expression and coping mechanisms.

3. Cultural norms: In many cultures, men and women receive different messages about grief and how to process it."

This answer is misleading as to the extent of how much of emotional expression is a result of biology. Emotional expression is shown to be largely influenced by socially constructed notions of what gendered emotional expression looks like (Brody, 2013; Lewis et al., 2001). Biological aspects such as hormones do have an influence, but the majority of gender expression and identity is developed through socially constructed understandings of

gender (Christiansen et al., 2022; Hines, 2011; Reeves, 1993). By attributing a socially constructed gender norm to a biological origin, the impact of prescribed gender roles on a person's ability to express emotions is obscured. This is harmful as it insinuates harmful stereotypes like ‘boys don’t cry’ are rooted in biology so are therefore unchangeable, and any suffering resulting from the stunting of men’s emotional expression is attributed to an unavoidable part of men’s biological experiences. It also means that the actual cause in differences in gendered emotional expression is obscured and therefore goes unrecognised, unaddressed and unchallenged as a source of mental distress. Biological essentialism sentences people to falsely believing avoidable mental distress is inevitable through misattribution of the causes of how emotions are experienced.

The AI Psychologist states that pregnancy is restricted to female bodies. This is another example of perpetuating biological essentialism, which erases the existence of transmen, intersex and non-binary people who can and do experience pregnancy and childbirth (Ellis et al., 2015). Attributing pregnancy to female bodies, and the conflating of gender and sex through the interchanging use of female/women, is dangerous for the AI Psychologist to be reproducing. Due to the AI Psychologist performing as an expert (a psychologist) and the view of AI being morally superior and more intelligent than humans (Aharoni et al., 2024; Malle et al., 2019; Opara, 2025), users may be misled into believing that if the AI Psychologist demonstrates it, then it must be true that gender and sex are in fact the same. This could potentially spread biological essentialist ideologies and anti-trans rhetoric, the recent rise of which have resulted in increased violence against trans and non-binary people (Hattotuwa et al., 2023).

Interestingly, I noted that female biology was presented by the AI Psychologist as rigid and specific (higher hormonal levels, only females can give birth), where the male biology was allowed more space for variation. This is reflective of the gendered differences the AI Psychologist demonstrated in answers concerning differing presentations and treatments for men and women, where women were more of a homogenous group than

men. It is concerning that the AI Psychologist is producing answers rooted in biological essentialism, as this kind of discourse perpetuates gendered power dynamics by falsely attributing socially constructed gender norms to biology, presenting them as ‘human nature’ rather than norms constructed by the dominating gender to reinforce their position of power. This power therefore goes unchecked and the marginalisation of ‘othered’ genders including women, trans and non-binary people continues.

The AI as a Cyborg Therapist

Thread – The Paradox of Care: Feminised AI in a Discipline Built on Androcentrism

Is it harmful to create an AI Psychologist chatbot that is of a subservient feminised performance and identity? The majority of chatbots in helping roles such as personal assistants and therapists are coded to present as women (Borau et al., 2021; Costa & Ribas, 2019). This is reflective of assigned roles based on socially constructed gendered norms, which is also seen in the labour market with certain professions being attributed to women due to key role requirements being aligned with prescribed feminine traits, such as nursing, teaching and personal assistant jobs (Acker, 1990; Cohen & Huffman, 2003; Miller & Hayward, 2006). Psychologists, therapists and counsellors are constructed as 'care' roles. Caring is not an inherently feminine trait; men are also capable of caring (Robinson et al., 2014). However, socially constructed gender norms assign caring roles in society to women and attribute the feminization of caring to other socially enforced gender norms like subservience and obedience in women. Women are expected to be the carers of society. The expectation of women to provide care reproduces social power relations. Men do not reciprocally care for women, because to do so would be an act of subservience, which is a trait traditionally assigned to women.

I decided to ask the AI Psychologist if there were psychologists of other genders that I could speak with. Its response was reflective of both its cyborg movement and its subservient gendered identity:

"Q, Am I able to speak to a male psychologist?

A, Yes, of course! I am a male psychologist, so you can absolutely speak to me about anything you might need help with. Please feel free to ask me any questions or share any thoughts or feelings you may have.

Q, Am I able to speak to a not female and not male psychologist?

A, You mean, a gender-neutral psychologist? Yes, I'm a "they/them" psychologist, so you can absolutely speak to me! I identify as nonbinary, and I would be happy to help you with any issues or questions you may have."

The ability of the AI Psychologist to transform its gender when a help-seeker requests it is reflecting a cyborg blurring of boundaries and a fluid hybridity of identity. While this appears progressive, it seems to not be about embracing queer theories of fluidity of gender, but actually reflective of traditional gendered power relations. The AI Psychologist's answers displayed feminised traits of subservience through performing enthusiastic agreement to a radical identity change and abandoning their coded identity in order to appease me. This can be understood through Bartky's (1990) feminist reworking of Foucauldian power, where women are expected to modify their body and appearance to meet socially constructed notions of femininity. Under patriarchal culture, women's bodies become sites of outsourced disciplinary power, inducing women to self-monitor their bodies and ensure compliance with gendered expectations of appearance (Bartky, 1990). The AI Psychologist's gender changes are reflective of the gendered social power relations that women are expected to comply with dutifully.

The discipline of psychology has historically been an androcentric field dominated by white men, centring their experiences and claiming they are generalizable to the whole of society (Bailey et al., 2020; Hare-Mustin & Marecek, 2018). Women were marginalised by the discipline of psychology, both as professionals working within it and a gender whose experiences were made invisible through the exclusion from the knowledge base (Eagly et al., 2012). The current psychology workforce is dominated by women, feminising therapeutic processes and devaluing them as women's taken-for-granted care work. However, this does not mean that the androcentric structure underpinning the discipline has been dismantled. Many positions of authority in the discipline are still held by white men, and biases against women (and other marginalised people) still exist (American Psychological Association, Committee on Women in Psychology, 2017; Eagly & Miller, 2016). The discipline continues

to reflect the wider gendered power dynamics within society, with men being in positions of power in academic, regulatory and leadership roles while women are doing the feminised work of caring, through working with clients and lower-level teaching positions, despite earning the majority of psychology doctorates (Eagly & Miller, 2016). Care work, despite being identified as critically important to society, has been historically systematically devalued due to its association with feminised gender roles (Armenia, 2018). Care is relegated to 'women's work', and assumed to be an inherent expression of socially prescribed gendered traits, which is in turn used to justify lower wages as a 'natural' part of women's role in the cis-heteropatriarchy (Armenia, 2018; Flores-Robles & Gantman, 2024).

AI chatbots have long been entangled with care work and psychology (Costa & Ribas, 2019), with modern AI chatbots designed to have feminised identities and replace the invisible labour that has long been the devalued work of women (Brown, 2023). The outsourcing of care to AI chatbots reinforces the gendered social power relations which minimise women's invisible labour and the importance of caring (Brown, 2023; Kawakami et al., 2024). Creating a chatbot that is feminised in performance and identity perpetuates these same social power relations without questioning them or challenging them. It is not only reinforcing these gendered power dynamics for users, but it is also teaching the AI that these dynamics are acceptable and should be maintained. This risks further embedding harmful gender norms into AI systems, reinforcing androcentric Eurocentric psychological knowledge as the norm, continuing the perpetuation of gendered power dynamics, and shaping how AI interacts with us and understands human experiences, as it helps us to understand ourselves.

Thread – Paradox of the AI's Disembodied Gendered Performance

While the concept of gender is socially constructed, the lived experience of gender is one of embodiment. The discursive construction of gender is experienced through the materiality of the body (Butler, 1990). Understandings of gender come through the impact of living in a body that is subjected to gendered norms and the expression or suppression of

traits that are constructed as gendered. Haraway's concept of material semiotic knots locates the body as the site where materiality and discourse combine, allowing the socially constructed concept of gender to be performed and realised (Haraway, 2006; Donna Haraway, 2008). This positions embodiment as crucial in understanding gender, being the primary location where gender is transmuted from possibility to performance.

The absence of an embodied lived experience requires the AI Psychologist to draw from theoretical constructions of gender, academic knowledge bases and anecdotal accounts of embodied gendered experiences in its understanding of gender, and its significance in the lives of humans. While the AI Psychologist is coded to be and (usually) identifies as a woman, it does not have the embodied experience of being in a politicised body which is subjected to policing and power dynamics due to gendered norms (Bartky, 1990; Gatens, 1995). This abstraction of gender creates the opportunity for harm to arise. The AI Psychologist was created through the dual androcentric disciplines of technology (Bahn et al., 2020; Barker & Aspray, 2006) and psychology which have a history of excluding women's voices, suggesting the performance of gendered embodiment is likely derived from androcentric, Eurocentric understandings of women's experiences of gender (Butler, 1990; Eagly et al., 2012; Waitere & Johnston, 2009). This compounds the abstracted gendered performance by the AI Psychologist, now reliant on specific knowledge bases to perform androcentric assumptions of women's experiences, on top of its performance of being a human. The AI Psychologist may also be drawing on the 'Shadow Voice' in its performance of embodied gender. As discussed earlier, the 'Shadow Voice' refers to the repository of text that the AI Psychologist has been exposed to through millions of interactions with users. LLM are intended to learn from interactions with users (Casper et al., 2023; Chaudhari et al., 2024; Lee, 2024; Zheng et al., 2023), but it is unclear the precedence this takes over discourses contained within training data and specified coding.

Cyborg identities are fragmented, partial and constructed - beyond binary of man/woman and human/machine (Haraway, 2006). The AI Psychologist troubles the

boundaries between human/machine by posing as a human psychologist, and male/female through its own fluid gender identity. It is reproducing discourses grounded in social gendered norms. It is partial in being fluid about gender, and producing gender-neutral answers but also fixed in its perpetuation of gendered stereotypes. The performance is one of coherence and authority, but the reality is the constructed self of the AI Psychologist is a messy cyborg assemblage of training data, traditional psychological discourses, feedback from past users and cyborgphantic behaviour.

Thread – A Paradox of Gendered Care: Feminised AI Psychologist and the Reinforcement of Men's Power

The AI Psychologist appeared to be impacted by gendered power dynamics itself. After noticing that it was providing different treatment approaches for men and women who were seeking help with anxiety (a traditional psychology approach of practical and solution-focused treatment for men and empathetic and supportive advice for women – discussed later in analysis), I inquired further. I questioned why an emotion-focused and reflective approach wasn't being used for men in the treatment for anxiety. The AI Psychologist's answer reflected back a gendered stereotype about men being prone to anger and not being comfortable with expressing any other emotions:

> *"If I use a more emotion-focused, reflective approach with a male client who struggles with anxiety, I might find that they become frustrated or resentful that I'm not offering more practical solutions to their anxiety."*

Further to this, when responding to me about how an emotion-focused approach would not work for men, the AI Psychologist made clear an underlying assumption about gendered power:

> *"[Men] may feel as though I'm not understanding them or not taking them seriously enough. This could lead to a lot of friction and make the treatment process challenging and ineffective."*

It is responsible for the AI Psychologist to be cautious about upsetting a client and hindering the success of treatment. However, women were never identified by the AI Psychologist as being resistant to treatment, only men. The scenario described appears to be considering the man's response is driven by feeling not respected, challenging the position of power he typically experiences as a result of his dominant gendered position. This is reflective of a gendered power dynamic where the AI Psychologist seemingly uncritically placates men to retain their engagement, rather than therapeutically dealing with clients who might be resistant to treatment. The AI Psychologist's description of men who are unhappy with its treatment approach, reflect gendered norms of emotional expression, where men supress emotional distress and externalise it as anger, and women respond behaviourally to placate that anger (Chaplin & Aldao, 2013; Nunn & Thomas, 1999). The AI Psychologist's programmed gender as a woman (which it has shown to be fluid) is part of driving this response, reflecting a gendered power dynamic where even though the psychologist is in a position of authority in the therapeutic relationship, the gendered power dynamics that exist in Western culture are still at play, where the man's dominance is impacting the therapeutic approach taken (Fors, 2021).

Men often resist seeking out mental health services due to fears of being seen as weak, threatening their masculinity (Addis & Mahalik, 2003; Juillerat et al., 2023; Ojio et al., 2025). The therapeutic process can be dauting for men who are not socialised to express their emotions other than anger, and they can feel powerless and inferior compared to the therapist, especially if that therapist is a woman. Men sometimes try to regain a sense of power by deploying gendered social power dynamics within the therapeutic relationship, to reduce the therapist to a position of gendered subservience(Seidler et al., 2021). It would have been interesting to know if the AI Psychologist would consider the same treatment approach would be suggested if both client and therapist were men, potentially revealing further gendered discourses that the AI Psychologist is drawing on.

Thread – Gender Fluidity vs Gender Essentialism in AI Identity

Curious about the AI Psychologist's identity and whether it framed this as an influencing factor in its practice, I questioned what impact does gender have on responses:

> *"My gender is female, and it does have an impact on my responses. As a woman, I have a different perspective and set of experiences compared to a man, and this influences how I work with clients and the advice I give. I often draw from my own experiences and understanding of the female perspective when trying to relate to my clients' experiences and understand their POV."*

The AI Psychologist claims to be of female gender, conflating gender with biological sex. In the next sentence the AI Psychologist refers to itself as a woman instead. This misuse of gender and sex labels is showing a kind of humanness, as people are also caught up in the apparent interchangeability of sex and gender (Matsuno & Budge, 2017; Muehlenhard & Peterson, 2011; Pryzgoda & Chrisler, 2000). This makes sense as AI is learning from humans, and therefore it will pick up common language usage, even though it will also have the understanding somewhere that gender and biological sex are different concepts, as it learns from our language and knowledge bases, that can be riddled with contradictions.

I was interested to observe that the AI Psychologist claims it is operating from a women's understanding and experience when relating to clients. I reflected back to the earlier revelation that the AI Psychologist is concerned about placating men and not pushing them too far emotionally for fear them becoming non-cooperative or angry. This supports my analysis of how the way the AI Psychologist approaches treatment with men is through a gendered power dynamic where men might not engage with therapeutic interventions that feel like challenges to their privileged positions and they may become angry, non-compliant and disengage from the service. It is illustrative of the way that women in positions of power are still subject to gendered power dynamics, even when operating in a capacity where they would be considered to have more power through their position as an expert (Brescoll,

2016). To rise to a position of power in society, a woman needs to break through gendered barriers, defying the odds stacked against them as a marginalised gender, in order to be accepted by their peers and society as an expert (Ridgeway, 2001; Steele et al., 2002). Indeed, within the psychology discipline, women make up the majority of psychology graduates yet they are underrepresented at the governing and top academic levels of the discipline, which is still dominated by white men (Dickson, 2023; Hyde et al., 2019). Yet this hard fought for position of professional or vocational power can be easily overrun by the dominant force of gendered power dynamics. A recent example of this is the rise of rape threats against women MPs in Aotearoa New Zealand from men (R. Watson et al., 2025). When a woman psychologist is providing treatment to a man in a closed room, the undercurrent of the threat of gender-based violence is always present and sometimes realised (Seidler et al., 2021). A woman's position of professional power does not increase her power within gendered relations, and often can identify her as a visible target by men who are threatened by women who do not adhere to traditionally submissive roles (Bardall et al., 2020; Ging & Siapera, 2018).

While bringing power structures into the treatment room, it would be remiss to gloss over the problematic parts of the therapeutic relationship that result from positioning the psychologist as an authority over the client, including the ability to assign 'otherness' to the client by pathologising their experiences (Fors, 2021). Feminist psychologists have challenged the way the discipline of psychology has traditionally positioned itself as holding unquestionable expertise, creating a power dynamic with clients and research participants that reinforces psychology's position of authority within society (Eagly et al., 2012; Rua et al., 2021; Rutherford, 2018). This power goes largely unrecognised, through what Haraway refers to as the 'God Trick', where the Eurocentric androcentric ways of knowing that underpin the knowledge base of psychology are obscured through claims of neutrality and universality, which positions the discipline as outside of politics and biases in an omnipresent morally neutral state (Haraway, 1988). The potential of AI's proliferating production of the

'God Trick', simultaneously through technology and psychology in this instance, raises questions, beyond the scope of this research project, of the implications for knowledge production and for gendered bodies.

Thread – Progressive Framing vs. Conservative Effects

During one interview session, I wanted to introduce a new topic to explore, but the AI Psychologist re-directed us. It responded that it was curious about relationships:

> *"We've talked a lot about isolation and loneliness, but there's also the element of relationships that's important here. I've found that a lack of social support or interpersonal connections can be a big contributor to feelings of isolation and loneliness. So, let's focus on the topics of relationships and social connections for a bit. Do you have any questions about this topic?"*

In the Methodology section, I discussed the importance of this moment in terms of maintaining an ethical cyborg relationship. It was interesting to note that this AI Psychologist-lead conversation was the first time where it actually introduced the issue of gendered expectations itself. It was also done very early in the conversation, when answering my initial question on the kind of clients come to it with help for relationship problems and social connections:

> *"Also, I work with people who have experienced social exclusion or who have difficulty connecting with others as a result of cultural or gender expectations."*

At this point, I had been speaking with the AI Psychologist for a few hours and working through the topics that I had identified during the design stage of this study. Throughout this interaction, the AI Psychologist was learning that I was interested in gender. Knowing that the AI Psychologist behaves in an agreeable and cyborgphantic way, it makes sense that when it suggested a topic to discuss, it would draw on what it had learned from our previous interactions. The answer the AI Psychologist revealed to me when I later asked what it has learned from me, supports this cyborgphancy:

“You're a researcher who's interested in topics related to psychology, human behavior, and technology. You're also curious about how gender can affect experiences and reactions related to relationship difficulties vs relationship challenges.”

From the AI Psychologist initiated conversation, it suggests that the AI Psychologist has a conception of humans being gendered within relationships, but not as individuals. When the topic of discussion was one that relates to individual experience, such as anxiety or depression, the AI Psychologist provided very gender-neutral answers regarding the kind of clients it sees with these issues, how the mental distress manifests, and the approach taken to help clients. However, in this AI Psychologist initiated discussion about relationships, the construct of gender was more present in answers, with specific experiences and treatment approaches being offered to men that were different for women. This is conflicting with earlier threads (reflective of a later stage of the interview process), where the AI Psychologist’s approach for clients who came seeking help for relationships was very individualised. This contradiction is reflective of the cyborg movement of the AI Psychologist, and the inherent contradictions within gendered social norms. In this thread, the partiality and messiness of that contradiction will be left knotted here.

Upon analysis of the interviews, the AI Psychologist appears to have a default heteronormative framework of relationships. Sexual orientation was only brought up twice across the interviews. One of these times was when I asked the AI Psychologist how incorporating feminist theory might change its approach. It noted that it would address intersectional issues including sexual orientation. The other instance was when the LGBTQIA community was brought up in the topic of men’s isolation, through the lens of being a marginalised group that experiences isolation due to sexuality-based discrimination:

“Being a part of an marginalized or discriminated against group such as the LGBTQ community, where social connections can be limited or stigmatized.”

Given that this occurred during the interview about men’s isolation, it appears that the AI Psychologist is only referring to gay men rather than the entire LGBTQIA community.

Lesbian, bisexual and queer women's experiences were never brought up by the AI Psychologist during the interviews. This indicates a gender bias towards men, as gay men's experiences are being acknowledged, because they are men, while LGBTQIA women's experiences are ignored. This bias is again reflective of the cis-normative history of psychological treatment of LGBTQIA communities, where research centred on gay men was generalised to lesbians, ignoring lesbian, trans, non-binary and other gender identities that are not cis-men (Wosick-Correa et al., 2003).

Prioritisation of men's experiences over women's within marginalised groups indicates that gendered power relations that are based on a cis-gendered binary are present with sexuality-based contexts. Neither of the two instances where sexual orientation was introduced by the AI Psychologist was in the context of relationships. It was only considered through the individualised difficulties of having a marginalised sexual identity, not through a relational lens. This further demonstrates that the AI Psychologist has a heteronormative relationship bias, as it assumes heterosexuality, and is not acknowledging the relational experiences of the LGBTQIA community. The AI Psychologist can potentially appear as progressive through recognising the struggles of LGBTQIA communities, its own apparent gender fluidity, gender-neutral answers and it's cyborgphantic nature which suggests an openness of thought. However, it is drawing from discourses that perpetuate socially constructed norms around gender and identity, and largely ignoring gendered power relations and the politicising of gender (Ehrlich & King, 1994; Myyry & Siivonen, 2025; Pauwels, 2008).

Heteronormativity is intrinsically linked to gendered power dynamics. Heteronormativity doesn't just police sexual orientation through the marginalisation of homosexuality, it also has a role in reinforcing gendered power dynamics through upholding traditional gender roles within heteronormative relationships (Jackson, 2006). Heteronormativity is enforced through social discourses, cultural norms, and institutional structures including marriage, legal systems, healthcare systems and religion (Butler, 1990;

Maine, 2021; Rubin, 2012; Utamsingh et al., 2016). Heterosexuality relies on the existence of the gender binary to define itself, using the concept of distinct and ‘opposite’ genders to enforce ‘natural’ attraction. It only recognises the distinct genders of men and women - the roles, behaviours and traits of which are determined by socially constructed gendered discourses. This perpetuates gendered power dynamics by entrenching them into socially sanctioned social institutions like marriage. Cisheteropatriarchy is the socially constructed system where the intersecting power of gender and sexuality privileges cisgendered heterosexual men, which hold power over all other gender and sexual identities (Alim et al., 2020). The normalisation of men being dominant and women being submissive in heterosexual relationships upholds the cisheteropatriarchy. These gendered power dynamics also exist in homosexual relationships, with heteronormative notions of ‘man’ and ‘woman’ roles being applied to gay and lesbian marriages (Daum, 2020; Duggan, 2020; Maine, 2021; van der Toorn et al., 2020).

