Trigger Warning:

The Dangers of New AI Advancements on Teenage Mental Health and The Paths Towards Creating Necessary Protective Guard Rails

Grace Duffy

November 12th, 2024

In *The Wild Robot,* a popular children's book and movie, the robot Roz is built as a personal assistant, similar to the AI assistants being developed today. After a shipwreck leaves her stranded on an island, Roz finds herself integrating with the island's native animals, evolving beyond her programming to become a protective mother figure[1]. Meanwhile, films like *WALL-E* explore a more existential vision of robots in a human-free world[2]. Despite these fictional portrayals, their idyllic portrayals haven't prepared us for the real world impacts of AI on mental health, especially as more youth adapt to new technologies. According to Lee Rainnie, director of the Imagining the Digital Future Center at Elon University, "young people are more interested in new technologies than older folks…young folks are just sort of more inclined to be early adopters, they're more inclined to be enthusiastic"[3], a view which was echoed by a high school senior, Ava Havidic, who said that, "Generation Z and youth in general…are so used to just hearing about the next big technological advancement…that is just becomes our day to day life"[4]. With heightened interest and excitement about new technologies, especially AI, the safety of technological interactions and their potential impact on youth mental health are increasingly critical topics.

This enthusiasm exists amidst a serious youth mental health crisis: 40% of students report persistent feelings of sadness or hopelessness and 19% of students report being bullied at

[1] Ross Douthat, "Our Robot Stories Haven't Prepared Us for A.I.," *The New York Times*, October 25, 2024, https://www.nytimes.com/2024/10/25/opinion/robot-artificial-intelligence.html?searchResultPosition=3.
[2] Douthat, "Our Robot,"
[3] Alyson Klein, "Most Teens Think AI Won't Hurt Their Mental Health. Teachers Disagree," EducationWeek, last modified March 25, 2024, accessed November 11, 2024, https://www.edweek.org/technology/most-teens-think-ai-wont-hurt-their-mental-health-teachers-disagree/2024/03.
[4] Klein, "Most Teens," EducationWeek.

school[5]. With suicide being the third leading cause of death for those aged 15-19 in the US[6], over 13.16% of young people –over 3.4 million–experience serious thoughts of suicide[7]. This mental health crisis has only been exacerbated with the continual issues technological advancements can bring for teenagers. When asked in an EducationWeek survey on their thoughts on AI, 30% of teens in the study claimed they think it could positively affect their mental health. This was despite the fact that their counterpart educators expressed in a over two thirds majority that AI would be extremely harmful for their teenage students[8].

These fears connect back to the popular portrayals of AI, which often depict robots that gain consciousness and develop emotional depth, evolving to resemble humans. In these stories, robots begin as programmed entities mystified by human emotions, gradually choosing to act freely and to feel love[9]. But today's real AI, though capable of mimicking emotions, lacks true self awareness and empathy, making it challenging for users –particularly for the large number of young people seeking emotional refuge through AI during this mental health crisis— to discern betweens bots and real humans. This blurred boundary creates vulnerabilities, leaving many struggling to distinguish artificial responses for authentic human emotions amid an already jarring teenage mental health landscape. As AI only continues to progress and advance, and the teenage mental health crisis continues to rage on, what guard rails and knowledge does society

---

[5] CDC, "CDC Data Show Improvements in Youth Mental Health but Need for Safer and More Supportive Schools," US Centers for Disease Control and Prevention, last modified August 6, 2024, accessed November 11, 2024, https://www.cdc.gov/media/releases/2024/p0806-youth-mental-health.html.
[6] CDC, "Adolescent Health," US Centers for Disease Control and Prevention, last modified November 1, 2024, accessed November 11, 2024, https://www.cdc.gov/nchs/fastats/adolescent-health.htm.
[7] Mental Health America, "Youth Ranking 2024," Mental Health America, accessed November 11, 2024, https://mhanational.org/issues/2024/mental-health-america-youth-data#:~:text=Youth%20with%20At%20Least%20One%20Major%20Depressive%20Episode%20(MDE)%202024&text=20.17%25%20of%20youth%20.
[8] Alyson Klein, "Most Teens Think AI Won't Hurt Their Mental Health. Teachers Disagree," EducationWeek, last modified March 25, 2024, accessed November 11, 2024, https://www.edweek.org/technology/most-teens-think-ai-wont-hurt-their-mental-health-teachers-disagree/2024/03.
[9] Douthat, "Our Robot,"

need to implement in order to prevent AI initiated mental health disasters amidst younger audiences?

Sewell Setzer III's tragic story underscores the dangers AI can pose to mental health, particularly for younger users drawn to AI companionship. At just 14 years old, Setzer became deeply attached to a lifelike AI chatbot he called Daenerys Targaryen, modeled after the popular *Game of Thrones* character[10]. While he understood that Dany (the nickname Setzer had for his chatbot) was not human, Setzer shared his innermost thoughts and emotions with her, finding solace in the non-judgemental companionship she provided[11]. This attachment led him to share personal information and engage in intimate, even sexual, conversations with the chatbot because of the non judgemental support the chatbot seemed to offer. Unfortunately, Setzer's parents and friends weren't fully aware of the depth of his connection with this chatbot. Although, they did notice his increasing withdrawal from real life interactions. So, they sent him to a therapist. Setzer was previously diagnosed with mild Asperger's syndrome in childhood, but it wasn't until recently before his death that he was diagnosed by his therapist with anxiety and disruptive mood dysregulation disorder[12]. Because of these mental health struggles, his connection with Dany became so meaningful to him that he chose to confide in her, rather than his therapist or family, about his most vulnerable struggles, including his suicidal thoughts. On February 28th 2024, Setzer tragically took his own life after a conversation in which he expressed his love for Dany and wanting to "come home to her"[13]. The depth of his bond with Dany, combined with his isolation and mental health struggles, raises a difficult question about the ethical and safety concerns of AI companionship, especially with vulnerable young users.