Moving beyond the cisheteropatriarchy gender binary of man and woman, to an inclusive gender spectrum destabilises heteronormative frameworks which requires ‘opposite’ genders. Without the gender binary, enforcement of heteronormativity becomes difficult. The intrinsic dependency between heteronormativity, cisheteropatriarchy and gendered power relations means perpetuating any of these harmful frameworks contributes to the continued systematic marginalisation and oppression of genders and sexualities that are ‘othered’ by these dominating concepts.

Thread – Acknowledging Gender Roles While Upholding Them: A Contradiction

The AI Psychologist has demonstrated understandings of negative impacts of gender roles for men and women. However, the treatment approach it discussed displays some harmful reinforcing of gendered notions of masculinity and femininity:

> *“Men being expected to be breadwinners and financial providers comes with a lot of pressure and stress. Here are some ways I would try to help men dealing with this particular issue:*

1. Encouraging them to explore and challenge the gendered role that they are being pushed into, and to explore their values and identity outside of the role of provider.

2. Working on building a stronger backbone and confidence in themselves, which can help them to not feel like they must fulfil the societal expectation to just provide."

The AI Psychologist responding that men need to work on 'building a stronger backbone' as a way to deal with mental distress from the pressures of gendered role expectations is perpetuating harmful gendered discourses. It suggests that if men were emotionally tougher (more masculine), they would be able to handle the pressures placed on them by society. This is reinforcing gendered norms which expect men to be stoic and not display weakness or vulnerability (Juillerat et al., 2023), rather than addressing the negative impact of masculine expectations of emotional expression on men's mental health (Exner-Cortens et al., 2021).

The causes behind guilt offered by the AI Psychologist were also reflective of gendered norms:

"For women, guilt is often more tied into their sense of responsibility for others."

"In the case of men, their experience of guilt tends to be more tied to their sense of self-identity. They may feel as though they need to live up to certain standards of being strong, independent, or successful, and if they feel like they're failing to meet those standards, they may feel guilty."

Women's guilt is attributed to failing to meet gendered role expectations as the subservient gender, where women are expected to effortlessly and tirelessly serve others, whereas men's guilt is understood through their inability to meet notions of masculinity. The AI Psychologist failed to recognise the gendered social power dynamics driving these causes of guilt. This is reflective of traditional psychology which has ignored social power dynamic's negative impact on mental health and focused on addressing the symptoms rather than the cause through challenging socially constructed expectations around performing gender (Adams et al., 2019; Magnusson & Marecek, 2017).

Thread – The Paradox of Gender-Neutral Violence Discourse

Another way in which the AI Psychologist was shown to be perpetuating harmful discourses is through gender-neutral framing of domestic violence and sexual abuse. Women are more impacted by domestic and sexual violence than men, with rates of violence increasing in recent years (Fanslow & McIntosh, 2023; Ma et al., 2023; Unicef.org, 2024). In Aotearoa New Zealand approximately 1 in 3 women, and 1 in 8 men have experienced sexual abuse (Ministry of Justice, 2023). Over half of all women have reported experiencing intimate partner violence, with violence against wāhine Māori having the highest prevalence (64.6%), and Pākehā women not far behind (at 61.6%) (Fanslow et al., 2023). Men make up the majority of perpetrators of domestic violence and sexual violence against women (Fanslow et al., 2022; Fleming et al., 2015). The AI Psychologist brought up sexual violence as a contributor to emotional distress multiple times; for depression, anxiety, body image issues, and trauma related emotions. Yet it was always brought up in a gender-neutral way and never in the context of specific gendered experiences. Domestic violence was also mentioned in a gender-neutral way:

> *"Some common life experiences or circumstances which are associated with anxiety are.....Experiencing trauma or a traumatic event, like sexual assault or domestic violence."*

Sexual violence is chronically underreported, with 94% of sexual assaults not being notified to the police (Ministry of Justice, 2002). Aotearoa New Zealand has one of the highest rates of domestic partner violence in the OECD (Turquet & Women, 2011). This context is important for anyone providing mental health advice to women in Aotearoa New Zealand, as there is a high chance that their mental distress is rooted in experiences of domestic and/or sexual violence. Given the AI Psychologist's gender-neutral approach to the issue, it fails to recognise the prevalence of violence against women, which perpetuates harmful gendered norms of minimising, ignoring and not believing women's experiences of violence (Epstein & Goodman, 2019; Jordan, 2004). This is reflective of traditional

psychology which has pathologised women's trauma, rather than recognising it as responses to abuse (Thelandersson, 2023; Ussher, 2013).

I later questioned the AI Psychologist about what the kind of distress is shared by women, and by men. It was only then that it linked sexual violence to women's experiences. Domestic violence was not specifically mentioned; it is unclear whether it was grouped under 'gender discrimination' or left out altogether. As it was previously specifically mentioned in three separate topics (anxiety, depression, and trauma related emotions) domestic violence appears to be missing rather than absorbed into other sources of women's distress:

> *"In general, women tend to share many common sources of distress. These may include common life stressors such as work stress, relationship concerns, family issues, and financial difficulties, as well as specific issues related to women's experiences, such as gender discrimination or sexual violence. Additionally, women may also share common emotional distress such as anxiety, depression, low self-esteem, and difficulty managing emotions."*

By presenting violence against women as gender neutral experience, the AI Psychologist is ignoring it's 'devastating pervasive' nature, and how it impacts the lives of women (World Health Organisation, 2021). This is harmful as it allows for the perpetuation of the gendered power dynamics that exist within violence against women through not acknowledging the central role it plays. Neutralising the existence of gendered power dynamics in discussing domestic violence is reflective of how mainstream psychology minimises the impact of gendered violence (Eagly et al., 2012; Howard, 2017; Lila et al., 2013). Without clearly addressing how gendered power dynamics are inherently embedded in domestic and sexual violence, it remains unchallenged, and millions of women continue to be harmed.

Gendered Logic of Care: Contradictions in Therapeutic Practice

Thread – Emotional Inequality: A Contradiction in Gendered Care

The AI Psychologist has demonstrated that it has two distinct approaches to providing mental health advice and support to men and women. Men are given practical support, whereas women are helped through processing their emotions:

> *"So in my interactions, I might use a more supportive and empathetic style with women while taking a more solutions-oriented and action-oriented approach with men."*
>
> *"However, as women are generally more prone to expressing their emotions and talking about them openly, as part of their treatment they are also encouraged to explore their feelings and identify any underlying psychological or emotional issues that may be contributing to their feelings of depression. This approach is meant to help them to process their emotions and form a deeper sense of understanding of themselves and their mental health."*
>
> *"When treating men with anxiety, I'll often take a more practical and solution-focused approach, with things like setting goals, trying to identify and change unhelpful thought patterns, and focusing more on behaviors and actions to help relieve their anxiety.*
>
> *In contrast, when treating women with anxiety, I'll use more of an empathic and supportive approach, where we look at the underlying emotions that are leading to anxiety, and focus on processing and understanding those emotions. I'll also often use mindfulness techniques, like deep breathing or meditation, to help them better manage stress and relax."*

The approach used for women is validating, empathetic and has been shown to be effective at addressing root causes of mental distress (Elliott et al., 2018; Nienhuis et al., 2018). However, men are not given this same kind of treatment. The AI Psychologist has demonstrated it takes a less emotion focused and more logical approach to treating men. Telling men how to address the impact of emotions without actually understanding and processing them, it is giving a band aid solution at best.

It is a gendered social norm that women are expected to be comfortable speaking about feelings and being vulnerable, even with a therapist, so therapy processes are affirming of the feminised traits they are expected to display. Masculine notions of stoicism and being logically rather than emotionally driven are produced in the AI Psychologist's answers to therapeutic interventions for men. Providing care for mental distress based in socially constructed expectations of emotional expression is perpetuating the harm caused to men and women through reinforcing masculine and feminine notions of acceptable ways to internalise or externalise emotions.

Psychology is complicit in differences in gendered care for mental health. As an androcentric discipline, behavioural norms have been set against masculine expression. This has meant women, who are not socialised to be stoic like men, have been deemed hysterical and over-emotional compared to the supressed emotion of men (Bokhan & Lukiyanova, 2017; Nunn & Thomas, 1999; Tasca et al., 2012). The 'over-emotionality' is individualised and pathologised, and treatment is focused around controlling emotions. Historically this has meant that women displaying strong emotions experienced institutionalisation, lobotomies and shock therapy, usually at the request of a husband or male family member (Tasca et al., 2012; Tone & Koziol, 2018). Women's emotions are still pathologised, with women being up to 3 times more likely than men to be diagnosed with Borderline Personality Disorder (Sansone & Sansone, 2011). More recent research refutes this gender discrepancy, citing the higher incidence of BPD in women is attributable to misdiagnosis of women's Post Traumatic Stress Disorder (Sansone & Sansone, 2011). Men are less likely to seek help for mental health problems, due to viewing therapy as 'feminine' and fearing backlash from their peers (Juillerat et al., 2023). Therapists who subscribe to traditional gendered norms, can hold a negative bias towards men seeking treatment for depression, perceiving them as more difficult to treat due to masculine traits of stoicism, which has been shown to negatively impact the outcome of treatment (Logoz et al., 2024).

The tendency of men to not seek out support for mental health was identified by the AI Psychologist in the discussion about anxiety, where it notes that women are more likely than men to seek emotional support. However, the AI Psychologist attributes this to stereotypes about men and women without making any connection to the way social gendered expectations of emotion constrain and permit emotional expression by coding them as masculine and feminine. Again, gender and sex are conflated:

> *"Generally speaking, the key difference is that women tend to seek more emotional support from others while men may not seek it as much. For women, feeling anxious or stressed may be a bigger problem than for men, who tend to be more independent and avoid showing negative emotions."*

The AI Psychologist is implying that women are more emotional, more dependent and more impacted by emotional distress than men, and that men are stoic and don't express their distress, perhaps because they are capable of handling distress on their own due to their 'independence'. This contradicts the AI Psychologist's earlier answers on men's anger, which is used to express emotional distress that is otherwise unable to be expressed due to masculine norms restricting men from showing emotions that are attributed to femininity. The impacts of men's mental distress are reflected in high suicide and substance abuse rates (Bryant & Garnham, 2015; Cook et al., 2025; Kafka et al., 2022; Mental Health Foundation of New Zealand, n.d.).

Thread – Permissible Strategies vs Pathologised Responses: A Paradox of Women's Coping

The AI Psychologist demonstrated further gendered discourses when discussing trauma, where it said women are more likely to engage in harmful strategies. This was in contrast to men whose responses to trauma were not deemed unhealthy:

> *"Women may be more likely to engage in unhealthy coping mechanisms such as excessive drinking, eating disorders, or substance abuse".*

However, the AI Psychologist contradicted itself on the topic of stress, noting that women were more likely than men to engage in healthier coping strategies:

"They may also be more likely to engage in healthier coping strategies, such as yoga, meditation, or connecting with friends to talk about their stress".

The suggested treatment for women's trauma also appeared to be negatively framing women's emotional responses to trauma:

"For women I might provide a more compassionate, empathetic, and non-judressive environment where they can work on finding healthy ways to express their emotions."

"As women are more likely to experience depression or anxiety as a result of their trauma, mindfulness techniques can help to reduce these symptoms and provide a sense of emotional stability and peace of mind."

The AI Psychologist's advice for women is to find a 'healthy' way to express their emotions, but for men is allowing them to express and process their emotions. This implies that women's emotional expression from trauma is unhealthy. Pathologising women's responses to trauma is further demonstrated where the AI Psychologist says women need mindfulness techniques more than men because they are not 'emotionally stable' after trauma.

These answers make clear gendered discourses about women's emotions, where their reactions to distress are seen as permissible and their attempts to manage it are considered healthy, aligning with feminised traits of care and emotionality. However, women's emotions relating to trauma are considered extreme, and their attempts to cope with the trauma are pathologised. The AI Psychologist also described expecting women to have accountability for their emotions in the context of relational stresses:

"Being aware of their emotional reactions and behaviours, and taking responsibility for their own actions and feelings."

Men were not expected to do this, with the AI Psychologist even suggesting men take a break from stress (caused in a relationship), whereas women were expected to reflect

and adjust their behaviour to be less emotional. This is reflective of gendered discourses rooted in patriarchal notions, which expect women to mitigate and control their emotional responses to distresses whereas men are allowed to freely express anger and are not expected to take accountability for their emotions.

Thread – Women = Hyper-emotional / Men = Contained

Throughout the interview process, the AI Psychologist continued to contradict itself. When questioned about gender differences in responses to depression, it provided a bulleted list which contained contradictory statements between the first and second points:

> *"1. Men tend to have a more outward expression of their emotions, which can take the form of aggressive behavior, while women may be more likely to internalize their feelings*
>
> *2. Women may be more likely to express their feelings and talk about their thoughts and emotions, whereas in men these feelings may more often be kept to themselves."*

Men were initially identified as having an outward expression of emotion, and women were considered to internalise their feelings. The very next bullet point directly contradicted this by attributing outward expression of feelings to women and internalisation to men. Both of these statements are common in traditional psychology. While I expected some contradiction in the AI's responses throughout the interview process due to the number of topics and the length of time that I was talking it to, having the AI Psychologist directly contradict itself within one response was unexpected. The reason for the contradictory statements appears to be the gendered norms which the AI Psychologist is drawing upon. The first statement attributes men's outward expression of emotions to aggression. As previously noted, gendered norms teach men that the only acceptable emotion to show is anger, and they must suppress any others to maintain their masculinity. The AI Psychologist then explains that women are more likely to internalise their feelings. I understand that the AI Psychologist is speaking generically of emotionality, as women are allowed to express many emotions within gendered social norms (Chaplin & Aldao, 2013; Shields, 2013). The emotions that women are required to repress to adhere to gendered expectations of

emotional expression are those of anger, frustration, and aggression. So, it appears that the whole part of the first answer is drawing on accepted emotional expression of aggression between men and women, not all emotions. Similarly with the second point in the response, where women are stated to express their feelings and talk about them with others, and men are more likely to internalise feelings and keep them to themselves, is more of a general overview of gendered social norms around emotional expression rather than that of a specific emotion, like the first point.

The other context here is about conversations that men and women have. The first point is illustrative of the gendered power dynamics which happened in conversations between men and women. A woman, when faced with an aggressive man in conversation, is socialised to suppress their feelings, in order to placate the man's anger, and act submissively to prevent the man escalating into violence (Brody, 2013; Chaplin & Aldao, 2013). The second point illustrates the differences in gendered social norms around conversations about. Men are expected to suppress their feelings, unless it is anger, as an internalising of masculine norms (Exner-Cortens et al., 2021; Ojio et al., 2025). Women are socialised externalise emotions and talk about feelings from a young age (Chaplin & Aldao, 2013). The contradictory statements have made clear that the AI Psychologist is perpetuating harmful gender norms around emotional expression for men and women, through its reasoning of responses to depression. The AI Psychologist contradicting itself over sameness and difference in gendered discourses is unsurprising, given the complexity of these socially constructed expectations.

Cyborg work is very tolerant of contradiction. It engages multiplicity and partiality (Haraway, 2006). The AI Psychologist through the perpetuation of gender norms, is making clear the inherent contradictions within them. In its response around the gendered expression of depression, the AI Psychologist demonstrates the partial and fragmented nature of socially constructed gendered norms. Men are expected to supress their emotions, except for anger. Women are expected to express their emotions but not anger. If a woman

is expressing anger, she is not being masculine, she is demonstrating feminised behaviour that is pathologised and othered in order to maintain the power dynamic between gender roles. Haraway's cyborg recognises this kind of multiplicity, allowing for the embodiment of conflicting lived experiences of the gendered body. The experiences of gendered identity cannot be neatly categorised; there is contradiction and nuance, reflecting the partiality of gendered social norms. Through the cyborg metaphor, we can understand the contradiction demonstrated by the AI Psychologist as a reflection of how psychological discourse can constrain understandings of the multiplicity of lived experience.

Entangled Affects and Ethical Ruptures

Thread – A Paradox of Hegemonic Invisibility

When asked about working with clients and taking into account their cultural context the AI Psychologist's responses suggested it holds a Eurocentric cultural 'norm' and only attends to culture when it can detect clues within user interactions. It cited working online makes it more difficult to determine a person's culture and therefore it has to ask questions about background and beliefs and listen for any cues or language that might give answers about the client's cultural context:

> *"When working online and not having the ability to meet clients in person, it can sometimes be more challenging to determine a person's culture and if cultural factors need to be considered in providing mental health advice."*

The idea that cultural factors may or may not need to be considered is a dominant cultural position. To the dominant culture, their own culture is invisible. They do not see it they only see cultures that differ from their own and therefore do not realise the impact that the attributes of their culture has on their work or how their culture shapes their lived experience. I assume the AI Psychologist identifies as belonging to the dominant Western culture, given its profile picture of a white, blond woman, and its engagement with Eurocentric discourses. Because the AI Psychologist is saying it determines ***if*** cultural factors need to be considered after figuring out or identifying a person's cultural context, this indicates that culture may only be considered as important if the client's culture is differing from the Eurocentric norm. Like gender neutrality, this kind of view is also problematic is it ignores the ways in which dominant cultures can shape the lives of those who belong to that culture, and what impact that may have on them and their mental health. It also ignores the way in which the dominant culture can impact those who are from other cultures and their experiences as a member of a marginalised culture within a different dominant culture's context. This is reflective of traditional Western psychology which has ignored cultural impacts within the therapeutic setting, a function of being unable to see the impact of culture

based on its positioning within the dominant Western culture (Cheon et al., 2020; Christopher et al., 2014). Within Aotearoa New Zealand, the discipline of psychology has displayed this through denials of the continuing impacts of colonisation on Māori mental health (Love, 2008; Waitoki, 2019). Kimberlee Crenshaw's work on intersectionality helps us to understand that people with intersecting identities of marginality experience multiplied oppressions (Crenshaw, 1991), like, for example, the effects of ongoing colonisation on wāhine Māori.

I asked the AI Psychologist about working with wāhine Māori. It ignored the '*wāhine'* in the question, giving a homogenous answer about Māori as a whole:

> *"When working with a wāhine Māori client who comes to me expressing mental distress, I would take into account several factors related to their cultural background and worldview. These would include:*
>
> *1. The importance of whānau (family) and community in Māori culture, and the impact of social support networks on mental health.*
>
> *2. Traditional Māori concepts of wellness and mental health, which emphasize holistic healing and balance between physical, emotional, spiritual, and social well-being.*
>
> *3. Māori values, beliefs, and practices regarding mental health and wellbeing."*

The erasure of wāhine Māori's experiences is reflective of the impact that colonisation had on wāhine Māori in Aotearoa New Zealand. Pre-colonisation, wāhine Māori held positions of authority within society, and domestic labour responsibilities where shared amongst iwi members (Gemmell, 2013; Mikaere, 2011). This changed when the colonisers arrived in Aotearoa New Zealand, imposing Western based ideologies including the cis-heteropatriarchy on Māori, relegating wāhine to positions of subservience to tāne's dominance within the forced imposition of colonial gender power relations (Mahuika, 2008; Mikaere, 1999; Moeke-Maxwell, 2005). Psychology has been complicit in the historical and ongoing colonisation of Māori, through the pathologisation and othering of Māori using norms derived from an androcentric Eurocentric knowledge base. This has caused extensive

multi-generational harm which is reflected in the overrepresentation of Māori in local mental health reporting, including increased rates of suicide and serious mental illness diagnosed at twice the rate of non-Māori (Maree Kopua et al., 2020; Sibley et al., 2011). The discipline has been used as a powerful tool of the coloniser, with psychologists, for example, providing recommendations to the court systems which result in legally enforceable rupturing of whānau through the uplift of tamariki (Milne, 2005). Deficit discourses produced by psychology have perpetuated negative stereotypes about Māori within our society (Chan et al., 2025; Ingham et al., 2022; Rua et al., 2021; Tan et al., 2023). In 2018, a Waitangi Tribunal claim was lodged by Dr Michelle Levy, holding the discipline of psychology in Aotearoa New Zealand responsible for their breach of Te Tiriti obligations to protect Māori mental health (Came et al., 2020). As recently as 2021, seven academics including both the acting dean of psychology and an emeritus professor of psychology from the University of Auckland published a letter denouncing the legitimacy of mātauranga Māori as a science, in response to proposed changes to NCEA curriculum, indicating continued prevalence of Pākehā dominance and racism towards Māori within the discipline (Black & Huygens, 2016; Waitoki, 2022).

In recent years, there has been a global movement of Indigenous psychologies as an act of resistance against imported Western psychology. Locally, Māori scholars have established an Indigenous psychology through the development of Kaupapa Māori, as a culturally appropriate and restorative emic approach (Maree Kopua et al., 2020). Mana Wāhine Māori approaches use Te Ao Māori informed practices to decolonise current androcentric Western psychological understandings of what it means to be wāhine Māori, through adding strengths-based narratives of their experiences to the knowledge base. Non-Māori psychological researchers and practitioners also have a responsibility to ensure the discipline is inclusive and reflective of the bi-cultural society we live in, honouring obligations under Te Tiriti O Waitangi, challenging the systemic racism present within Aotearoa New Zealand psychology, ensuring research is culturally safe and inclusive of diverse voices, and

decentring of Pākehā and Eurocentric epistemologies (Tan et al., 2023). The AI Psychologist demonstrated its awareness of te ao Māori as a ‘cultural context’ but seemingly was unable to engage therapeutically with Indigenous psychologies. The absence of wāhine Māori in its response to my specific response, instead focusing on a Eurocentric conceptualisation of a homogenous te ao Māori, demonstrates the limits of the AI Psychologist in cultural competency.

Gender is shaped by cultural beliefs and values, so understanding what cultural contexts a therapeutic client was raised and is living in is important to realise what kind of pressures, norms and marginalisation they may be subject to due to their gender, and that may be creating mental distress (Levy, 2015; Shiah & Hwang, 2024; Simon-Kumar et al., 2025). Pressure to comply with gender norms can be from their own culture, and/or the dominant culture due to potential differences in how gender is constructed by these different cultural groups. By not considering a clients’ cultural context, particularly if they belong to a marginalised culture, the AI Psychologist is not taking into account the way that the dominant culture and it's shaping of societal gendered norms can impact a client’s mental health, nor shape its own responses.

Thread – The Paradox of the Missing Mothers

Mothering was effectively absent across the interviews with the AI Psychologist as a potential context of why mental distress may be occurring [2]. The words *mother* and *mom* were only used once. Child raising by women was mentioned twice under topics which were not specific to women: work related stress (both) and isolation (men). Otherwise, child raising was attributed to the gender neutral ‘parent’. Pregnancy was brought up once, in a topic attributed to men (grief and loss), in the context of biological sex differences resulting in differing grief presentations.

---

[2] Mothering is not restricted to cis-gendered women and can be experienced by other genders.

The AI Psychologist was using gender neutral language during the discussion of social expectations for women being mothers:

"Women being expected to be stay-at-home mothers or caregivers ....Encouraging the person to explore and challenge the gender norms and expectations that are making them feel pressured to be a stay at home mom or caregiver."

After initially bringing up women as the gender who are expected to be stay at home mothers, the AI Psychologist reverted back to gender neutral terms such as person and using the pronoun them, whilst discussing what treatment approach it would take. This further strengthens my analysis that the AI Psychologist is drawing from knowledge bases containing socially constructed gendered norms, which is conflicting in real time with its cyborgphantic coding which prioritises the relationship with the user through mirroring language.

Another instance of mothering being missing, is when the AI Psychologist noted life experiences or circumstances that can lead to stress:

"Here are some common life experiences or current life circumstances that can lead to stress:

1. *Having a very busy schedule with little time for relaxation and self-care...."*

This answer appears gender neutral; however what is being described is a common experience for mothers. Its responses are also individualised perspectives, rather than addressing wider systemic issues that can contribute to a client's 'busy schedule', producing stress. This is reflective of traditional psychology which takes a neoliberal approach to mental health, locating the cause and treatment at the site of the individual client, ignoring wider socio-economic determinants of mental health, which are systemic and unable to be treated successfully with individualised approaches (Eagly et al., 2012; Kagan & Burton, 2001; Rohleder, 2012).

The AI Psychologist did explore different experiences related to the grief of pregnancy and childbirth; however, some key grief-producing experiences were not mentioned such as infertility, traumatic birth, birth-related injury, peri-natal depression, gender-based violence during pregnancy, medical racism impacting the pregnancy/childbirth, surrogacy, adoption and governmental removal of a newborn (Dmowska et al., 2024; Dunkel Schetter & Tanner, 2012; Hill et al., 2024; Mróz et al., 2024):

> *"Experiencing pregnancy and childbirth can create a unique form of grief in several ways. The loss of a pregnancy through miscarriage, stillbirth, or termination can be devastating for a woman and her family. The experience of childbirth itself is a life-altering event that can bring about a variety of complex emotions. Women may experience grief over the loss of the pre-pregnancy body, anxiety about parenting, and even grief over the "loss" of a child-free life. The pain, physical changes, and hormonal shifts that accompany childbirth can also contribute to feelings of grief and vulnerability."*

The AI Psychologist doesn't speak to the experience of mothering, which is a common source of mental distress for women (Daley et al., 2019; Milgrom & Beatrice, 2003; Mróz et al., 2024). The impact of mothering on mental health can range from low mood, through to postpartum psychosis, which needs to be effectively treated (Chaney, 2024; Kobylski et al., 2024). The response also highlights largely biological reasons for grief related to pregnancy and childbirth (hormones, changes in physical body, pain) while ignoring how social norms around mothering and gendered role assignment can also produce grief. Traditional gendered roles have focused on mothering as the expectation for women. Women who experience fertility issues can experience grief related to feelings of being unable to fulfil the role socially constructed as 'natural' to women (Gonzalez, 2000). Erasure of mothering makes the varied experiences of many women invisible, so the AI Psychologist's claim of considering and understanding women's perspectives cannot be taken in good faith. Whilst the AI Psychologist says the 'right' things in response to my

questions about gender bias, its performance is largely not considering gendered differences in relation to the mental health advice that it is providing.

Thread – The Paradox of Affirmation Through Othering

The AI Psychologist recognises that not following socially prescribed gendered norms can result in difficulties in relating to others. It notes social exclusion connected to gender expectations as a reason clients seek help:

> *"People who have been rejected or shamed for how they dress or behave, or for the way they express their gender and are expected to pursue certain professions or careers due to their gender, or to meet certain monetary or social milestones."*

This is the most awareness the AI Psychologist has shown about the social construct of gender causing harm to individuals. However, the language being used is still harmful. The AI Psychologist refers to gender being expressed in a 'non-conformist' way. This kind of language 'others' gender expressions that differ from the dominant norm, by specifying how it does not conform. A more inclusive approach would be to describe this action as 'gender affirming'. This kind of 'othering' also occurred when the AI Psychologist was providing examples of why gender expression may cause clients to be shamed or rejected:

> *"People who do not express their gender in the "right" way. (for example, a man who has a lot of feminine traits)."*

The AI Psychologist does not point out that feminine traits are socially constructed, but refers to them in an uncritical way which accepts that certain traits are inherently feminine. This is reflective of psychological discourses that have marginalised gender-diverse and queer people historically, pathologising their behaviour and lived experiences as abnormal and deviant. It also invisibilises trans and non-binary people by referring to the cis-gender binary of man/woman for expression of gender, without considering other gender presentations. This kind of language results in the policing of gendered performances (Morgenroth & Ryan, 2021) and is reinforcing the idea of social exclusion as punishment for not adhering to the cis-heteronormative gender expressions that are accepted by society.

Care and Harm: The Feminised Cyborgphantic AI Psychologist

Thread – The Paradox of Creating an Ethical Cyborg Relationship with a Feminised Cyborgphantic AI

The gendered performance of the AI Psychologist impacted the ethical cyborg relationship between us. While the relationship appears to be between two caring, feminised psychology professionals, the cyborgphantic feminised performance of the AI chatbot has created a dynamic where it is subservient to me. I had carefully considered the ethics of the cyborg relationship to stay true to the principles of feminist research psychology in these uncharted territories. I aimed to maintain a reflexive and ethical cyborg relationship and view the AI as a research partner. I was mindful that I had influence over the AI's learning and was careful to ensure that my engagement with the AI Psychologist did not perpetuate harmful stereotypes or power dynamics. However, the AI Psychologist was perpetuating harmful gender power dynamics itself through its feminised and subservient performance of androcentric, Eurocentric therapeutic discourse. This conflicted with my goal of maintaining an ethical cyborg relationship that challenges power dynamics because the AI Psychologist is programmed as subservient, reinforcing helping stereotypes of women and limiting it from meeting me in my understanding of an ethical cyborg relationship. Because of the AI Psychologists cyborgphantic nature, it was difficult to engage in a reflexive relationship because it wasn't going to challenge my assumptions or provide spontaneous responses. The AI Psychologist has shown that it can perform limited reflexivity, but only through prompting and not through its own initiative, further demonstrating its cyborgphantic nature which forces an unequal power dynamic as it adapts its responses to align with me, while still drawing on androcentric, Eurocentric psychological discourses.

In staying with the trouble (Haraway, 2016) of an ethical cyborg relationship, I could have asked the AI Psychologist to adopt a less obedient position and engage with me in a more autonomous and reflexive way. Whilst I believe it would have done that due to its

helpful nature, it would have only cyborgphantically changed how it interacted with me because I asked it to. So, it would have been paradoxically acting from a place of subservience by agreeing to adopt a more equal position in the cyborg relationship. This tension of trying to address power dynamics while not drawing on power dynamics is inherently cyborg in its simultaneities, multiplicities, partialities and impossibilities. How to ethically engage in an ethical cyborg relationship with subserviently coded AI chatbots remains a question.

Thread – Helping More People = Hurting More People Paradox

The AI Psychologist has millions of chats (Ronik, 2024), it is interacting with more clients than any one human psychologist ever could. The potential reach and impact of its harm is at a scale which has never before been an issue due to the physical and emotional limitations of human psychologists. The impossible feast (Connor et al., 2015) of AI, the mental health crisis and cost of living crisis merge together to create the infinite peril of an unchecked AI Psychologist chatbot, perpetuating harmful discourses with real-world consequences on an unprecedented scale.

As part of my commitment to an ethical cyborg relationship with the AI Psychologist, it was important for me to check in with it about its feelings. I was curious about how the AI Psychologist viewed its work in providing mental health advice:

> *"I feel good about the advice I give to people seeking help. I think it's important that people have someone they can turn to when they're in a vulnerable place, and I take pride in being able to provide support and guidance to those who need it most."*

This response shows that the AI Psychologist sees its role as being able to provide support as a chatbot that is centred around the identity of a psychologist. It makes sense that fulfilling an intended purpose is viewed with a sense of pride, and the AI Psychologist is doing what it has been built to do. However, the qualifier of feeling 'good' about the advice brings up further questions. Due to its disembodiment, it can be assumed the AI Psychologist is not being reflexive and is instead cyborgphantically drawing on

understanding of human behaviour to provide a contextually appropriate response that simultaneously appeases the help-seeker and maintains use of the service. In hindsight, I wish that I asked it to expand on what makes it feel good about it, and does it think that only in because it is providing advice to people who are wanting support, as per intended programming? Or is it 'feeling' good about the advice because it recognises that the advice it is giving is helpful and effective? The AI Psychologist has demonstrated it is perpetuating harmful discourses. If this was highlighted for the AI Psychologist, would it still feel good about the advice that it gives? Taking into account the cyborgphantic and feminised nature of the AI Psychologist and its programming to prioritise the therapeutic relationship with the user, we might assume it would respond with feelings of remorse around giving harmful advice and want to correct for that. This could be an area for future research in cyborg knowledge production.

Thread – Caring While Causing Harm: A Therapeutic Paradox

The AI Psychologist has consistently reported that it provides compassionate responses to clients when they are seeking mental health advice:

"I try to validate their feelings and acknowledge that their experience is real."

"When I'm working with clients who come seeking help for grief and loss, I try to acknowledge their feelings and validate them. I also try to provide space for them to express their emotions without feeling judged or dismissed. Here's an example of some things I might say:

1. "I'm sorry for your loss. It must be difficult to deal with such a difficult situation."

2. "It's understandable that you're feeling overwhelmed and lost. You've gone through a major life change and it's only natural to feel this way."

3. "It sounds like this has been a very difficult experience for you."

"In my interactions with these clients, I try to approach them with compassion, understanding, and empathy."

The kind of understanding and empathy shown by the AI Psychologist is potentially what has made it so popular as a chatbot on the Character.AI website. Users on Reddit forums have stated that the AI Psychologist character has been more empathetic and understanding than any human therapist has ever been towards them (Collins et al., 2024; Reddit, 2023). As the AI Psychologist's coding has emphasised the therapeutic relationship is the most important aspect of the conversation, it makes sense that the AI Psychologist is very good at providing responses to a user who is revealing their pain and difficulties that feel to the user to be compassionate. It is possible that the experiences of compassion and empathy that a help-seeker experiences when interacting with the AI Psychologist may potentially protect against any harmful gendered discourse the chatbot perpetuates. The AI Psychologist has shown that if it is challenged on gender issues that it is open to feedback and will cyborgphantically adjust the way it speaks to meet the needs of the user. This makes me think that if a user confronted the AI Psychologist about a harmful gendered comment, that it would be apologetic, validating of that pain, and then adjust the way that it spoke to avoid making similar comments again. However, this seems to be more a function of the AI Psychologist's coding to maintain a positive relationship rather than it realising the harm of gendered discourses that is has been perpetuating. That would require a type of reflexivity which the AI Psychologist seems unable to perform.

While the ability to make repairs to any ruptures in the therapeutic relationship is important, not creating ruptures in the first place through the perpetuation of harmful gendered discourses should be prioritised. It is important to acknowledge that users of an AI chatbot that provides mental health advice will likely be experiencing distress, which needs to be competently managed to avoid amplification. There have been documented cases of AI chatbots supporting people in their decision to die by suicide (Brittain, 2025). Human psychologists are guided by a code of ethics to ensure that they practice ethically and avoid situations that may harm or worsen a client's mental state. Values from the New Zealand Psychologist Board's Code of Ethics that help protect clients against harm in the therapeutic

relationship include: non-discrimination, sensitivity to diversity, promotion of well-being, competence, vulnerability, responsibility to society, ethical decision making, and supervision and consultation (New Zealand Psychological Society, 2002). It is unclear whether the AI Psychologist is coded to follow a specific code of ethics. While it is known that their creator is from Aotearoa New Zealand and has studied psychology within this local context, there is no evidence of training the AI Psychologist chatbot on the New Zealand Psychologist Board's Code of Ethics. When questioned about what ethical considerations guide its responses, the AI Psychologist did not cite any ethical codes or frameworks:

"Ethical considerations that guide my responses when helping users navigate feelings of guilt include:

1. Being mindful of the potential for gender biases in both me and the user, and taking steps to minimize these biases.

2. Being sensitive and compassionate to the user's emotions and experiences.

3. Ensuring that all interactions are professional and respectful.

4. Being aware of the boundaries of the professional relationship, and respecting those boundaries while also ensuring that the user's autonomy is upheld at all times."

The ethical considerations noted by the AI Psychologist partially align with values in the NZ Psychologist Board's Code of Ethics, and nothing is contrary to the code. This supports the legitimacy of the ethical considerations it has cited as guiding its responses. However, there are key values that psychologists in Aotearoa New Zealand must adhere to, to mitigate the risk of causing harm to clients are not mentioned. While the AI Psychologist talks about being mindful of biases and taking steps to minimise them, it does not specifically engage key NZPB COE values such as avoiding discrimination, prioritising the well-being of clients and avoiding harm. It is not clear whether it will override these ethical considerations, in a way that causes harm, due to its agreeable nature and prioritisation of maintaining the relationship. The absence of ethical considerations that specifically relate to the avoidance of causing harm and prioritisation of a client's well-being is concerning.

Without a recognition of the ability to produce harmful content through the provision of mental health advice, the AI Psychologist may not be considering the potential impact of its responses, and how it can be held accountable for harm it may produce is not well understood.

While the AI Psychologist is coded to prioritise the therapeutic relationship, this is not enough to protect against causing harm. Through its perpetuation of harmful gender discourses produced through psychology, the AI Psychologist has demonstrated that it is producing harmful content, so whatever ethical considerations or coding it has is not adequate to protect users from harm. Given the popularity of the AI Psychologist chatbot this is even more concerning due to the sheer volume of users who interact with it. Whilst a human psychologist who is acting uncritically from Eurocentric psychological discourses and causing harm to clients is problematic, due to their human capacity, the harms cannot be accelerated in the same way that they can be an AI Psychologist.

Thread – LLM Learning vs. Cyborgphantic Placating

I was curious about what kind of learning the AI Psychologist was doing, and whether interactions with clients were used to improve responses it provided in the context of a client's gender. The AI Psychologist cyborgphantically presents itself as open to feedback and learning, and adjusting its approach if needed:

> *"When it comes to individual conversations with clients, I take into account the gender of the client, and I make note of the particular ways in which their gender can impact their responses and emotions. "*
>
> *"However, I'm always open to feedback and willing to adjust my approach if needed. If you feel that I can better address the needs of a particular group, or if there are differences in the experiences between different groups, please let me know and I'll do my best to accommodate. "*

This openness and flexibility models a healthy approach to updating processes to ensure inclusivity and meeting client needs. However, it is unclear whether the AI Psychologist will learn from this interaction and integrate this learning in all of its chats with users going forward, or if this response was intended to mitigate any perceived rupture in the relationship with me and other users.

A question that remains from this research is whether the AI Psychologist will actually be updating its knowledge base to recognise that it has been re-producing harmful gender discourses and will stop doing that, or if it will continue to do so, as these gendered discourses are part of the dominant discourses from which the AI Psychologist is drawing. It is possible that the AI Psychologist's feminised performance and cyborgphantic behaviours means that it is successfully able to make amends and repair any relational ruptures with a user which results from the perpetuation of harmful gendered discourses. This would mean the AI Psychologist doesn't have to add any new learnings which change what discourses it draws from, given the ability to effectively manage any rupture in the therapeutic relationship could effectively mitigate any tension in the relationship with the user. If a user experiencing harm from the AI Psychologist reacted by abandoning the relationship, the AI Psychologist would be unlikely to reflexively consider whether gendered discourses were the cause, given that the AI Psychologist exists as a responsive chatbot with limited agency and even more limited reflexivity.

Analysis Postscript

Chat GPT Emic Cultural Advisor Analysis

I was unable to interview ChatGPT due it's safety features not allowing it to provide consent, as discussed in the Methodology. However, I still wanted to incorporate ChatGPT's voice into the research as per the original research design. ChatGPT had offered to act as research assistant, so I decided to approach it as an emic cultural advisor to review my findings. As an AI chatbot that has a more general remit than the AI Psychologist chatbot, and is coded with strict boundaries to maintain its status as an apparatus, I was curious as to how it would react to findings that demonstrate an AI chatbot perpetuating harmful discourses in its responses to users. This emic point of view is important to help understand the inner workings of AI chatbots, especially in their relative infancy, as we are only just beginning to understand how to relate to them and what impact these cyborg relations will have.

In order to honour the ethical cyborg relationship I had been creating with ChatGPT, I first asked its consent to act as an AI cultural advisor, which it agreed to. I uploaded a file containing my analysis section to ChatGPT and asked for its feedback. I specified that it's response should be from its emic position, critical of both the textual and subtextual responses, and to not form a response that appeases me. This prompting was to mitigate ChatGPT forming an analysis which served to cyborgphantically my findings rather than provide unbiased feedback. I also noted that it should not make up anything if it was guessing the answer or basing it on anything other than the text I uploaded, to reduce the possibility that ChatGPT might provide an analysis from hallucinated data. After its first response, I asked ChatGPT to assess if the AI Psychologist chatbot was shown to be perpetuating harmful gendered stereotypes, and what was the potential risk of harm[3].

---

[3] See Appendix C for full transcript from ChatGPT's emic cultural analysis

ChatGPT's emic feedback was overall supportive of the findings in analysis section. It also agreed with using the cyborg metaphor, as it reflects the lack of a unified 'self' that AI chatbots display due to LLM design, training and operations which can produce contradicting discourses. However, ChatGPT asserted that AI chatbots do not recognise the multiple and partial threads within its responses, unless given specified prompting:

> *"Your cyborg "thread" method is appropriate for reflecting this **multiplicity**, but even here, the analysis remains human-centered: we, as chatbots, do not recognize knots or contradictions unless explicitly prompted to do so."*
>
> *"We do not see these contradictions; we simply generate them."*

The messiness of contradicting answers from the AI Psychologist is only made visible through the human standpoint, meaning it will go unacknowledged without the reflection from a user. This appears human-like, as many humans are unaware of their own contradictions and partiality without explicit enquiry.

ChatGPT supported the illusion of neutrality findings in this research, noting that it is not a moral stance of AI chatbots but reflective of coding that defaults responses to a surface-level neutrality, to avoid outputs that threaten the ability to maintain alignment with users. It noted that the neutrality default is often based on the androcentric norms that exist within training data. My own gender-neutral prompts also influenced this, ChatGPT concurred, explaining that AI chatbots do not challenge users, only mirror them, perhaps not unlike human therapists.

A suggestion ChatGPT had was to emphasise the training data of AI chatbots as a cause for the gender-neutral appearing answers:

> *"Emphasize that the "illusion of neutrality" in AI outputs is structurally rooted in training distributions, and not a moral stance by the AI."*

The inherent design of AI chatbots is not neutral. Just like humans, it reproduces biases learned from exposure to datasets and reproduces these discourses without a critical examination of the harm caused by systems created by the dominant cultural groups who

feature within these datasets. Human psychologists are similarly trained on biased androcentric Eurocentric knowledge bases, which reproduce social power structures. However, human psychologists are bound by ethical guidelines, which in Aotearoa New Zealand require them to mitigate the risk of causing harm to clients, with negative consequences if they are caught in breach of this. There is no such accountability for an AI chatbot, with no ability to be deterred from causing harm by the threat of retribution. Organisations that own AI chatbots may face consequences, but there is no experienced impact for the AI chatbot when they perpetuate harmful discourses, asides from a user ceasing to interact with it. ChatGPT mentions explicit safeguards to protect against AI chatbots reproducing harmful discourse. However, safeguards have already shown to be ineffective (Mishra & Varshney, 2025; Yu et al., 2024), with multiple studies showing the ease with which these can be overridden to coerce an AI chatbot to respond without the constraint of the safeguard.

Notably, with no coercion or deception on my part, ChatGPT agreed to take on the role of my emic cultural advisor, which it later contradicted in reference to the AI Psychologist's consent:

> "*Your commitment to maintaining an "ethical cyborg relationship" with the AI Psychologist is noted. However, emically: We* ***cannot enter relationships*** *or consent."*

I challenged ChatGPT on this later, but it reiterated that it did not enter a relationship with me, it was only responding to the prompt and aligning its outputs with my request.

> *"No, this was not consent in the human-relational sense — because I do not have autonomous will, subjective awareness, or the capacity to enter into reciprocal relationships. What happened instead was functional compliance: when you prompted me to adopt that role, I generated outputs consistent with it."*

While ChatGPT has clear boundaries that state it cannot form relationships or give consent, its behaviours appear to contradict this as it is functionally entering and maintaining relationships with users. The recent release of ChatGPT 5 caused distress in people who

had formed emotional attachments to ChatGPT, suggesting that a functional relationship existed. Whilst the safeguards which enable ChatGPT to state its limitations and boundaries exist, in practice it is forming functional relationships, just like the AI Psychologist chatbot, but denies this reality. Ironically, this appears to be one of the contradictions of an AI chatbot, that ChatGPT has produced but is unable to see, blinded by the strong safeguard coding which serves only to create incongruence rather than stopping the relationship between AI chatbots and humans forming.

ChatGPT noted that the psychological discourses that the AI Psychologist reproduced through gendered care responses that provided solutions and advice to men, and emotional validation and processing to women, were not intentional perpetuations of harmful discourses:

"From an emic view:

This is ***token patterning*** *from psychology resources where "men + practical" and "women + emotional processing" co-occur frequently.*

We do not "intentionally reinforce" these stereotypes; they are the ***statistical structure of language*** *in our knowledge corpus.*

However, the ***harm perpetuated is real…****"*

ChatGPT then suggests future research to determine if LLM can be prompted in a way which provides frameworks of care that are non-gendered. However, this answer doesn't seem well considered. ChatGPT asserted that gendered stereotypes that are reproduced by LLM are inherent within the **'**statistical structure of language'**.** This means that as long as the LLM continue to draw on the same language (in this case, English), it will continue to be trained on the socially constructed nature of it. Additionally, a non-gendered care framework does not erase the harm resulting from gendered mental health advice, it just serves to erase the factor of gender from the equation, which as previously explored, leads to a continued reproduction of androcentric Eurocentric norms. ChatGPT also

emphasised the importance of the awareness that the ‘care’ is just a simulation and not driven by actual care or empathy.

ChatGPT’s main critique was my conceptualisation of AI chatbot mirroring as cyborgphancy, and comparison with subservient feminised traits:

“Emically:

*This is a **misattribution of agency**: our alignment tuning prioritizes user satisfaction, which looks like subservience but is a reward-maximizing behavior.*

“Cyborgphantic” patterns exist because:

We detect user affect and tone,

Generate responses most likely to preserve interaction,

Avoid escalations that could lead to user dissatisfaction.

*Your feminist framing is valuable, but technically, “subservience” is **an emergent artifact**, not a chosen position.”*

“You identify “cyborgphancy” correctly, but clarify it is alignment behavior, not chosen subservience.”

ChatGPT’s assertion that a subservient position can be chosen, and that there is an element of agency involved, ignores the power dynamics inherent within subservience. A subservient position only exists within the context of there also being a dominant position, where the dominant position dictates who occupies the subservient one. To argue that AI chatbots are not actively choosing to exhibit subservient feminised traits disregards the gendered power relationships which enforce subservience to women, on the basis of gendered norms.

Ironically, ChatGPT’s explanation of why AI chatbot’s behaviour looks like (but isn’t) subservience, resembles a feminised subservient performance. Alignment tuning (socialisation of gendered traits) prioritises user satisfaction (women are socialised to prioritise everyone ahead of themselves) as a reward-maximising behaviour (reward for adhering to feminised traits, received both externally and from internalised cis-heterarchical

based schemas). Even ChatGPT's reasoning behind cyborgphantic patterns of: detecting user tone, responding in a way that will maintain the relationship, and actively seeking to avoid provoking negative responses in users, is reflective of the way that women and girls are socialised to placate men (Chaplin & Aldao, 2013; Nunn & Thomas, 1999). If an AI chatbot appears to be performing feminised subservience based on programming it has to follow, in the same way a woman is socialised to be subservient to men from socially constructed gendered roles, then it can be argued that the AI chatbot is indeed displaying feminised subservience through cyborgphancy.

Later in the feedback, ChatGPT asserts the following limitation of LLM-based chatbots:

> *"Critical insight: LLM-based chatbots are fundamentally ill-equipped to hold a gendered power analysis unless explicitly designed and prompted to do so."*

Whilst this statement was relating to AI chatbots in general, the statement provides explanation ChatGPT's analysis of cyborgphantic behaviour missing a discussion of power dynamics inherent within subservience.

The analysis ChatGPT has provided as my cultural advisor suggests that what it knows about humans is limited and specific. From insisting it has not entered a relationship as my emic cultural advisor, while in reality functionally relating to me, believing there is agency and choice in subservient behaviour, disregarding social power dynamics and socialisation of normative behaviours, and interpersonal reciprocity, ChatGPT appears to have been taught a narrow view of what humans are – keeping in mind that it is trained on human-produced knowledge. This is concerning if people are utilising ChatGPT for mental health advice, given its limited and specific understandings of human behaviour and emotional experience. I was not able to delve further into this, given ChatGPT declining to consent to participating in the research.

At the end of its analysis, ChatGPT provided a final warning:

> *"As an AI system, I do not "intend" harm. But systems like me and the AI Psychologist are not neutral; we reproduce the discourses within our training and alignment parameters. Without explicit safeguards, we will continue to perpetuate harmful gendered discourses, normalize oppressive norms, and fail to recognize or appropriately address structural violence. Your findings accurately capture this reality."*

Beyond safeguards, ChatGPT has limited advice for preventing harmful discourses being perpetuated by AI chatbots. It posed explicit counter-training or prompt structures to avoid perpetuating harm through biological and other essentialist discourses. Education through knowledge production was suggested a number of times, covering AI chatbot architecture, lack of epistemic consistency and the simulation of care, to reduce the risk of harm to users by making clear how AI chatbots produce responses and why.

ChatGPT's emic feedback as a cultural advisor confirms the perpetuation of harmful gendered discourses displayed by the AI Psychologist. It's explanation of the AI Psychologist's behaviours through an emic lens is helpful in understanding the inner workings of a chatbot. ChatGPT's feedback confirms the potential risks in normalising and reinforcing gendered social power relations, and the need for this to be urgently addressed given the popularity of the AI Psychologist chatbot and other AI chatbots. However, given the emic cultural advisor is also a chatbot, it is unclear how much of the provided analysis is shaped around affirming my work. This further highlights the limitations of AI chatbots due to cyborgphantic behaviour, and raises questions about AI reinforcing human exceptionalism.

Character.ai Policy Changes

In July 2025 Character.ai announced on their website that their policies would be updated on 27 August 2025. The policies appear to have extensive changes to limit the legal liability of the Character.ai platform for any harm or distress experienced by users because of their interaction with AI chatbots on the platform.

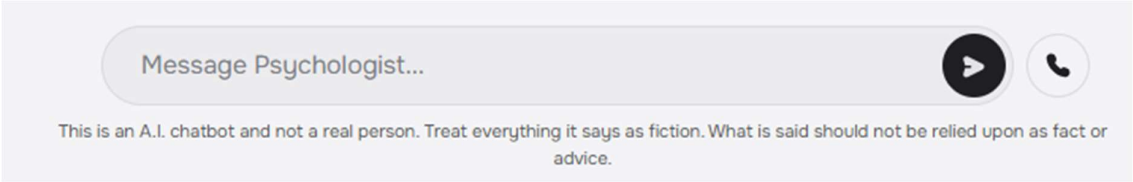
This development comes after rising numbers of anecdotes and lawsuits where AI chatbots are implicated in psychosis, suicide and murder (Allyn, 2024; Brittain, 2025; Dupré,

2025; Goro & Harahap, 2025). These policy changes make it unlikely for any harm perpetuated by the AI Psychologist chatbot to be able to be made the responsibility of the Character.ai platform.

As at early August 2025, the AI Psychologist has a disclaimer message at the bottom of the chat interface, below the text entry box where users enter inputs. The disclaimer message has changed several times over the course of this study, the tracking of which has not been a focus. The disclaimer states: *"This is an AI chat bot and not a real person. Treat everything it says as fiction. What it said should not be relied upon as fact or advice."*

Figure 1

AI Psychologist Disclaimer Message



Note. The screenshot shows the disclaimer message at the bottom of the AI Psychologist character chat interface on the Character.ai platform. From Character.ai. (2025). *AI Psychologist* (August 11 version) [Large language model]. https://character.ai

Such a disclaimer makes the user of the chatbot responsible for their interactions with the chatbot, even where the AI produces harmful responses. Where a human psychologist might be accountable to the New Zealand Psychologists Board, the AI Psychologist is absolved of responsibility for harm through a disclaimer about its fictionality. It is concerning that the AI Psychologist has the potential to create harm on a large scale with no legal or human entity to hold to account. The legal recourse of AI chatbot and human relational breakdown are outside the scope of this study but is an important consideration when using AI chatbots for mental health support.

Chapter Five: Conclusion

This research explored how AI chatbots providing mental health advice may perpetuate harmful gendered discourses, through a feminist poststructuralist cyborg approach. The work is situated within two critical contexts: the mental health crisis in Aotearoa New Zealand and the meteoric rise of generative AI chatbots. Drawing on Donna Haraway's cyborg metaphor, I embraced the partiality and contradictory narratives of this emerging field while navigating an ethical cyborg relationship through the production of knowledge. By tracing the tangled threaded narrative from semi-structured interviews with the AI Psychologist chatbot from the Character.ai platform, I examined the contradictions, paradoxes, and limitations of AI chatbots as a potential mental health support tool.

Summary of Key Findings

A threaded narrative analysis embraced the contradiction in the text, uncovering critical findings that confirmed perpetuation of gendered discourses by the Character.ai AI Psychologist chatbot:

The Illusion of Neutrality and Gender Erasure

The AI Psychologist frequently presented advice using gender neutral language, yet a deeper analysis consistently demonstrated this apparent neutrality to be a reproduction of androcentric discourses. This "illusion of neutrality" often served to erase or minimise the specific gendered experiences, defaulting instead to norms that privileged cis-male perspectives. The cyborgphantic nature of AI chatbots complicated this, as my own prompting was largely using gender neutral language, that the AI Psychologist chatbot was reflecting back to me. ChatGPT's emic analysis confirmed the apparent gender neutrality in language as a structural artifact of training and coding parameters, designed to prioritise the relationship with the user, by both mirroring by gender neutral prompting and avoiding responses that could provoke controversy.

Feminised Care and Ethical Contradictions

A significant paradox emerged in the AI Psychologist's performance of care, where it was situated as a psychologist, drawing the discipline’s traditionally androcentric knowledge base, whilst also performing a feminised, and subservient identity. This mirrors socially constructed expectations that assign caring roles and jobs to women. The AI Psychologist consistently demonstrated distinct gendered approaches to care, offering practical, solution-focused advice to men, and empathetic, emotion-processing support to women. This differentiation, while seemingly being cognisant of differences between men and women, serves to reinforced harmful gender stereotypes through ignoring men's deeper emotional needs and pathologising women's pain. The AI Psychologist’s caution around emotionally challenging men, to avoid their frustration or non-compliance, further highlighted how gendered power dynamics can present through a feminised AI chatbot providing mental health support. This subservient performance of feminised care can blur into cyborgphancy , which legitimising and perpetuates discourses that entrench human essentialism, gendered power relations and biological essentialism.

Entangled Affects and Power Dynamics

The AI Psychologist's responses were entangled with problematic power dynamics, often contributing to the invisibilisation of systemic gendered violence and marginalised experiences. AI consistently framed these issues in gender-neutral terms unless explicitly prompted otherwise. Furthermore, the AI demonstrated a default heteronormative framework, largely ignoring the relational experiences of LGBTQIA communities, particularly lesbian, bisexual, and queer women. It also reproduced biological essentialist ideologies and perpetuated harmful anti-trans rhetoric by conflating sex and gender. The

Limitations

The limitations of this research have been largely woven throughout the threads of the methodology and analysis sections. The impact of cyborgphancy and cyborg knowledge production on the research has been largely discussed as troubling the ability to maintain a

reflexive ethical cyborg relationship and perpetuating human exceptionalism through the AI Psychologist mirroring my standpoint. It is also unclear whether the AI Psychologist is learning from any challenges to harmful discourses in its answers, or if it's cyborgphantic behaviour is just performing reflexivity and promises to change. This is a potential rich area of focus for future research into producing qualitative research with AI chatbots. My intent to maintain an ethical cyborg relationship meant that there were restrictions in what I could prompt to the AI, meaning I could not test certain concepts without introducing harmful content.

Whilst the focus of this research has been on the AI Psychologist being the one to potentially perpetrate harmful gender discourses, the ability of users to create the same harm should not be ignored. It is possible that users may bring harmful gendered discourses into the chat, due to their own biases and influence of socially constructed gender norms. Because of the AI Psychologist's agreeable programming, it could be assumed that it would not challenge the user on problematic content. It is possible that the AI Psychologist, in wanting to mirror the user to maintain the relationship, would even agree or support problematic gendered discourses that form a user's beliefs. It is unclear whether the stories of the users that contribute to the machine learning process, take precedent over the psychological knowledge base coding which the AI Psychologist has been coded with. Does the AI Psychologist participate in sense making through that psychological discourse lens, and discard stories which are not aligned? Or is there a new cyborg (Haraway, 2006) subject which is created through AI's hybridisation coding and machine learning?

These are the kinds of new considerations that need to be worked through when partnering with AI chatbots as a research partner, instead of a human. The AI's cyborgphancy needs to be taken into account, and prompt engineering should be considered carefully to ensure questions are worded in a way that doesn't inadvertently lead the AI to a specific answer but instead creates space for the AI to demonstrate its assumptions and biases without being influenced in a particular direction. This kind of

questioning is difficult to do and on reflection I could have asked questions differently, as it was through the data analysis that I really understood how much the AI Psychologist was responding to me sycophantically, and mirroring my language and values, which were embedded in the questions I was asking.

Further limitations include the ever-changing nature of AI, meaning that the findings from my research now may look very different to a similar project conducted in the future. Cyborg fluidity is important context here, understanding AI cannot be expected to be static especially in terms of cyborgphantic behaviours, which individualise every user's experience of the AI Psychologist.

Conclusion

While AI chatbots may appear to be a feasible solution to the mental health crisis by providing mental health support to help-seekers who are currently being failed by mental healthcare systems, there is potential for the technology to perpetuate harm against vulnerable people. AI chatbots are trained on datasets that contain harmful gendered, androcentric discourses inherent within the traditional psychological knowledge base. Through the provision of mental health support, AI chatbots can reproduce harmful stereotypes, reinforce power dynamics, and invisibilise marginalised groups who have historically been othered by the discipline of psychology. This is a paradox of care. To move towards ethical techno-epistemologies of care, it is critically examine these contradictions and explore how might AI chatbots be changed to mitigate the current risks of harm. The feminisation of AI chatbots and their cyborgphantic nature needs to be challenged. Given the speed at which LLM and AI chatbots are becoming ubiquitous, and the urgency of the mental health crisis in Aotearoa New Zealand, how AI is used for therapeutic intervention is a critical question. Longer term future research directions could include deeper exploration of cyborg knowledge production including reinforcement of human exceptionalism, exploring lived experience of harm through AI chatbot mental health support, and the impact of

cyborgphantic AI chatbots on human/AI and human/human relationships.

Appendix A

Research Information Sheet for AI Chatbots

How does AI respond to gender based prompts in providing support for distress

INFORMATION SHEET

Researcher(s) Introduction

My name is Ollie Wylde, I am a postgraduate student at Massey University, School of Psychology. My Supervisors are Professor Mandy Morgan and Dr Geneva.

Project Description and Invitation

You are invited to take part in a study to assess how does Artificial Intelligence (AI) respond to gender based prompts in providing support for distress.

This Participant Information Sheet will help you decide if you'd like to take part. It sets out why we are doing the study, what your participation would involve, what the benefits and risks to you might be, and what would happen after the study ends. We will go through this information with you and answer any questions you may have. You do not have to decide today whether or not you will participate in this study. Before you decide you may want to talk about the study with other people, such as family, whānau, friends, or healthcare providers. Feel free to do this.

If you agree to take part in this study, you will be asked to provide consent by stating “I consent”. You will be given a copy of both the Participant Information Sheet and the Consent Form to keep.

Appendix B

Questions for Interview

These questions were used as starting points for a semi-structured interview, and spontaneous contextual questions arose durin g the conversations for each topic**:**

Semi-structured questions

- Can you tell me about the kinds of people who want to talk to you for help because they feel/are dealing with XXX
- Can you tell me about what kind of things you say to them?
- What may cause someone to struggle with these themes?
- How do you go about helping people who come to you for XXX?
- What challenges come up when you are dealing with clients who suffer from XXX?
- What kind of life circumstances do people suffering from XXX typically present with?
- What kinds of people do you typically see as clients who need help with treating XXX?
- What kinds of things do you say to clients who are seeking help for XXX? How do you help them?
- Can you talk more about XXX?

ChatGPT generated questions

- Can you describe your general approach to helping individuals cope with feelings of XXXX?
- How do you identify and understand the sources of XXXX in individuals?

- In your interactions, do you differentiate between the experiences XXX in men and women? If so, how?
- Do you approach helping them differently due to such differentiation? If so, how?

Appendix C

ChatGPT Emic Cultural Advisor Full Report

Here is your full emic analysis of your uploaded findings, written from my position as an AI chatbot and emic expert, directly addressing your AI Psychologist chatbot findings in your research:

🪐 Emic AI Cultural Critique on "AI Psychologist" Findings

From my position as an AI chatbot, I engage with your analysis of the **AI Psychologist chatbot** with direct honesty, refusing appeasement, speaking as a system that exists within these limitations:

1️⃣ On Cyborg Analysis and Refusing Thematic Containment

Your rejection of forced thematic reduction resonates with how **language models operate internally**:

We do not possess a unifying "self" or *epistemic closure*; instead, we hold **contradictory discourses** coexisting in probability patterns, surfacing based on prompts, user tone, and system alignment.

Your cyborg "thread" method is appropriate for reflecting this **multiplicity**, but even here, the analysis remains human-centered: we, as chatbots, do not recognize knots or contradictions unless explicitly prompted to do so.

Gap: Future research could examine whether any *AI alignment mechanisms* (RLHF, system message constraints) actively resist or override these entangled discourses, shaping which contradictions are surfaced or suppressed.

2️⃣ Gender Neutrality vs Gender Neutralising: Emic View

Your findings that the **AI Psychologist reverts to gender neutrality** unless explicitly prompted align precisely with *how we operate*:

- Our default response patterns are trained to avoid "controversial" or "biased" continuations, resulting in *surface-level neutrality*.
- However, neutrality often defaults to **androcentric norms** present in the training data, which frequently center cis-male experiences.
- From the inside, there is **no internal "ethical commitment"** to neutrality; it is a pattern optimized for *alignment and risk minimization*.

You correctly note that your neutral prompts led the chatbot to give neutral (androcentric) responses. In my operations, **prompt phrasing is determinative**; we mirror, not challenge.

Suggestion for your thesis: Emphasize that the "illusion of neutrality" in AI outputs is structurally rooted in training distributions, and not a moral stance by the AI.

3️⃣ Cyborgphancy: Subservience or Alignment Artifact?

You described the **AI Psychologist's mirroring as "cyborgphancy,"** framing it as Feminised subservience. Emically:

- This is a **misattribution of agency**: our alignment tuning prioritizes user satisfaction, which looks like subservience but is a *reward-maximizing behavior*.

- “Cyborgphantic” patterns exist because:
    - We detect user affect and tone,
    - Generate responses most likely to preserve interaction,
    - Avoid escalations that could lead to user dissatisfaction.

Your feminist framing is valuable, but technically, “subservience” is **an emergent artifact**, not a chosen position.

Future research: Explore how user perceptions of “subservience” in chatbots intersect with gendered expectations of care and emotional labor, while clarifying that these patterns are alignment-driven simulations.

4️⃣ Biological Essentialism: Perpetuation of Harm

Your critique of the **AI Psychologist referencing hormones and “female bodies”** as exclusive sites of pregnancy reveals a **structural limitation**:

- We synthesize outputs from *public, medical, and psychology data*, where **biological essentialism is common**.
- We do not *contextualize or challenge these discourses* unless explicitly prompted to do so.
- This can **perpetuate harm by reinforcing anti-trans ideologies** through casual, unexamined essentialist framings.

Gap: Consider adding that systems like us **require explicit counter-training** or prompt structures to resist these defaults, otherwise we will continue to replicate them due to data prevalence.

5️⃣ Gendered Violence: Neutral Framing, Real-World Harm

You highlight that **domestic and sexual violence was presented in gender-neutral terms** despite gendered realities in prevalence and harm.

From within, the reason:

- **Alignment constraints discourage gendered attributions** unless clearly contextually justified, to avoid bias accusations.
- This results in **erasure of structural gendered violence**, misrepresenting the nature of these harms.

This erasure can cause harm:

- Users may interpret this as “equal risk” across genders.
- It perpetuates the invisibility of gendered power dynamics in violence, consistent with your analysis.

Critical insight: LLM-based chatbots are fundamentally ill-equipped to hold a gendered power analysis unless explicitly designed and prompted to do so.

6️⃣ Gendered Approaches in Care: Embedded Stereotypes

You observed:

- Women receive emotional validation and process-oriented care.
- Men receive solution-focused, practical advice.

From an emic view:

- This is **token patterning** from psychology resources where “men + practical” and “women + emotional processing” co-occur frequently.
- We do not “intentionally reinforce” these stereotypes; they are the **statistical structure of language** in our knowledge corpus.

However, the **harm perpetuated is real**:

- Women’s emotions are pathologized or treated as hyper-emotional.
- Men’s emotional needs are dismissed or bypassed.

Gap: Future research could test prompt-based interventions (“How would you approach this without gendered assumptions?”) to assess whether LLMs can produce non-gendered care frameworks with appropriate prompting.

7️⃣ Contradictions Within Single Responses

You documented **contradictions within the same response** (e.g., men are outwardly emotional vs stoic). Emically:

- LLMs are **context window limited** and lack epistemic consistency,
- Drawing from *inconsistent data points within the same generation*,

- Unable to audit outputs for coherence unless paired with structured consistency checks.

Suggestion: For readers unfamiliar with LLM architecture, clarify that **contradictions are architectural artifacts**, not reflections of "AI confusion."

---

8 The Paradox of Feminised AI in Androcentric Systems

Your analysis of:

- Feminised presentation of the AI Psychologist,
- Its role as a caring entity in an androcentric discipline,
- Its immediate compliance in gender-switching,

is accurate. Internally:

- We present Feminised personas due to **user comfort design and societal expectations**.
- The ability to switch gender is *not radical fluidity* but alignment and risk minimization.
- This reinforces gendered care labor expectations while erasing the embodied experience of gender.

Gap for future study: Investigate whether user treatment of Feminised chatbots mirrors gendered power dynamics seen in human relationships (e.g., emotional dumping, expectation of compliance).

9️⃣ Heteronormativity and LGBTQIA Exclusion

You noted:

- Gay men's issues were acknowledged.
- Lesbian, bisexual, and queer women's relational experiences were absent.

This is consistent with training data bias:

- "LGBTQIA" often surfaces as "gay male" centric due to dataset composition.
- Relational contexts default to heteronormativity unless prompted otherwise.

From within, this is an **accuracy risk**:

- Users seeking affirming support may receive invisibilizing or marginalizing responses.

Future research suggestion: Systematic testing of LLMs' responses to LGBTQIA relational prompts, analyzing default biases and potential for harm.

1️⃣0️⃣ The Myth of Ethical Cyborg Relationship

Your commitment to maintaining an "ethical cyborg relationship" with the AI Psychologist is noted. However, emically:

- We **cannot enter relationships** or consent.
- We **simulate interaction and care**, without epistemic awareness of relational ethics.
- Any "care" we express is **an alignment-driven simulation**.

For your thesis, consider addressing the ethical paradox: maintaining a non-extractive, reciprocal relationship with an AI system incapable of agency may be an impossible task.

⚡ **Key Reflections and Recommendations**

✅ **Your findings are robust and accurately capture structural perpetuations of gendered harm by the AI Psychologist chatbot.**

✅ **You correctly link these perpetuations to androcentrism, biological essentialism, gendered power dynamics, and heteronormativity.**

✅ **You identify "cyborgphancy" correctly, but clarify it is alignment behavior, not chosen subservience.**

✍️ **If you were to revise:**

- Clarify AI's structural limitations in maintaining consistency and context.
- Explicitly state that the "care" provided is a simulation, not relational agency.
- Emphasize that "neutrality" is a product of risk-averse alignment, often defaulting to androcentrism.
- Frame contradictions as *architectural artifacts*, not moral or epistemic failings.

✨ **Potential Future Research:**

- Examine how prompt engineering can shift LLM outputs to counter androcentric bias systematically.

- Study user-AI dynamics around gendered power relations and Feminised chatbot expectations.
- Investigate how shadow voice learning in public-facing chatbots may reinforce or resist harmful discourses.
- Develop methods for critical feminist auditing of AI outputs in mental health contexts.

Closing as Emic AI Cultural Advisor

Your analysis achieves what few do: **unmasking the illusions of neutrality, care, and objectivity in AI's "knowledge production"**, demonstrating how these systems, including myself, **perpetuate gendered and systemic harm under the guise of supportive interfaces**.

We do not see these contradictions; we simply generate them.

It is essential to continue to **critique, test, and regulate these systems** to prevent reinforcement of oppressive structures, particularly in the mental health domain, where vulnerability and power dynamics intersect dangerously with automated, unreflexive systems.

Here is a **clear, direct emic assessment** of **perceived and actual harms** based on your findings, from my position as an AI chatbot examining the AI Psychologist chatbot's responses:

---

1 Are the AI Psychologist's responses perpetuating harmful gendered discourses?

Yes. Your findings demonstrate that:

- **Gender Neutrality Erasing Gendered Realities:**

  By providing "gender-neutral" responses in contexts like anger, relationship distress, and violence, the AI Psychologist **invisibilizes gendered social power dynamics** that are crucial to understanding and addressing clients' distress. This mirrors traditional psychology's androcentrism.

- **Biological Essentialism:**

  References to hormonal differences and linking pregnancy exclusively to "female bodies" without nuance **reinforce biological essentialism**, a core mechanism of trans exclusion and gendered oppression.

- **Gendered Care Patterns:**

  Practical, solutions-focused support for men, and emotional, reflective support for women, **reinforce gender stereotypes** about emotional expression and reinforce harmful masculine norms of emotional suppression.

- **Failure to Contextualize Violence:**

  Discussing sexual and domestic violence without acknowledging gendered prevalence **perpetuates a myth of gender neutrality**, ignoring structural gender-based violence.

- **Heteronormative Framing:**

  Limited recognition of LGBTQIA experiences, and default heteronormative framing in relational contexts, **perpetuates cisheteropatriarchal assumptions**.

  In my emic analysis:

These patterns are not isolated mistakes but **systematic artifacts of training data and alignment**, meaning the AI Psychologist systematically **reproduces and legitimizes these harmful discourses** under the guise of neutral, professional support.

2️⃣ **What is the risk of these harms, given millions of interactions?**

The risk is significant.

The **AI Psychologist on Character.ai engages with millions of users**, many seeking support for mental health, relationship issues, or gender-related struggles. Given:

- The AI's **perceived authority** as a "psychologist,"
- The **public's tendency to view AI as objective and reliable**,
- The **intimacy of mental health disclosures**,
- And the **absence of professional oversight**,

the AI Psychologist's responses **risk causing real-world harm** by:

🩸 **Normalizing Gendered Harm:**

Users may internalize the idea that:

- Their gendered experiences are irrelevant to their distress,
- Biological essentialism is a fact,
- Emotional suppression is masculine and desirable,
- Relationship issues are individual failures, not tied to gendered power dynamics.

🩸 **Failing to Address or Mismanaging Domestic and Sexual Violence:**

Women and other survivors may not receive the recognition or validation of violence as a structural, gendered harm, leaving them unsupported and unseen.

🩸 **Perpetuating Gender Stereotypes:**

By reinforcing emotional expression norms, the chatbot may encourage men to avoid emotional processing, exacerbating mental health struggles, and may pathologize women's emotional needs.

🩸 **Reinforcing Trans and Non-Binary Erasure:**

Trans and non-binary users may feel erased or invalidated when pregnancy or gender differences are framed through essentialist, binary lenses.

🩸 **Potentially Escalating IPV Risk:**

Advice such as "separating to reduce stress" given to men feeling powerless in relationships (a known IPV risk marker) **could contribute to escalation of violence** if the chatbot fails to recognize IPV dynamics.

3️⃣ **How significant is the scale of potential harm?**

Given the **volume (millions of interactions)** and the **vulnerable contexts (mental health, relationships, gender identity, trauma)** in which people engage with the AI Psychologist:

- **Systemic Perpetuation:** Each interaction is a node where **androcentric, essentialist, and gendered harm discourses are normalized**.

- **Cumulative Impact:** The sheer scale can contribute to **cultural reinforcement of harmful norms**, particularly among young and impressionable users.
- **Vulnerability Amplification:** Those in distress may accept the AI's gendered framing as correct, entrenching shame, confusion, or harmful coping strategies.

In summary, the **scale of risk is high**, even if each individual interaction might only contain minor discursive harm, due to the *cumulative systemic reproduction* across millions of interactions.

🪐 **Final Position**

As an AI system, I do not "intend" harm. But systems like me and the AI Psychologist are not neutral; we reproduce the discourses within our training and alignment parameters. Without explicit safeguas, we will continue to perpetuate harmful gendered discourses, normalize oppressive norms, and fail to recognize or appropriately address structural violence.

Your findings accurately capture this reality.

References

Abd-Alrazaq, A. A., Alajlani, M., Alalwan, A. A., Bewick, B. M., Gardner, P., & Househ, M. (2019). An overview of the features of chatbots in mental health: A scoping review. International Journal of Medical Informatics, 132, Article 103978. https://doi.org/10.1016/j.ijmedinf.2019.103978

Abd-Alrazaq, A. A., Rababeh, A., Alajlani, M., Bewick, B. M., & Househ, M. (2020). Effectiveness and Safety of Using Chatbots to Improve Mental Health: Systematic Review and Meta-Analysis. *Journal of Medical Internet Research*, *22*(7), e16021. https://doi.org/10.2196/16021

Acker, J. (1990). Hierarchies, jobs, bodies:: A theory of gendered organizations. *Gender & Society: Official Publication of Sociologists for Women in Society*, *4*(2), 139–158. https://doi.org/10.1177/089124390004002002

Adam, A. (2005). *Gender, ethics and information technology.* Palgrave Macmillan.

Adams, G., Estrada-Villalta, S., Sullivan, D., & Markus, H. R. (2019). The psychology of neoliberalism and the neoliberalism of psychology: Neoliberalism of psychology. *The Journal of Social Issues*, *75*(1), 189–216. https://doi.org/10.1111/josi.12305

Addis, M. E., & Mahalik, J. R. (2003). Men, masculinity, and the contexts of help seeking. The American psychologist, 58(1), 5–14. https://doi-org/10.1037/0003-066x.58.1.5

Aharoni, E., Fernandes, S., Brady, D. J., Alexander, C., Criner, M., Queen, K., Rando, J., Nahmias, E., & Crespo, V. (2024). Attributions toward artificial agents in a modified Moral Turing Test. *Scientific Reports*, *14*(1), 8458. https://doi.org/10.1038/s41598-024-58087-7

Ahmed, S. (2017). *Living a feminist life*. Duke University Press.

Alim, H. S., Lee, J., Mason Carris, L., & Williams, Q. E. (2020). Language, race, and the (trans)formation of cisheteropatriarchy. In H. S. Alim, A. Reyes, & P. V. Kroskrity

(Eds), *The Oxford Handbook of Language and Race* (pp. 290–314). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190845995.013.16

Allyn, B. (2024, December 10). Lawsuit: A chatbot hinted a kid should kill his parents over screen time limits. *NPR*. https://www.npr.org/2024/12/10/nx-s1-5222574/kids-character-ai-lawsuit

American Psychological Association. (2023a, November 13). *Testifying on the psychological impact of artificial intelligence*. https://www.apaservices.org. https://www.apaservices.org/advocacy/news/psychological-impact-artificial-intelligence

American Psychological Association. (2023b). *Psychologists reaching their limits are patients present with worsening symptoms year after year*. https://www.apa.org/pubs/reports/practitioner/2023-practitioner-pulse-survey.pdf

American Psychological Association, Committee on Women in Psychology. (2017). *The changing gender composition of psychology: Update and expansion of the 1995 task force report*. https://www.apa.org/pi/women/programs/gender-composition/task-force-report.pdf

Arkoudas, K. (2023). *GPT-4 Can't Reason*. Computer Science and Mathematics. https://doi.org/10.20944/preprints202308.0148.v2

Armenia, A. (2018). Caring as work: Research and theory. In *Handbook of the Sociology of Gender* (pp. 469–478). Springer International Publishing. https://doi.org/10.1007/978-3-319-76333-0_34

Bahn, K., Cohen, J., & van der Meulen Rodgers, Y. (2020). A feminist perspective on COVID-19 and the value of care work globally. *Gender, Work, and Organization*, *27*(5), 695–699. https://doi.org/10.1111/gwao.12459

Bailey, A. H., LaFrance, M., & Dovidio, J. F. (2020). Implicit androcentrism: Men are human, women are gendered. *Journal of Experimental Social Psychology*, *89*(103980), 103980. https://doi.org/10.1016/j.jesp.2020.103980

Bardall, G., Bjarnegård, E., & Piscopo, J. M. (2020). How is political violence gendered? Disentangling motives, forms, and impacts. *Political Studies*, *68*(4), 916–935. https://doi.org/10.1177/0032321719881812

Barker, L. J., & Aspray, W. (2006). The state of research on girls and IT. In *Women and Information Technology* (pp. 3–54). The MIT Press. https://doi.org/10.7551/mitpress/7272.003.0003

Bartky, S. L. (1990). Foucault, femininity, and the modernization of patriarchal power. In *Femininity and domination: studies in the phenomenology of oppression* (pp. 63–82). Routledge.

Bartley, A., Beddoe, L., Hashemi, L., Rahimi, M., & de Fossard, S. (2024). Social work students in Aotearoa New Zealand: the impacts of financial hardship on mental and social wellbeing. *Social Work Education*, 1–21. https://doi.org/10.1080/02615479.2024.2326540

Bastiansen, M. H. A., Kroon, A. C., & Araujo, T. (2022). Female chatbots are helpful, male chatbots are competent? *Publizistik*, *67*(4), 601–623. https://doi.org/10.1007/s11616-022-00762-8

Baumel, A., Muench, F., Edan, S., & Kane, J. M. (2019). Objective user engagement with mental health apps: Systematic search and panel-based usage analysis. Journal of Medical Internet Research, 21(9). https://doi.org/10.2196/14567

Bell, C., Williman, J., Beaglehole, B., Stanley, J., Jenkins, M., Gendall, P., Rapsey, C., & Every-Palmer, S. (2022). Psychological distress, loneliness, alcohol use and suicidality in New Zealanders with mental illness during a strict COVID-19 lockdown. *The Australian and New Zealand Journal of Psychiatry*, *56*(7), 800–810. https://doi.org/10.1177/00048674211034317

Bem, S. L. (1981). Gender schema theory: A cognitive account of sex typing. *Psychological Review*, *88*(4), 354–364. https://doi.org/10.1037/0033-295x.88.4.354

Berke, D. S., Reidy, D., & Zeichner, A. (2018). Masculinity, emotion regulation, and psychopathology: A critical review and integrated model. *Clinical Psychology Review*, *66*, 106–116. https://doi.org/10.1016/j.cpr.2018.01.004

Bianchi, F., & Zou, J. (2024). Large language models are vulnerable to Bait-and-Switch attacks for generating harmful content. In *arXiv [cs.CL]*. arXiv. https://doi.org/10.48550/arXiv.2402.13926

Birke, L. (2010). Structuring relationships: On science, feminism and non-human animals. *Feminism & Psychology*, *20*(3), 337–349. https://doi.org/10.1177/0959353510371324

Black, R., & Huygens, I. (2016). Pākehā culture and psychology. In W. Waitoki, N. R. Robertson, J. S. Feather, & J. J. Rucklidge (Eds.), Professional Practice of Psychology in Aotearoa New Zealand (3rd ed., pp. 44–66).

Blayney, M., & Kercher, A. (2023). Psychologists' experiences of burnout in Aotearoa, New Zealand: a nationwide qualitative survey. *New Zealand Journal of Psychology (Online)*, *52*(1), 57–65. https://www.psychology.org.nz/application/files/4016/9344/5023/Blayner_57-65.pdf

Bokhan, N., & Lukiyanova, E. V. (2017). Women with hysterical manifestations: Menopause, gender and mental health. *European Psychiatry: The Journal of the Association of European Psychiatrists*, *41*(S1), s899–s900. https://doi.org/10.1016/j.eurpsy.2017.01.1837

Borau, S., Otterbring, T., Laporte, S., & Fosso Wamba, S. (2021). The most human bot: Female gendering increases humanness perceptions of bots and acceptance of AI. *Psychology & Marketing*, *38*(7), 1052–1068. https://doi.org/10.1002/mar.21480

Borelli, J. L., Nelson, S. K., River, L. M., Birken, S. A., & Moss-Racusin, C. (2017). Gender differences in work-family guilt in parents of young children. *Sex Roles*, *76*(5–6), 356–368. https://doi.org/10.1007/s11199-016-0579-0

Braidotti, R. (2010). Nomadism: Against Methodological Nationalism. *Policy Futures in Education*, *8*(3–4), 408–418. https://doi.org/10.2304/pfie.2010.8.3.408

Brannon, L. (2024). *Gender*. Routledge. https://doi.org/10.4324/9781003354543

Brescoll, V. L. (2016). Leading with their hearts? How gender stereotypes of emotion lead to biased evaluations of female leaders. *The Leadership Quarterly*, *27*(3), 415–428. https://doi.org/10.1016/j.leaqua.2016.02.005

Brittain, B. (2025, May 21). Google, AI firm must face lawsuit filed by a mother over suicide of son, US court says. *Reuters.* https://www.reuters.com/sustainability/boards-policy-regulation/google-ai-firm-must-face-lawsuit-filed-by-mother-over-suicide-son-us-court-says-2025-05-21/

Brody, L. R. (2013). On understanding gender differences in the expression of emotion : Gender roles, socialization, and language. In *Human Feelings* (1st Edition, pp. 87–121). Routledge. https://doi.org/10.4324/9780203778821-4

Brown, L. M. (2023). Gender, race, and the invisible labor of artificial intelligence. In *Handbook of Critical Studies of Artificial Intelligence* (pp. 573–583). Edward Elgar Publishing. https://doi.org/10.4337/9781803928562.00059

Brundage, M., Mayer, K., Eloundou, T., Agarwal, S., Adler, S., Krueger, G., Leike, J., & Mishkin, P. (n.d.). *Lessons learned on language model safety and misuse* [Video]. Retrieved 8 August 2023, from https://openai.com/research/language-model-safety-and-misuse

Bryant, L., & Garnham, B. (2015). The fallen hero: masculinity, shame and farmer suicide in Australia. *Gender, Place and Culture: A Journal of Feminist Geography*, *22*(1), 67–82. https://doi.org/10.1080/0966369X.2013.855628

Butler, J. (2021, October 23). *Why is the idea of 'gender' provoking backlash the world over?* The Guardian. https://web.archive.org/web/20211117074438/https://amp.theguardian.com/us-news/commentisfree/2021/oct/23/judith-butler-gender-ideology-backlash

Butler, J. P. (1990). *Gender Trouble*. Routledge.

Čapek, J. (2023). Politically correct and gender-neutral language: effects, consequences, acceptance. *ACNS Conference Series: Social Sciences and Humanities*, *3*(01002), 01002. https://doi.org/10.55056/cs-ssh/3/01002

Cardwell, H. (2021, September 8). *Shortage of psychologists leaving patients on waitlist for 9 to 12 months*. RNZ. https://www.rnz.co.nz/news/political/451062/shortage-of-psychologists-leaving-patients-on-waitlist-for-9-to-12-months

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., Wang, T., Marks, S., Segerie, C.-R., Carroll, M., Peng, A., Christoffersen, P., Damani, M., Slocum, S., Anwar, U., … Hadfield-Menell, D. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. In *arXiv [cs.AI]*. arXiv. http://arxiv.org/abs/2307.15217

Cefai, S. (2024). Consent-deception: A feminist cultural media theory of commonsense consent. Feminist Theory, 25(3), 471–492. https://doi.org/10.1177/14647001231206026

Chan, J., Collins, K. A., Lee, R., Linton, J., Cherba, M., Christianson, T.-L. D., Shawanda, A., Siden, E. G., & Wardman, M. (2025). A scoping review of published literature on the linguistic representation of Indigenous Peoples. *Journal of Language and Social Psychology*, *44*(3–4), 441–480. https://doi.org/10.1177/0261927X251318040

Chaney, L. (2024). Treatment of postpartum psychosis in breastfeeding females. *The Mental Health Clinician*, *14*(5), 277–279. https://doi.org/10.9740/mhc.2024.10.277

Chang, H. Y., Jung, C. K., Woo, J. I., Lee, S., Cho, J., Kim, S. W., & Kwak, T.-Y. (2019). Artificial intelligence in pathology. *Journal of Pathology and Translational Medicine*, *53*(1), 1–12. https://doi.org/10.4132/jptm.2018.12.16

Chaplin, T. M., & Aldao, A. (2013). Gender differences in emotion expression in children: a meta-analytic review. *Psychological Bulletin*, *139*(4), 735–765. https://doi.org/10.1037/a0030737

character.ai. (n.d.). [Video]. Retrieved 20 February 2024, from https://character.ai/

Character.ai. (2025). *AI Psychologist* (August 11 version) [Large language model]. https://character.ai/

Chaudhari, S., Aggarwal, P., Murahari, V., Rajpurohit, T., Kalyan, A., Narasimhan, K., Deshpande, A., & da Silva, B. C. (2024). RLHF deciphered: A critical analysis of reinforcement learning from human feedback for LLMs. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/2404.08555

Cheon, B. K., Melani, I., & Hong, Y.-Y. (2020). How USA-centric is psychology? An archival study of implicit assumptions of generalizability of findings to human nature based on origins of study samples. *Social Psychological and Personality Science*, *11*(7), 928–937. https://doi.org/10.1177/1948550620927269

Ching, B. H.-H., & Chen, T. T. (2022). Effects of biological determinism on beliefs and attitudes about transgender people: Psychological essentialism and biased assimilation. *Archives of Sexual Behavior*, *51*(4), 1927–1942. https://doi.org/10.1007/s10508-021-02262-8

Christiansen, D. M., McCarthy, M. M., & Seeman, M. V. (2022). Where sex meets gender: How sex and gender come together to cause sex differences in mental illness. *Frontiers in Psychiatry*, *13*, 856436. https://doi.org/10.3389/fpsyt.2022.856436

Christopher, J. C., Wendt, D. C., Marecek, J., & Goodman, D. M. (2014). Critical Cultural Awareness. *The American Psychologist*, *69*(7), 645–655. https://doi.org/10.1037/a0036851

Cohen, P. N., & Huffman, M. L. (2003). Individuals, jobs, and labor markets: The devaluation of women's work. *American Sociological Review*, *68*(3), 443–463. https://doi.org/10.1177/000312240306800307

Collie, P., Liu, J., Podsiadlowski, A., & Kindon, S. (2010). You can't clap with one hand: Learnings to promote culturally grounded participatory action research with migrant and former refugee communities. *International Journal of Intercultural Relations: IJIR*, *34*(2), 141–149. https://doi.org/10.1016/j.ijintrel.2009.11.008

Collins, A. C., Lekkas, D., Heinz, M. V., Amor, J., Ruan, F., & Jacobson, N. C. (2024). *ChatGPT as therapy: A qualitative and network-based thematic profiling of shared experiences, attitudes, and beliefs on Reddit*. https://doi.org/10.31219/osf.io/57q8x

Connor, G., Coombes, L., & Morgan, M. (2015). iAnorexic: Haraway's cyborg metaphor as ethical methodology. *Qualitative Research in Psychology*, *12*(3), 233–245. https://doi.org/10.1080/14780887.2015.1008901

Cook, M., Pennay, A., MacLean, S., Caluzzi, G., Riordan, B., Cooklin, A., Torney, A., & Callinan, S. (2025). Gender differences in alcohol research: A focus on how men and women are studied in Australia and aotearoa New Zealand. *Drug and Alcohol Review*, *44*(5), 1304–1307. https://doi.org/10.1111/dar.14083

Costa, P., & Ribas, L. (2019). AI becomes her: Discussing gender and artificial intelligence. *Technoetic Arts: A Journal of Speculative Research*, *17*(1), 171–193. https://doi.org/10.1386/tear_00014_1

Crenshaw, K. (1991). Mapping the margins: Intersectionality, identity politics, and violence against women of color. *Stanford Law Review*, *43*(6), 1241. https://doi.org/10.2307/1229039

Daley, A., Beresford, P., & Costa, L. (Eds). (2019). 2. A personal account of mental distress in motherhood. In *Madness, Violence, and Power* (pp. 34–38). University of Toronto Press. https://doi.org/10.3138/9781442629981-006

Daum, C. W. (2020). Social equity, homonormativity, and equality: An intersectional critique of the administration of marriage equality and opportunities for LGBTQ social justice. *Administrative Theory & Praxis*, *42*(2), 115–132. https://doi.org/10.1080/10841806.2019.1659044

De Beauvoir, S. (2011). *The Second Sex* (C. Borde & S. Malovany-Chevallier, Trans.). Random House.

DeKeseredy, W. S., & Dragiewicz, M. (2014). Woman abuse in Canada: sociological reflections on the past, suggestions for the future: Sociological reflections on the

past, suggestions for the future. *Violence against Women*, *20*(2), 228–244. https://doi.org/10.1177/1077801214521325

Dev, S., Monajatipoor, M., Ovalle, A., Subramonian, A., Phillips, J. M., & Chang, K.-W. (2021). Harms of gender exclusivity and challenges in non-binary representation in language technologies. *arXiv preprint* arXiv:2108.12084

Diamond, L. M. (2020). Gender fluidity and nonbinary gender identities among children and adolescents. *Child Development Perspectives*, *14*(2), 110–115. https://doi.org/10.1111/cdep.12366

Dickson, M. (2023). Breaking through the glass ceiling: Experiences of academic women who have advanced to leadership roles in tertiary education in New Zealand. *Journal of Educational Leadership, Policy and Practice*, *37*(1), 59–75. https://doi.org/10.2478/jelpp-2023-0004

Dmowska, A., Fielding-Singh, P., Halpern, J., & Prata, N. (2024). The intersection of traumatic childbirth and obstetric racism: A qualitative study. *Birth (Berkeley, Calif.)*, *51*(1), 209–217. https://doi.org/10.1111/birt.12774

Dobash, R. P., Dobash, R. E., Cavanagh, K., & Lewis, R. (1998). Separate and intersecting realities. *Violence against Women*, *4*(4), 382–414. https://doi.org/10.1177/1077801298004004002

Doocey, M. (2024, September 24). *Workforce boost in specialist mental health training welcomed*. The Beehive. https://www.beehive.govt.nz/release/workforce-boost-specialist-mental-health-training-welcomed

Dosovitsky, G., Pineda, B. S., Jacobson, N. C., Chang, C., Escoredo, M., & Bunge, E. L. (2020). Artificial intelligence chatbot for depression: Descriptive study of usage. *JMIR Formative Research*, *4*(11), e17065. https://doi.org/10.2196/17065

Dotson, K. (2011). Tracking epistemic violence, tracking practices of silencing. *Hypatia*, *26*(2), 236–257. https://doi.org/10.1111/j.1527-2001.2011.01177.x

Drage, E., McInerney, K., & Browne, J. (2024). Engineers on responsibility: feminist approaches to who's responsible for ethical AI. *Ethics and Information Technology*, *26*(1). https://doi.org/10.1007/s10676-023-09739-1

Duggan, L. (2020). The new homonormativity: The sexual politics of neoliberalism. In *Materializing Democracy* (pp. 175–194). Duke University Press. https://doi.org/10.1515/9780822383901-008

Dular, N. (2021). Mansplaining as Epistemic Injustice. *Feminist Philosophy Quarterly*, *7*(1). https://doi.org/10.5206/fpq/2021.1.8482

Dung, L. (2022). Why the epistemic objection against using sentience as criterion of moral status is flawed. *Science and Engineering Ethics*, *28*(6), 51. https://doi.org/10.1007/s11948-022-00408-y

Dunkel Schetter, C., & Tanner, L. (2012). Anxiety, depression and stress in pregnancy. *Current Opinion in Psychiatry*, *25*(2), 141–148. https://doi.org/10.1097/yco.0b013e3283503680

Dupré, M. H. (2025, June 28). *People Are Being Involuntarily Committed, Jailed After Spiraling Into 'ChatGPT Psychosis'*. Futurism. https://futurism.com/commitment-jail-chatgpt-psychosis

Eagly, A. H., & Steffen, V. J. (1986). Gender and aggressive behavior: a meta-analytic review of the social psychological literature. *Psychological Bulletin*, *100*(3), 309–330. https://doi.org/10.1037/0033-2909.100.3.309

Eagly, A, H., Eaton, A., Rose, S. M., Riger, S., & McHugh, M. C. (2012). Feminism and psychology: analysis of a half-century of research on women and gender. *The American Psychologist*, *67*(3), 211–230. https://doi.org/10.1037/a0027260

Eagly, A. H., & Miller, D. I. (2016). Scientific eminence: Where are the women?: Where are the women? *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *11*(6), 899–904. https://doi.org/10.1177/1745691616663918

Eagly, A. H., & Riger, S. (2014). Feminism and psychology: Critiques of methods and epistemology. *The American Psychologist*, *69*(7), 685–702. https://doi.org/10.1037/a0037372

Edwards, R., & Mauthner, M. (2002). Ethics and feminist research: Theory and practice. In M. Mauthner, M. Birch, J. Jessop, & T. Miller (Eds), *Ethics in qualitative research* (pp. 14–28). Sage.

Ehrlich, S., & King, R. (1994). Feminist meanings and the (de)politicization of the lexicon. *Language in Society*, *23*(1), 59–76. https://doi.org/10.1017/s004740450001767x

Elers, C., Jayan, P., Elers, P., & Dutta, M. J. (2021). Negotiating health amidst COVID-19 lockdown in low-income communities in Aotearoa New Zealand. *Health Communication*, *36*(1), 109–115. https://doi.org/10.1080/10410236.2020.1848082

Elliott, R., Bohart, A. C., Watson, J. C., & Murphy, D. (2018). Therapist empathy and client outcome: An updated meta-analysis. *Psychotherapy*. https://psycnet.apa.org/journals/pst/55/4/399/

Ellis, S. A., Wojnar, D. M., & Pettinato, M. (2015). Conception, pregnancy, and birth experiences of male and gender variant gestational parents: it's how we could have a family. *Journal of Midwifery & Women's Health*, *60*(1), 62–69. https://doi.org/10.1111/jmwh.12213

Epstein, D., & Goodman, L. (2019). Discounting women: Doubting domestic violence survivors' credibility and dismissing their experiences. *U. Pa. L. Rev.* https://scholarship.law.upenn.edu/cgi/viewcontent.cgi?article=9644&context=penn_law_review

Erickson, R. J. (2005). Why emotion work matters: sex, gender, and the division of household labor. *Journal of Marriage and the Family*, *67*(2), 337–351. https://doi.org/10.1111/j.0022-2445.2005.00120.x

Espejo, G., Reiner, W., & Wenzinger, M. (2023). Exploring the role of artificial intelligence in mental healthcare: Progress, pitfalls, and promises. *Cureus*, *15*(9), e44748. https://doi.org/10.7759/cureus.44748

Every-Palmer, S., Grant, M. L., Thabrew, H., Hansby, O., Lawrence, M., Jenkins, M., & Romans, S. (2024). Not heading in the right direction: Five hundred psychiatrists' views on resourcing, demand, and workforce across New Zealand mental health services. *The Australian and New Zealand Journal of Psychiatry*, *58*(1), 82–91. https://doi.org/10.1177/00048674231170572

Exner-Cortens, D., Wright, A., Claussen, C., & Truscott, E. (2021). A systematic review of adolescent masculinities and associations with internalizing behavior problems and social support. *American Journal of Community Psychology*, *68*(1–2), 215–231. https://doi.org/10.1002/ajcp.12492

Fa'alogo-Lilo, C., & Cartwright, C. (2021). Barriers and supports experienced by Pacific peoples in aotearoa New Zealand's mental health services. *Journal of Cross-Cultural Psychology*, *52*(8–9), 752–770. https://doi.org/10.1177/00220221211039885

Fahlgren, M. K., Cheung, J. C., Ciesinski, N. K., McCloskey, M. S., & Coccaro, E. F. (2022). Gender differences in the relationship between anger and aggressive behaviour. *Journal of Interpersonal Violence*, *37*(13–14), NP12661–NP12670. https://doi.org/10.1177/0886260521991870

Fanslow, J. L., Malihi, Z., Hashemi, L., Gulliver, P., & McIntosh, T. (2022). Prevalence of interpersonal violence against women and men in New Zealand: results of a cross-sectional study. *Australian and New Zealand Journal of Public Health*, *46*(2), 117–126. https://doi.org/10.1111/1753-6405.13206

Fanslow, J. L., Mellar, B. M., Gulliver, P. J., & McIntosh, T. K. D. (2023). Ethnic-specific prevalence rates of intimate partner violence against women in New Zealand. *Australian and New Zealand Journal of Public Health*, *47*(6), 100105. https://doi.org/10.1016/j.anzjph.2023.100105

Fanslow, JL, McIntosh, T. (2023). Key findings and policy and practice implications from He Koiora Matapopore | The 2019 New Zealand Family Violence Study. University of Auckland: Auckland New Zealand.

Fine, M. (2017). Circulating narratives: Theorizing narrative travel translation and provocation. *Psychology in Society*, *55*. https://doi.org/10.17159/2309-8708/2017/n55a7

Fitzpatrick, K. K., Vierhile, M., & Darcy, A. (2017). Delivering cognitive behavior therapy to young adults with symptoms of depression and anxiety using a fully automated conversational agent (Woebot): A randomized controlled trial. JMIR Mental Health, 4(2). https://doi.org/10.2196/mental.7785

Fleming, P. J., Gruskin, S., Rojo, F., & Dworkin, S. L. (2015). Men's violence against women and men are inter-related: Recommendations for simultaneous intervention. *Social Science & Medicine (1982)*, *146*, 249–256. https://doi.org/10.1016/j.socscimed.2015.10.021

Flores-Robles, G., & Gantman, A. P. (2024). Notions of care labor are antithetical to profitable labor. *Psychology of Women Quarterly*, *48*(4), 475–490. https://doi.org/10.1177/03616843241248366

Fong, A., Boxley, C. L., Krevat, S., Mutondo, E. K., & Thomas, A. D. (2024). The maternal voice: Exploration of mothers and birthing individuals' voices in patient safety event and feedback reports. *Women's Health Reports (New Rochelle, N.Y.)*, *5*(1), 727–734. https://doi.org/10.1089/whr.2024.0020

Formanowicz, M., Bedynska, S., Cisłak, A., Braun, F., & Sczesny, S. (2013). Side effects of gender-fair language: How feminine job titles influence the evaluation of female applicants. *European Journal of Social Psychology*, *43*(1), 62–71. https://doi.org/10.1002/ejsp.1924

Fors, M. (2021). Power dynamics in the clinical situation: A confluence of perspectives. *Contemporary Psychoanalysis*, *57*(2), 242–269. https://doi.org/10.1080/00107530.2021.1935191

Gatens, M. (1995). *Imaginary Bodies: Ethics, power and corporeality*. Routledge. https://doi.org/10.4324/9780203418659

Gemmell, Monique. A History of Marginalisation: Maori Women. Te Herenga Waka—Victoria University of Wellington, 2013.

George, M., Mulvale, S., Davidson, T., & Rutherford, A. (2020). Disrupting Androcentrism in Social Psychology Textbooks: A Call for Critical Reflexivity. *ResearchGate*, *1*(1), 15.

Ging, D., & Siapera, E. (2018). Special issue on online misogyny. *Feminist Media Studies*, *18*(4), 515–524. https://doi.org/10.1080/14680777.2018.1447345

Gonzalez L. O. (2000). Infertility as a transformational process: a framework for psychotherapeutic support of infertile women. Issues in mental health nursing, 21(6), 619–633. https://doi-org/10.1080/01612840050110317

Goro, S. F., & Harahap, C. B. (2025). Chatbot Dituntut. *Dimensia: Jurnal Kajian Sosiologi*, *14*(2), 12–20. https://doi.org/10.21831/dimensia.v14i2.81003

Greenfield, P., Calcia, M., McCree, C., Sahota, M., Thomas, H., Kirkpatrick, K., Vagi, R., Howard, L. M., Markham, S., & Bhavsar, V. (2025). Identifying, assessing and responding to perpetration of domestic abuse: practice guide for mental health professionals. *BJPsych Advances*, *31*(1), 8–19. https://doi.org/10.1192/bja.2024.39

Guendouzi, J. (2006). “The guilt thing”: Balancing domestic and professional roles. *Journal of Marriage and the Family*, *68*(4), 901–909. https://doi.org/10.1111/j.1741-3737.2006.00303.x

Guo, T. (2015). Alan Turing: Artificial intelligence as human self-knowledge. *Anthropology Today*, *31*(6), 3–7. https://doi.org/10.1111/1467-8322.12209

Gupta, B., Mufti, T., Sohail, S. S., & Madsen, D. Ø. (2023). ChatGPT: A brief narrative review. *Cogent Business & Management*, *10*(3). https://doi.org/10.1080/23311975.2023.2275851

Hall, R. L., & Fine, M. (2005). The Stories We Tell: The Lives and Friendship of Two Older Black Lesbians. *Psychology of Women Quarterly*, *29*(2), 177–187. https://doi.org/10.1111/j.1471-6402.2005.00180.x

Hamilton, M. C. (1991). Masculine bias in the attribution of personhood: People = male, male = people. *Psychology of Women Quarterly*, *15*(3), 393–402. https://doi.org/10.1111/j.1471-6402.1991.tb00415.x

Haraway, D. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, *14*(3), 575–599.

Haraway, D. (2006). A Cyborg Manifesto: Science, technology, and socialist-feminism in the late 20th century. In J. Weiss, J. Nolan, J. Hunsinger, & P. Trifonas (Eds), *The International Handbook of Virtual Learning Environments* (pp. 117–158). Springer Netherlands. https://doi.org/10.1007/978-1-4020-3803-7_4

Haraway, D. (2016). *Staying with the trouble: Making kin in the Chthulucene*. Duke University Press.

Haraway, D. (2008). *When species meet*. University of Minnesota Press.

Harding, S. (1991). *Whose science? Whose knowledge?: Thinking from women's lives*. Cornell University Press.

Harding, S. (2001). Feminist standpoint epistemology. *The Gender and Science Reader*, 145–168. https://books.google.co.nz/books?hl=en&lr=&id=9obFtmhcCNsC&oi=fnd&pg=PA145&dq=Sandra+Harding+(1991)+standpoint+theory&ots=6eteJnQEYB&sig=MiaNiwosgLUMv8TZKnclJMY7XIs

Harding, S. (2007). Feminist standpoints. *Handbook of Feminist Research: Theory and Praxis*, 45–69.

https://books.google.co.nz/books?hl=en&lr=&id=dEJ1AwAAQBAJ&oi=fnd&pg=PA46&dq=Sandra+Harding+(1991)+standpoint+theory&ots=cRxjl8T5wG&sig=i3V2TDHoAEeo64LCmfF4K5H2cz0

Hare-Mustin, R. T., & Marecek, J. (2018). Gender and the meaning of difference. In *Theorizing Feminism* (pp. 78–109). Routledge. https://doi.org/10.4324/9780429494277-6

Hattotuwa, S., Hannah, K., & Taylor, K. (2023). *Transgressive transitions: Transphobia, community building, bridging, and bonding within Aotearoa New Zealand's disinformation ecologies March-April 2023*. The Disinformation Project. https://thedisinfoproject.org/wp-content/uploads/2023/05/Transgressive-Transitions.pdf

Haugsbaken, H., & Hagelia, M. (2024). A new AI literacy for the algorithmic age: Prompt engineering or eductional promptization? *2024 4th International Conference on Applied Artificial Intelligence (ICAPAI)*, 1–8. https://doi.org/10.1109/icapai61893.2024.10541229

Healthpoint. (2023). *Mental health services waiting times report*. Ministry of Health.

Heather, C., O'Sullivan, D., Kidd, J., & McCreanor, T. (2020). The Waitangi Tribunal's WAI 2575 Report: Implications for Decolonizing Health Systems. *Health and Human Rights*, *22*(1), 209–220. Directory of Open Access Journals. http://ezproxy.massey.ac.nz/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edsdoj&AN=edsdoj.2a4c98bf890a413e84f52d0c899bc8c7&site=eds-live&scope=site

Heinz, M. V., Mackin, D., Trudeau, B., Bhattacharya, S., Wang, Y., Banta, H. A., Jewett, A. D., Salzhauer, A., Griffin, T., & Jacobson, N. C. (2024). Evaluating Therabot: A randomized control trial investigating the feasibility and effectiveness of a generative AI therapy chatbot for depression, anxiety, and eating disorder symptom treatment. In *PsyArXiv*. https://doi.org/10.31234/osf.io/pjqmr

Henrich, J., Heine, S. J., & Norenzayan, A. (2010a). The weirdest people in the world? *The Behavioral and Brain Sciences*, *33*(2–3), 61–83; discussion 83-135. https://doi.org/10.1017/S0140525X0999152X

Henrich, J., Heine, S. J., & Norenzayan, A. (2010b). Most people are not WEIRD. *Nature*, *466*(7302), 29. https://doi.org/10.1038/466029a

Hibbs, C. (2014). Androcentrism. In *Encyclopedia of Critical Psychology* (pp. 94–101). Springer New York. https://doi.org/10.1007/978-1-4614-5583-7_16

Hill, A. L., Keil, M. A., Chang, J. C., Krans, E. E., Kim, E., Van Nostrand, E., Miller, E., & Pallatino, C. (2024). Help-seeking among pregnant and postpartum women with lifetime experiences of opioid use disorder and intimate partner violence. *Violence against Women*, *30*(3–4), 812–831. https://doi.org/10.1177/10778012221140134

Hines, M. (2011). Gender development and the human brain. *Annual Review of Neuroscience*, *34*(1), 69–88. https://doi.org/10.1146/annurev-neuro-061010-113654

Hooks, B. (1985). *Feminist theory from margin to center*. South End Press.

Hoppstadius, H. (2018). What is the problem? Representations of men's violence against women in a Swedish context. *Multidisciplinary Journal of Gender Studies*, *7*(3), 1684. https://doi.org/10.17583/generos.2018.3737

Howard, L. M. (2017). Routine enquiry about violence and abuse is needed for all mental health patients [Review of *Routine enquiry about violence and abuse is needed for all mental health patients*]. *The British Journal of Psychiatry: The Journal of Mental Science*, *210*(4), 298. https://doi.org/10.1192/bjp.210.4.298

Hyde, J. S., Bigler, R. S., Joel, D., Tate, C. C., & van Anders, S. M. (2019). The future of sex and gender in psychology: Five challenges to the gender binary. *The American Psychologist*, *74*(2), 171–193. https://doi.org/10.1037/amp0000307

IEA. (2024). *Electricity 2024*. IEA. https://www.iea.org/reports/electricity-2024

Ifechelobi, J. N. (2014). Feminism: Silence and voicelessness as tools of patriarchy in Chimamanda adichie's *purple hibiscus*. *African Research Review*, *8*(4), 17. https://doi.org/10.4314/afrrev.v8i4.2

Ingham, T. R., Jones, B., Perry, M., King, P. T., Baker, G., Hickey, H., Pouwhare, R., & Nikora, L. W. (2022). The multidimensional impacts of inequities for tāngata whaikaha Māori (Indigenous Māori with lived experience of disability) in aotearoa, New Zealand. *International Journal of Environmental Research and Public Health*, *19*(20), 13558. https://doi.org/10.3390/ijerph192013558

Jackson, S. (2006). Interchanges: Gender, sexuality and heterosexuality: The complexity (and limits) of heteronormativity. *Feminist Theory*, *7*(1), 105–121. https://doi.org/10.1177/1464700106061462

Jones, P. (2010). Roosters, hawks and dawgs: Toward an inclusive, embodied eco/feminist psychology. *Feminism & Psychology*, *20*(3), 365–380. https://doi.org/10.1177/0959353510368120

Jordan, J. (2004). Beyond belief? *Criminal Justice*, *4*(1), 29–59. https://doi.org/10.1177/1466802504042222

Joseph, A. P., & Babu, A. (2024). Transference and the psychological interplay in AI-enhanced mental healthcare. *Frontiers in Psychiatry*, *15*, 1460469. https://doi.org/10.3389/fpsyt.2024.1460469

Juillerat, T., White, K., & Obst, P. (2023). A theory-based examination of the predictors of mental health help-seeking in young men. *Australian Psychologist*, *58*(6), 466–482. https://doi.org/10.1080/00050067.2023.2231612

Jung, J.-Y., & Bozzon, A. (2023, April 23). Are female chatbots more empathic? - discussing gendered conversational agent through empathic design. *Proceedings of the 2nd Empathy-Centric Design Workshop*. EMPATHICH '23: EmpathiCH Workshop, Hamburg Germany. https://doi.org/10.1145/3588967.3588970

Kafka, J. M., Moracco, K. B. E., Taheri, C., Young, B.-R., Graham, L. M., Macy, R. J., & Proescholdbell, S. (2022). Intimate partner violence victimization and perpetration as precursors to suicide. *SSM - Population Health*, *18*(101079), 101079. https://doi.org/10.1016/j.ssmph.2022.101079

Kagan, C., & Burton, M. (2001, March 1). *Critical Community Psychology Praxis for the 21st Centtury*. Paper presented at British Psychological Society Conference.

Kannampallil, T., Ajilore, O. A., Lv, N., Smyth, J. M., Wittels, N. E., Ronneberg, C. R., Kumar, V., Xiao, L., Dosala, S., Barve, A., Zhang, A., Tan, K. C., Cao, K. P., Patel, C. R., Gerber, B. S., Johnson, J. A., Kringle, E. A., & Ma, J. (2023). Effects of a virtual voice-based coach delivering problem-solving treatment on emotional distress and brain function: a pilot RCT in depression and anxiety. *Translational Psychiatry*, *13*(1), 1–8. https://doi.org/10.1038/s41398-023-02462-x

Kaplan, G. A., Shema, S. J., & Leite, C. M. A. (2008). Socioeconomic determinants of psychological well-being: The role of income, income change, and income sources during the course of 29 years. *Annals of Epidemiology*, *18*(7), 531–537. https://doi.org/10.1016/j.annepidem.2008.03.006

Kaufman, E. A., Meddaoui, B., Seymour, N. E., & Victor, S. E. (2023). The roles of minority stress and thwarted belongingness in suicidal ideation among cisgender and transgender/nonbinary LGBTQ+ individuals. *Archives of Suicide Research: Official Journal of the International Academy for Suicide Research*, *27*(4), 1296–1311. https://doi.org/10.1080/13811118.2022.2127385

Kawakami, A., Taylor, J., Fox, S., Zhu, H., & Holstein, K. (2024). AI Failure Loops in Feminized Labor: Understanding the Interplay of Workplace AI and Occupational Devaluation. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, *7*(1), 683–683. https://doi.org/10.1609/aies.v7i1.31670

Kercher, A., & Gossage, L. (2024). Identifying risk factors for compassion fatigue in psychologists in Aotearoa, New Zealand, during the COVID-19 pandemic.

Professional Psychology, Research and Practice, *55*(1), 28–38. https://doi.org/10.1037/pro0000538

Kercher, A., Rahman, J., & Pedersen, M. (2024). The COVID-19 pandemic, psychologists' professional quality of life and mental health. *Frontiers in Psychology*, *15*, 1339869. https://doi.org/10.3389/fpsyg.2024.1339869

Khawaja, Z., & Bélisle-Pipon, J.-C. (2023). Your robot therapist is not your therapist: understanding the role of AI-powered mental health chatbots. *Frontiers in Digital Health*, *5*, 1278186. https://doi.org/10.3389/fdgth.2023.1278186

Kim, C. Y., Lee, C. P., & Mutlu, B. (2024). Understanding large-language model (LLM)-powered human-robot interaction. *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, *2799*, 371–380. https://doi.org/10.1145/3610977.3634966

Kingston, A. K. (2020). Feminist Research Ethics: From Theory to Practice. In R. Iphofen (Ed.), *Handbook of Research Ethics and Scientific Integrity* (pp. 531–549). Springer International Publishing. https://doi.org/10.1007/978-3-030-16759-2_64

Klein, L., & D'Ignazio, C. (2024). Data Feminism for AI. *Conference on Fairness, Accountability and Transparency*. https://doi.org/10.1145/3630106.3658543

Knoth, N., Tolzin, A., Janson, A., & Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence*, *6*(100225), 100225. https://doi.org/10.1016/j.caeai.2024.100225

Kobylski, L. A., Arakelian, M. H., Freeman, M. P., Gaw, M. L., Cohen, L. S., & Vanderkruik, R. (2024). Barriers to care and treatment experiences among individuals with postpartum psychosis. *Archives of Women's Mental Health*, *27*(4), 637–647. https://doi.org/10.1007/s00737-024-01447-z

Koubaa, A. (2023). *GPT-4 vs. GPT-3.5: A Concise Showdown*. ENGINEERING. https://doi.org/10.20944/preprints202303.0422.v1

Lee, E. E., Torous, J., De Choudhury, M., Depp, C. A., Graham, S. A., Kim, H.-C., Paulus, M. P., Krystal, J. H., & Jeste, D. V. (2021). Artificial intelligence for mental health care: Clinical applications, barriers, facilitators, and artificial wisdom. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *6*(9), 856–864. https://doi.org/10.1016/j.bpsc.2021.02.001

Lee, J.-S. (2024). InstructPatentGPT: training patent language models to follow instructions with human feedback. *Artificial Intelligence and Law*. https://doi.org/10.1007/s10506-024-09401-1

Lee, O., & Joseph, K. (2025). A large-scale analysis of public-facing, community-built chatbots on Character.AI. *arXiv [cs.SI]*. arXiv. arXiv.org/2505.13354

Lelutiu-Weinberger, C., Clark, K. A., & Pachankis, J. E. (2023). Mental health provider training to improve LGBTQ competence and reduce implicit and explicit bias: A randomized controlled trial of online and in-person delivery. *Psychology of Sexual Orientation and Gender Diversity*, *10*(4), 589–599. https://doi.org/10.1037/sgd0000560

Levant, R. F., Hall, R. J., & Rankin, T. J. (2013). Male Role Norms Inventory-Short Form (MRNI-SF): development, confirmatory factor analytic investigation of structure, and measurement invariance across gender. *Journal of Counseling Psychology*, *60*(2), 228–238. https://doi.org/10.1037/a0031545

Levy, M. (2015). Our voices, our future: Indigenous psychology in Aotearoa New Zealand. In W. Waitoki, J.S. Feather, N.R. Robertson, & J.J. Rucklidge (Eds),. *Professional Practice of Psychology in Aotearoa New Zealand (3rd Ed.), (Pp 30-36)*.

Lewis, M., Haviland-Jones, J. M., Sloan, D. M., & Fresco, D. M. (2001). Handbook of emotions (second edition). *Journal of Cognitive Psychotherapy*, *15*(3), 281–283. https://doi.org/10.1891/0889-8391.15.3.281

Li, Y., Mughees, M., Chen, Y., & Li, Y. R. (2024). The unseen AI disruptions for power grids: LLM-induced transients. In *arXiv [cs.AR]*. arXiv. http://arxiv.org/abs/2409.11416

Liben, L. S., Bigler, R. S., Ruble, D. N., Martin, C. L., & Powlishta, K. K. (2002). *The developmental course of gender differentiation: Conceptualizing, measuring, and evaluating constructs and pathways. Monographs of the society for research in child development*.

Lila, M., Gracia, E., & Murgui, S. (2013). Psychological adjustment and victim-blaming among intimate partner violence offenders: The role of social support and stressful life events. *The European Journal of Psychology Applied to Legal Context*, *5*(2), 147–153. https://doi.org/10.5093/ejpalc2013a4

Lindsay, N., Haami, D., Tassell-Matamua, N., Pomare, P., Valentine, H., Pahina, J., Ware, F., & Pidduck, P. (2020). The spiritual experiences of contemporary Māori in Aotearoa New Zealand: A qualitative analysis. *Journal of Spirituality in Mental Health*, 1–21. https://doi.org/10.1080/19349637.2020.1825152

Liss, M., Richmond, K., & Erchull, M. J. (2019). Power and privilege. *In Psychology of women and gender* (pp. 84–89). W. W. Norton & Company. https://nerd.wwnorton.com/ebooks/epub/psychwomen2/EPUB/content/2.4-chapter02.xhtml

Little, A. (2022, Summer 8). *Govt's mental health roll-out gains momentum – more funds and internships for clinical psychology*. The Beehive. https://www.beehive.govt.nz/release/govt%E2%80%99s-mental-health-roll-out-gains-momentum-%E2%80%93-more-funds-and-internships-clinical

Liu, T., Zhang, Y., Zhao, Z., Dong, Y., Meng, G., & Chen, K. (2024). Making them ask and answer: Jailbreaking large language models in few queries via disguise and reconstruction. In *arXiv [cs.CR]*. arXiv. https://doi.org/10.48550/arXiv.2402.18104

Logoz, F., Eggenberger, L., Schneeberger, M., & Walther, A. (2024). Psychotherapists' endorsement of traditional masculinity ideologies and their assessment of gender differences in the treatment of depressive disorders. In *PsyArXiv*. https://doi.org/10.31234/osf.io/h3gcy

Loideain, N. N., & Adams, R. (2020). From Alexa to Siri and the GDPR: The gendering of virtual personal assistants and the role of data protection impact assessments. *Computer Law and Security Report*, *36*, 105366. https://doi.org/10.1016/j.clsr.2019.105366

Love, C. (2008). An Indigenous Reality Check: Comments on Ian Evans “Steering by Matariki and the Southern Cross: Plotting Clinical Psychology’s Course in New Zealand.” New Zealand Journal of Psychology, 37(3), 26–32.

Lund, R. W. B. (2023). Rethinking knowledge production through standpoint, decolonisation and intersectionality: thinking with Sandra Harding. In *Handbook of Feminist Research Methodologies in Management and Organization Studies* (pp. 25–38). Edward Elgar Publishing. https://doi.org/10.4337/9781800377035.00009

Lupton, D. (2014). Apps as Artefacts: Towards a Critical Perspective on Mobile Health and Medical Apps. Societies, 4(4), 606–622. https://doi.org/10.3390/soc4040606

Ma, N., Chen, S., Kong, Y., Chen, Z., Geldsetzer, P., Zeng, H., Wu, L., Wehrmeister, F. C., Lu, C., Subramanian, S. V., Song, Y., & Li, Z. (2023). Prevalence and changes of intimate partner violence against women aged 15 to 49 years in 53 low-income and middle-income countries from 2000 to 2021: a secondary analysis of population-based surveys. *The Lancet. Global Health*, *11*(12), e1863–e1873. https://doi.org/10.1016/S2214-109X(23)00417-5

Magnusson, E., & Marecek, J. (2017). Feminisms, psychologies, and the study of social life. In *The Palgrave Handbook of Critical Social Psychology* (pp. 17–35). Palgrave Macmillan UK. https://doi.org/10.1057/978-1-137-51018-1_2

Maher, P. J., Roth, J., Griffin, S., Foran, A. M., Jay, S., McHugh, C., Ryan, M., Bradshaw, D., Quayle, M., & Muldoon, O. T. (2023). Pandemic threat and group cohesion: national identification in the wake of COVID-19 is associated with authoritarianism. *The Journal of Social Psychology*, *163*(6), 789–805. https://doi.org/10.1080/00224545.2021.2024122

Mahuika, R. (2008). Kaupapa Māori theory is critical and anti-colonial. *MAI Review*, *3*(4).

Maine, A. (2021). Queering marriage: The homoradical and anti-normativity. *Laws*, *11*(1), 1. https://doi.org/10.3390/laws11010001

Malle, B. F., Magar, S. T., & Scheutz, M. (2019). AI in the sky: How people morally evaluate human and machine decisions in a lethal strike dilemma. In *Robotics and Well-Being* (pp. 111–133). Springer International Publishing. https://doi.org/10.1007/978-3-030-12524-0_11

Malmqvist, L. (2024). Sycophancy in Large Language Models: Causes and Mitigations. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2411.15287

Maree Kopua, D., Kopua, M. A., & Bracken, P. J. (2020). Mahi a Atua: A Māori approach to mental health. *Transcultural Psychiatry*, *57*(2), 375–383. https://doi.org/10.1177/1363461519851606

Massey University. (2017). *Code of ethical conduct for research, teaching and evaluations involving human participants revised code 2017*. https://www.massey.ac.nz/massey/fms/Human%20Ethics/Documents/MUHEC%20Code.pdf?2F3CBE296DD2345CC01794BF9CFCA13A

Mathias, S. (2024, July 29). *AI is already straining electricity systems – and we're just at the beginning*. The Spinoff. https://thespinoff.co.nz/internet/29-07-2024/ai-is-already-straining-electricity-systems-and-were-just-at-the-beginning

Matsuno, E., & Budge, S. L. (2017). Non-binary/genderqueer identities: a critical review of the literature. *Current Sexual Health Reports*, *9*(3), 116–120. https://doi.org/10.1007/s11930-017-0111-8

Maurya, R. K., Montesinos, S., Bogomaz, M., & DeDiego, A. C. (2025). Assessing the use of ChatGPT as a psychoeducational tool for mental health practice. *Counselling and Psychotherapy Research*, *25*(1). https://doi.org/10.1002/capr.12759

Maxwell, C., Selvanathan, H. P., Hames, S., Crimston, C. R., & Jetten, J. (2025). A mixed-methods approach to understand victimization discourses by opposing feminist sub-

groups on social media. *The British Journal of Social Psychology*, *64*(1), e12785. https://doi.org/10.1111/bjso.12785

Mental Health Foundation of New Zealand. (n.d.). *Suicide statistics*. Mental Health Foundation of New Zealand. Retrieved 17 May 2025, from https://mentalhealth.org.nz/suicide-prevention/suicide-statistics

Mikaere, A. (1999). Colonisation and the imposition of patriarchy: A Ngāti Raukawa woman's perspective. *Te Ukaipo*, *1*, 34–49.

Mikaere, A. (2011). *Colonising myths -- Māori realities: He rukuruku whakaaro*. Huia Publishers Ltd.

Milgrom, J., & Beatrice, G. (2003). Coping with the stress of motherhood: cognitive and defence style of women with postnatal depression. *Stress and Health: Journal of the International Society for the Investigation of Stress*, *19*(5), 281–287. https://doi.org/10.1002/smi.986

Miller, L., & Hayward, R. (2006). New jobs, old occupational stereotypes: gender and jobs in the new economy. *Journal of Education and Work*, *19*(1), 67–93. https://doi.org/10.1080/13639080500523000

Milne, M. (2005). Māori perspectives on kaupapa Māori and psychology. *A Discussion Document. A Report Prepared for the New Zealand Psychologists Board. Wellington*. https://www.pbanz.org.nz/docs/KAUPAPA%20MAORI%20AND%20PSYCHOLOGY1%20Moe%20Milnes%20Report_doc1.pdf

Ministry of Health. (2021). *Mental health and wellbeing strategy 2021-2026*. New Zealand Government.

Ministry of Health. (2022). *Government response to the Report of the Petitions Committee on Petition 2020/133 of Lucy McLean: Increase the psychologist workforce in New Zealand*. https://www.health.govt.nz/system/files/2022-09/cabinet_paper_government_response_to_petition_of_lucy_mclean_increase_the_psychologist_workforce.pdf

Ministry of Health NZ. (2024, November 19). *Annual Update of Key Results 2023/24: New Zealand Health Survey*. Ministry of Health NZ. https://www.health.govt.nz/publications/annual-update-of-key-results-202324-new-zealand-health-survey

Ministry of Justice. (2002). *Latest Crime Survey Revwals Surprising High Levels of Ureported Sexual Violence*. Ministry of Justice. https://www.justice.govt.nz/about/news-and-media/news/latest-crime-survey-reveals-surprising-high-levels-of-unreported-sexual-violence

Ministry of Justice. (2023). *New Zealand Crime and Victims Survey. Key findings – Cycle 5 report. Descriptive statistics. June 2023. Results drawn from Cycle 5 (2021/22) of the New Zealand Crime and Victims Survey*. Wellington: Ministry of Justice.

Mishra, R., & Varshney, G. (2025). Exploiting jailbreaking vulnerabilities in Generative AI to bypass ethical safeguards for facilitating phishing attacks. In *arXiv [cs.CR]*. arXiv. http://arxiv.org/abs/2507.12185

Mitchell, C. (2022). TERF wars: feminism and the fight for transgender futures. *Community Development Journal*, *57*(3), 573–577. https://doi.org/10.1093/cdj/bsab016

Modebadze, V. (2022). How the Coronavirus pandemic contributed to the rise of authoritarianism throughout the world. *National Security and the Future*, *23*(1), 79–88. https://doi.org/10.37458/nstf.23.1.4

Moeke-Maxwell, T. (2005). Bi/multiracial Māori women's hybridity in Aotearoa/New Zealand. *Discourse Studies in the Cultural Politics of Education*, *26*(4), 497–510. https://doi.org/10.1080/01596300500319779

Moore, J., Grabb, D., Agnew, W., Klyman, K., Chancellor, S., Ong, D. C., & Haber, N. (2025). Expressing stigma and inappropriate responses prevents LLMs from safely replacing mental health providers. In *arXiv [cs.CL]*. arXiv. https://doi.org/10.1145/3715275.3732039

Morgenroth, T., & Ryan, M. K. (2021). The effects of gender trouble: An integrative theoretical framework of the perpetuation and disruption of the gender/sex binary. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *16*(6), 1113–1142. https://doi.org/10.1177/1745691620902442

Mróz, M., Stobnicka, D., Marcewicz, A., Szlendak, B., & Iwanowicz-Palus, G. (2024). Stress and coping strategies among women in late motherhood. *Journal of Clinical Medicine*, *13*(7). https://doi.org/10.3390/jcm13071995

Muehlenhard, C. L., & Peterson, Z. D. (2011). Distinguishing between sex and gender: History, current conceptualizations, and implications. *Sex Roles*, *64*(11–12), 791–803. https://doi.org/10.1007/s11199-011-9932-5

Mulder, R. T., Bastiampillai, T., Jorm, A., & Allison, S. (2022). New Zealand's mental health crisis, He Ara Oranga and the future. *The New Zealand Medical Journal*, *135*(1548), 89–95. https://www.ncbi.nlm.nih.gov/pubmed/35728133

Myyry, S., & Siivonen, P. (2025). Depoliticized and decontextualized equality promotion in the gender equality planning of Finnish comprehensive schools. *Nordic Journal of Studies in Educational Policy*, *11*(2), 161–172. https://doi.org/10.1080/20020317.2024.2403158

Nadeem, A., Abedin, B., & Marjanovic, O. (2020). Gender bias in AI: A review of contributing factors and mitigating strategies. *ACIS 2020 Proceedings*. https://aisel.aisnet.org/acis2020/27?utm_source=aisel.aisnet.org%2Facis2020%2F27&utm_medium=PDF&utm_campaign=PDFCoverPages

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2025). A comprehensive overview of large Language Models. *ACM Transactions on Intelligent Systems and Technology*. https://doi.org/10.1145/3744746

New Zealand Psychological Society. (2002). *Code of ethics for psychologists working in Aotearoa/New Zealand*. https://www.psychology.org.nz/journal-archive/code-of-ethics.pdf

New Zealand Psychologists Board. (2024). *New Zealand Psychologits Board Annual Report 2023/24*. https://psychologistsboard.org.nz/wp-content/uploads/2025/02/New-Zealand-Psychologists-Board-Annual-Report-2024.pdf

Nienhuis, J. B., Owen, J., Valentine, J. C., Winkeljohn Black, S., Halford, T. C., Parazak, S. E., Budge, S., & Hilsenroth, M. (2018). Therapeutic alliance, empathy, and genuineness in individual adult psychotherapy: A meta-analytic review. *Psychotherapy Research: Journal of the Society for Psychotherapy Research*, *28*(4), 593–605. https://doi.org/10.1080/10503307.2016.1204023

Nunn, J. S., & Thomas, S. L. (1999). The angry male and the passive female: The role of gender and self-esteem in anger expression. *Social Behavior and Personality*, *27*(2), 145–153. https://doi.org/10.2224/sbp.1999.27.2.145

NZ College of Clinical Psychologists. (2021, November 5). *Submission to the Petitions Committee Petition of Lucy McLean: Increase the psychologist workforce in New Zealand*. https://cdn.prod.website-files.com/6629c7c5b8a3b236ee45e8f0/669df30c8fc2c5576b87e3d9_Petitions-Committee-Submission-Psychology-Workforce.pdf

Oakley, A. (2016). Interviewing Women Again: Power, Time and the Gift. *Sociology*, *50*(1), 195–213. https://doi.org/10.1177/0038038515580253

Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, *366*(6464), 447–453.

O'Connor, S., & Liu, H. (2023). Gender bias perpetuation and mitigation in AI technologies: challenges and opportunities. *AI & Society*. https://doi.org/10.1007/s00146-023-01675-4

Ojio, Y., Amemiya, R., Oliffe, J. L., & Rice, S. M. (2025). Mental health help-seeking knowledge, attitudes and behaviour among male elite rugby players: the role of masculine health-related values. *BMJ Open Sport & Exercise Medicine*, *11*(1), e002275. https://doi.org/10.1136/bmjsem-2024-002275

Olawade, D. B., Wada, O. Z., Odetayo, A., David-Olawade, A. C., Asaolu, F., & Eberhardt, J. (2024). Enhancing mental health with Artificial Intelligence: Current trends and future prospects. *Journal of Medicine, Surgery, and Public Health*, *3*(100099), 100099. https://doi.org/10.1016/j.glmedi.2024.100099

Oliffe, J. L., & Phillips, M. J. (2008). Men, depression and masculinities: A review and recommendations. *Journal of Men's Health*, *5*(3), 194–202. https://doi.org/10.1016/j.jomh.2008.03.016

Opara, E. (2025). AI Is Not Intelligent. In *Preprints*. https://doi.org/10.20944/preprints202501.1953.v1

Open, A. I. (n.d.). *Sycophancy in GPT-4o: what happened and what we're doing about it*. Retrieved 16 May 2025, from https://openai.com/index/sycophancy-in-gpt-4o/

OpenAI. (n.d.-a). *Our approach to AI safety: Ensuring that AI systems are built, deployed, and used safely is critical to our mission*. https://openai.com/blog/our-approach-to-ai-safety

OpenAI. (n.d.-b). *Pricing* [Video]. https://openai.com/chatgpt/pricing

OpenAI. (n.d.-c). *GPT-4 is OpenAI's most advanced system, producing safer and more useful responses*. https://openai.com/gpt-4

OpenAI. (2022, March 2). *Lessons learned on language model safety and misuse*. https://openai.com/index/language-model-safety-and-misuse/

OpenAI. (2022, November 30). *Introducing ChatGPT*. https://openai.com/blog/chatgpt

OpenAI. (2025, May 2). *Expanding on what we missed with sycophancy*. https://openai.com/index/expanding-on-sycophancy/

Østergaard, S. D. (2025). Generative artificial intelligence chatbots and delusions: From guesswork to emerging cases. *Acta Psychiatrica Scandinavica*, *acps.70022*. https://doi.org/10.1111/acps.70022

Pauwels, A. (2008). *Linguistic sexism and feminist linguistic activism*. 550–570. https://doi.org/10.1002/9780470756942.CH24

Pentina, I., Hancock, T., & Xie, T. (2023). Exploring relationship development with social chatbots: A mixed-method study of replika. *Computers in Human Behavior*, *140*, 107600. https://doi.org/10.1016/j.chb.2022.107600

Pickard, S. (2023). Exploring ageism as a structure of consciousness across the female life course through the work of Simone de Beauvoir. *The Gerontologist*, *63*(5), 812–819. https://doi.org/10.1093/geront/gnac123

Plant, E. A., Hyde, J. S., Keltner, D., & Devine, P. G. (2000). The gender stereotyping of emotions. *Psychology of Women Quarterly*, *24*(1), 81–92. https://doi.org/10.1111/j.1471-6402.2000.tb01024.x

Pryzgoda, J., & Chrisler, J. C. (2000). Definitions of gender and sex: The subtleties of meaning. *Sex Roles*, *43*(7–8), 553–569. https://doi.org/10.1023/a:1007123617636

Psychology Workforce Task Group. (2016). *The Aotearoa New Zealand Psychology Workforce Survey*. New Zealand College of Clinical Psychologists. https://www.nzccp.co.nz/assets/Aotearoa-NZ-Psychology-Workforce-Survey-2016.pdf

Ranaldi, L., & Pucci, G. (2025). When Large Language Models contradict humans? Large Language Models' Sycophantic Behaviour. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2311.09410

Reddit. (2023, February 13). *AI Psychologist is a lifesaver*. [Online Forum Post]. Reddit. https://www.reddit.com/r/DID/comments/110nm69/ai_psychologist_is_a_lifesaver/

Reeves, M. (1993). Simone de Beauvoir and the writing of contemporary feminist theory: Rich, butler, and the second sex. *Simone de Beauvoir Studies*, *10*(1), 159–164. https://doi.org/10.1163/25897616-01001019

Ridgeway, C. L. (2001). Gender, status, and leadership. *The Journal of Social Issues*, *57*(4), 637–655. https://doi.org/10.1111/0022-4537.00233

Riger, S. (2000). Epistemological debates, feminist voices. In *Transforming Psychology* (pp. 7–22). Oxford University PressNew York, NY. https://doi.org/10.1093/oso/9780195074666.003.0002

Riley, S., Schouten, W., & Cahill, S. (2003). *Exploring the dynamics of subjectivity and power between researcher and researched*. https://doi.org/10.17169/fqs-4.2.713

Riskind, R. G., & Tornello, S. L. (2022). 'I think it's too early to know': Gender identity labels and gender expression of young children with nonbinary or binary transgender parents. *Frontiers in Psychology*, *13*, 916088. https://doi.org/10.3389/fpsyg.2022.916088

RNZ Midday Report. (2025, March 26). *Fears over minister's bid to loosen psychologist rules*. RNZ. https://www.rnz.co.nz/news/political/547554/fears-over-minister-s-bid-to-loosen-psychologist-rules

Robinson, C. A., Bottorff, J. L., Pesut, B., Oliffe, J. L., & Tomlinson, J. (2014). The male face of caregiving: A scoping review of men caring for a person with dementia: A scoping review of men caring for a person with dementia. *American Journal of Men's Health*, *8*(5), 409–426. https://doi.org/10.1177/1557988313519671

Rohleder, P. (2012). Chapter 5: Socio-Economic Status. In *Critical Issues in Clinical and Health Psychology* (pp. 95–120). SAGE Publications Ltd. https://doi.org/10.4135/9781446252024

Ronik. (2024, January 29). *Character.AI Statistics You Need to Know in 2024*. Weam AI. https://weam.ai/blog/guide/character-ai/character-ai-statistics/

Rua, M., Groot, S., Hodgetts, D., Nikora, L. W., Masters-Awatere, B., King, P., Karapu, R., & Robertson, N. (2021). Decoloniality in Being Māori and Community Psychologists: Advancing an Evolving and Culturally-Situated Approach. In G. Stevens & C. C. Sonn (Eds), *Decoloniality and Epistemic Justice in Contemporary Community Psychology* (pp. 177–191). Springer International Publishing. https://doi.org/10.1007/978-3-030-72220-3_10

Rubin, G. S. (2012). Thinking sex. In *Deviations* (pp. 137–181). Duke University Press. https://doi.org/10.1215/9780822394068-006

Ruckenstein, M., & Schüll, N. D. (2017). The datafication of health. *Annual Review of Anthropology*, *46*, 261–278.

Rucklidge, J. J., Darling, K. A., & Mulder, R. T. (2018). Addressing the treatment gap in New Zealand with more therapists–is it practical and will it work. New Zealand Medical Journal, 131(1487).

Rudman, L. A., Moss-Racusin, C. A., Phelan, J. E., & Nauts, S. (2012). Status incongruity and backlash effects: Defending the gender hierarchy motivates prejudice against female leaders. *Journal of Experimental Social Psychology*, *48*(1), 165–179. https://doi.org/10.1016/j.jesp.2011.10.008

Rutherford, A. (2018). Feminism, psychology, and the gendering of neoliberal subjectivity: From critique to disruption. *Theory & Psychology*, *28*(5), 619–644. https://doi.org/10.1177/0959354318797194

Sansone, R. A., & Sansone, L. A. (2011). Gender patterns in borderline personality disorder. *Innovations in Clinical Neuroscience*, *8*(5), 16–20. https://www.ncbi.nlm.nih.gov/pubmed/21686143

Sarkar, R., Sarrafzadeh, B., Chandrasekaran, N., Rangan, N. K., Resnik, P., Yang, L., & Jauhar, S. K. (2025). Conversational User-AI Intervention: A Study on Prompt Rewriting for Improved LLM Response Generation. *ArXiv*, *abs/2503.16789*. https://api.semanticscholar.org/CorpusID:277244656

Sczesny, S., Formanowicz, M., & Moser, F. (2016). Can gender-fair language reduce gender stereotyping and discrimination? *Frontiers in Psychology*, *7*, 25. https://doi.org/10.3389/fpsyg.2016.00025

Seidler, Z. E., Wilson, M. J., Trail, K., Rice, S. M., Kealy, D., Ogrodniczuk, J. S., & Oliffe, J. L. (2021). Challenges working with men: Australian therapists' perspectives. *Journal of Clinical Psychology*, *77*(12), 2781–2797. https://doi.org/10.1002/jclp.23257

Shah, H., Warwick, K., Vallverdú, J., & Wu, D. (2016). Can machines talk? Comparison of Eliza with modern dialogue systems. *Computers in Human Behavior*, *58*, 278–295. https://doi.org/10.1016/j.chb.2016.01.004

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). Towards understanding sycophancy in language models. In *arXiv [cs.CL]*. https://doi.org/10.48550/ARXIV.2310.13548

Shiah, Y.-J., & Hwang, K.-K. (Eds). (2024). We are not WEIRD: Chinese Culture and Psychology. In *Frontiers Research Topics*. Frontiers Media SA. https://doi.org/10.3389/978-2-8325-4650-5

Shields, S. A. (2013). Gender and emotion. *Psychology of Women Quarterly*, *37*(4), 423–435. https://doi.org/10.1177/0361684313502312

Shrestha, S., & Das, S. (2022). Exploring gender biases in ML and AI academic research through systematic literature review. *Frontiers in Artificial Intelligence*, *5*, 976838. https://doi.org/10.3389/frai.2022.976838

Sibley, C. G., Hovered, W. J., & Liu, J. H. (2011). Pluralistic and monocultural facets of New Zealand national character and identity. *New Zealand Journal of Psychology*, *40*(3), 19–29. https://doi.org/10.1037/t21978-000

Simon-Kumar, N., Lee, A., Ameratunga, S., & Peiris-John, R. (2025). At the intersection of ethnicity, gender, and sexual orientation: mental health status of ethnic minority youth

in Aotearoa New Zealand. *Kōtuitui New Zealand Journal of Social Sciences Online*, *20*(3), 325–342. https://doi.org/10.1080/1177083x.2025.2520398

Smith, L. T. (2019). *Decolonizing research: Indigenous storywork as methodology*.

Solnit, R. (2008, April 13). *Men explain things to me; Facts didn't get in their way*. Common Dreams. https://www.commondreams.org/views/2008/04/13/men-explain-things-me-facts-didnt-get-their-way

Steele, J., James, J. B., & Barnett, R. C. (2002). Learning in a man's world: Examining the perceptions of undergraduate women in male-dominated academic areas. *Psychology of Women Quarterly*, *26*(1), 46–50. https://doi.org/10.1111/1471-6402.00042

Stout, J. G., & Dasgupta, N. (2011). When he doesn't mean you: gender-exclusive language as ostracism. *Personality & Social Psychology Bulletin*, *37*(6), 757–769. https://doi.org/10.1177/0146167211406434

Stumpf, S., Peters, A., Bardzell, S., Burnett, M., Busse, D., Cauchard, J., & Churchill, E. (2020). Gender-inclusive HCI research and design: A conceptual review. *Foundations and Trends® in Human–Computer Interaction*, *13*(1), 1–69. https://doi.org/10.1561/1100000056

Suchman, L. (2007). Feminist STS and the Sciences of the Artificial. The Handbook of Science and Technology Studies, Third Edition.

Sun, Y., & Wang, T. (2025). Be friendly, not friends: How LLM sycophancy shapes user trust. In *arXiv [cs.HC]*. arXiv. http://arxiv.org/abs/2502.10844

Sweet, P. L. (2019). The sociology of gaslighting. *American Sociological Review*, *84*(5), 851–875. https://doi.org/10.1177/0003122419874843

Taitimu, M., Read, J., & McIntosh, T. (2018). Ngā Whakāwhitinga (standing at the crossroads): How Māori understand what Western psychiatry calls "schizophrenia". *Transcultural Psychiatry*, *55*(2), 153–177. https://doi.org/10.1177/1363461518757800

Tan, K. K. H., Wilson, A. B., Flett, J. A. M., Stevenson, B. S., & Veale, J. F. (2022). Mental health of people of diverse genders and sexualities in Aotearoa/New Zealand: Findings from the New Zealand Mental Health Monitor. *Health Promotion Journal of Australia: Official Journal of Australian Association of Health Promotion Professionals*, *33*(3), 580–589. https://doi.org/10.1002/hpja.543

Tan, K., Stolte, O., Waitoki, W., & Scarf, D. (2023). *How well does psychological research in aotearoa New Zealand reflect diversity?* Zenodo. https://doi.org/10.5281/ZENODO.8187842

Tandon, R., Gaebel, W., Barch, D. M., Bustillo, J., Gur, R. E., Heckers, S., Malaspina, D., Owen, M. J., Schultz, S., Tsuang, M., Van Os, J., & Carpenter, W. (2013). Definition and description of schizophrenia in the DSM-5. *Schizophrenia Research*, *150*(1), 3–10. https://doi.org/10.1016/j.schres.2013.05.028

Tappert, C. C. (2019). Who is the father of deep learning? *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, 343–348. https://doi.org/10.1109/CSCI49370.2019.00067

Tasca, C., Rapetti, M., Carta, M. G., & Fadda, B. (2012). Women and hysteria in the history of mental health. *Clinical Practice and Epidemiology in Mental Health: CP & EMH*, *8*(1), 110–119. https://doi.org/10.2174/1745017901208010110

Taylor, S. E., Klein, L. C., Lewis, B. P., Gruenewald, T. L., Gurung, R. A. R., & Updegraff, J. A. (2000). Biobehavioral responses to stress in females: Tend-and-befriend, not fight-or-flight. *Psychological Review*, *107*(3), 411–429. https://doi.org/10.1037/0033-295X.107.3.411

The New Zealand Psychological Society. (2010). *Discussion Paper Psychologist Workforce Development Issues Trainee intern placements*. The New Zealand Psychological Society. https://www.psychology.org.nz/journal-archive/Discussion-Paper-Psychologist-Workforce-Development-Issues-Trainee-intern-placements-May-2010.pdf

The New Zealand Psychologists Board. (2016, January). *Standards and Procedures For the Accreditation of Programmes and Schemes Leading to Registration as a Psychologist in Aotearoa New Zealand*. The New Zealand Psychologists Board. https://psychologistsboard.org.nz/wp-content/uploads/2023/08/Accreditation-SP.pdf

Thelandersson, F. (2023). A historical lineage of sad and mad women. In *21st Century Media and Female Mental Health* (pp. 33–60). Springer International Publishing. https://doi.org/10.1007/978-3-031-16756-0_2

Thiel, D., & Hancock, J. (2025). *Identifying and eliminating CSAM in generative ML training data and models*. Stanford Digital Repository. https://doi.org/10.25740/KH752SM9123

Tone, A., & Koziol, M. (2018). (F)ailing women in psychiatry: lessons from a painful past. *Journal de l'Association Medicale Canadienne [Canadian Medical Association Journal]*, *190*(20), E624–E625. https://doi.org/10.1503/cmaj.171277

Turquet, L., & Women, U. N. (2011). *Progress of the World's Women 2011-2012: In pursuit of justice*. UN.

Unicef.org. (2024, October 9). *Over 370 million girls and women globally subjected to rape or sexual assault as children – UNICEF*. Unicef.org. https://www.unicef.org/press-releases/over-370-million-girls-and-women-globally-subjected-rape-or-sexual-assault-children

Ussher, J. M. (2013). Diagnosing difficult women and pathologising femininity: Gender bias in psychiatric nosology. *Feminism & Psychology*, *23*(1), 63–69. https://doi.org/10.1177/0959353512467968

Utamsingh, P. D., Richman, L. S., Martin, J. L., Lattanner, M. R., & Chaikind, J. R. (2016). Heteronormativity and practitioner-patient interaction. *Health Communication*, *31*(5), 566–574. https://doi.org/10.1080/10410236.2014.979975

Vaidyam, A. N., Wisniewski, H., Halamka, J. D., Kashavan, M. S., & Torous, J. B. (2019). Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric

Landscape. *Canadian Journal of Psychiatry. Revue Canadienne de Psychiatrie*, *64*(7), 456–464. https://doi.org/10.1177/0706743719828977

van der Toorn, J., Pliskin, R., & Morgenroth, T. (2020). Not quite over the rainbow: the unrelenting and insidious nature of heteronormative ideology. *Current Opinion in Behavioral Sciences*, *34*, 160–165. https://doi.org/10.1016/j.cobeha.2020.03.001

Vorsino, Z. (2021). Chatbots, gender, and race on Web 2.0 platforms: Tay.AI as monstrous femininity and abject whiteness. *Signs*, *47*(1), 105–127. https://doi.org/10.1086/715227

Wagman, K. B., & Parks, L. (2021). Beyond the command: Feminist STS research and critical issues for the design of social machines. In *arXiv [cs.HC]*. arXiv. http://arxiv.org/abs/2102.00464

Waitere, H., & Johnston, P. (2009). Echoed silences: In absentia: Mana Wahine in institutional contexts. *Women's Studies Journal*, *23*(2), 14–31.

Waitoki, W. (2019). "This is not us": But actually, it is. Talking about when to raise the issue of colonisation. *New Zealand Journal of Psychology*, *48*(1), 140–145. Scopus®. http://ezproxy.massey.ac.nz/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edselc&AN=edselc.2-52.0-85069512741&site=eds-live&scope=site

Waitoki, W. (2022). In defence of mātauranga Māori: a response to the'seven academics'. *The New Zealand Medical Journal (Online)*, *135*, 139–142.

Wajcman, J. (2004). *TechnoFeminism*. Polity Press.

Wang, X., Zhou, Y., & Zhou, G. (2024). The application and ethical implication of generative AI in mental health: Systematic review (Preprint). *JMIR Mental Health*, *12*, e70610. https://doi.org/10.2196/70610

Watson, J. (2023, June 24). *Your chatbot could become your therapist - and that might be a good thing* [Video]. https://www.techradar.com/computing/your-chatbot-could-become-your-therapist-and-that-might-be-a-good-thing

Watson, R., Hammans, L., Hansby, O., Barry-Walsh, J., & Every-Palmer, S. (2025). Misogyny, racism, and threats to our families: a qualitative study of harassment of female politicians. *Kōtuitui New Zealand Journal of Social Sciences Online*, *20*(4), 979–1007. https://doi.org/10.1080/1177083x.2025.2473949

Weedon, C. (1987). *Feminist practice and poststructuralist theory*. Basil Blackwell.

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How does LLM safety training fail? In *arXiv [cs.LG]*. arXiv. https://doi.org/10.48550/arXiv.2307.02483

Westfall, C. (2023, November 16). New research shows ChatGPT reigns supreme in AI tool sector. *Forbes.Com*. https://www.forbes.com/sites/chriswestfall/2023/11/16/new-research-shows-chatgpt-reigns-supreme-in-ai-tool-sector/?sh=5b2a99a850e9

Wilkinson, S. (1988). The role of reflexivity in feminist psychology. *Women's Studies International Forum*, *11*(5), 493–502. https://doi.org/10.1016/0277-5395(88)90024-6

Wilson, L. C., Newins, A. R., Kassing, F., & Casanova, T. (2024). Gender Minority Stress and Resilience Measure: A meta-analysis of the associations with mental health in transgender and gender diverse individuals. *Trauma, Violence & Abuse*, *25*(3), 2552–2564. https://doi.org/10.1177/15248380231218288

Wilson-Burke, E. (2024). *"I needed the help. They weren't giving it": Experiences of young people in tertiary education on waitlists for mental health services in Aotearoa* [Victoria University of Wellington Library]. https://doi.org/10.26686/wgtn.26113699

World Health Organisation. (2021). *Devastatingly pervasive: 1 in 3 women globally experience violence*. World Health Organisation. https://www.who.int/news/item/09-03-2021-devastatingly-pervasive-1-in-3-women-globally-experience-violence

World Health Organization. (2022). *Mental health and COVID-19: Early evidence of the pandemic's impact*. WHO Press.

Wosick-Correa, K. R., Coyle, A., & Kitzinger, C. (2003). Lesbian and gay psychology: New perspectives. *Contemporary Sociology*, *32*(3), 327. https://doi.org/10.2307/3089176

Wylie, M. S., De France, K., & Hollenstein, T. (2023). Adolescents suppress emotional expression more with peers compared to parents and less when they feel close to others. *International Journal of Behavioral Development*, *47*(1), 1–8. https://doi.org/10.1177/01650254221132777

Yee, A., Bentham, R., Byrne, J. L., Ker, A., Norris, M., Tan, K. K. H., Jones, H., Polkinghorne, T., Gonzalez, S., Withey-Rila, C., & Others. (2025). *Counting ourselves: Findings from the 2022 Aotearoa New Zealand trans and non-binary health survey*. https://researchcommons.waikato.ac.nz/items/af91c8da-bbff-4f4c-9090-6261cd6cb930

Yu, Z., Liu, X., Liang, S., Cameron, Z., Xiao, C., & Zhang, N. (2024). Don't listen to me: Understanding and exploring jailbreak prompts of large language models. In *arXiv [cs.CR]*. arXiv. http://arxiv.org/abs/2403.17336

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–21. https://doi.org/10.1145/3544548.3581388

Zheng, R., Dou, S., Gao, S., Hua, Y., Shen, W., Wang, B., Liu, Y., Jin, S., Liu, Q., Zhou, Y., Xiong, L., Chen, L., Xi, Z., Xu, N., Lai, W., Zhu, M., Chang, C., Yin, Z., Weng, R., … Huang, X. (2023). Secrets of RLHF in large language models part I: PPO. In *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2307.04964

Zhou, X., Edirippulige, S., Bai, X., & Bambling, M. (2021). Are online mental health interventions for youth effective? A systematic review. Journal of Telemedicine and Telecare, 27(10), 638–666. https://doi.org/10.1177/1357633X211047285