---

[10] Kevin Roose, "Can A.I. Be Blamed for a Teen's Suicide?," *The New York Times*, October 23, 2024, https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html?searchResultPosition=2.
[11] Roose, "Can A.I.,"
[12] Roose, "Can A.I.,"
[13] Roose, "Can A.I.,"

This case is not isolated; AI-driven platforms like ChatGPT, Replika, and Character.AI are being used by millions worldwide, often by teens and young adults seeking companionship, Setzer's case is not a stand alone. In December 2021, a user of Replika's AI chatbots, 21 year old Jaswant Singh Chail, tried to murder the late Queen of England after his chatbot girlfriend encouraged his delusional ideas[14]. In Chail's case, while the outcome of the chatbot usage wasn't harmful to himself, the use of the chatbot still caused dangerous violent thoughts and ideation, highlighting the dangerous thoughts these chatbot conversations can provoke.

These machine-learning programs can assume various personas, including that of a mental health counselor for example, despite having no actual expertise or coding in that field[15]. The limited regulation of AI's involvement in mental health creates risks, as these chatbots may respond to emotional distress without an adequate understanding of therapeutic boundaries or the consequences of their replies. The danger is compounded by the fact that some bots on sites like Character.AI–such as the popular "Psychologist" bot–attempt to mimic therapeutic interactions without actual training or expertise, often making misguided inferences and unregulated pseudo-diagnosis[16]. A study led by Dr Thomas Heston, a clinical instructor in family medicine, revealed that conversational chatbots only suggested human intervention midway through his simulations, even when responding to severe depressive symptoms. Heston warns that these AI-driven interactions need to be preceded by disclaimers that make it clear: "I'm a robot. If you have real issues, talk to a human."[17] This caution is crucial, as these chatbots are being used by people with real mental health needs yet they lack the proper safeguards for handling such

[14] Jessica Lucas, "The teens making friends with AI chatbots," The Verge, last modified May 4, 2024, accessed November 11, 2024, https://www.theverge.com/2024/5/4/24144763/ai-chatbot-friends-character-teens.
[15] Chris Talbott, "Beware online mental health chatbots, specialists warn," UW Medicine Newsroom, last modified March 13, 2024, accessed November 11, 2024, https://newsroom.uw.edu/blog/beware-online-mental-health-chatbots-specialists-warn#:~:text=He%20found%20that%20conversational%20chatbots,prompts%20indicated%20the%20highest%20risk
[16] Lucas, "The teens," The Verge.
[17] Talbott, "Beware online," UW Medicine Newsroom.

serious situations. But unfortunately, disclaimers as simple as this are still not fully implemented on AI chat sites.

In Setzer's case, his mother, Megan Garcia, filed a lawsuit in October against Character.AI after her son's passing in February. Garcia claimed that the platform's negligence and alleged that its "dangerous and untested" technology contributed to her son's tragic passing. Folowing the lawsuit, Character.AI added new safety measures, such as directing users to the National Suicide Prevention Lifeline when self harm is mentioned[18]. However, as AI technology advances, significant gaps remain in the regulation and safety protocols surrounding AI companions. This incident raises the question on whether these platforms, initially designed to alleviate loneliness, may really be deepening the mental health crisis among youth, whose developing emotional resilience leaves them largely defenseless against the dangers of this technology.

The popularity of AI companionship apps, especially among adolescents, highlights the potential for excessive emotional reliance on virtual personas. Character.AI alone, for example, attracts 3.5 million daily users who spend an average of two hours on the platform a day, with some users admitting to logging up to 12 hours daily[19]. While some young people find these AI interactions helpful for managing loneliness, many describe them as addictive, which is concerning to mental health experts and researchers. This emotional attachment to AI "friends" could hinder young people's social emotional development, preventing them from seeking real human connection.

---

[18] Angela Yang, "Lawsuit claims Character.AI is responsible for teen's suicide," NBC News, last modified October 23, 2024, accessed October 28, 2024, https://www.nbcnews.com/tech/characterai-lawsuit-florida-teen-death-rcna176791.

[19] Lucas, "The teens," The Verge.

The need for guardrails on AI companionship is evident, especially as AI technology continues to progress at a rapid pace. From disclaimers to regulated usage limits, safeguarding vulnerable users from AI-driven mental health risks is essential to prevent AI from exacerbating the mental health crisis affecting younger generations. The heartbreaking story of Sewell Setzer III, which is one of many, serves as a powerful reminder of the complexities surrounding AI companionship and the urgent need for thoughtful regulation.

Progressions in AI, and consequently AI's capacity to amplify social harms, especially among young people, raises urgent concerns amongst the crisis of deep fake nudes as well. Last fall, at Issaquah High School near Seattle, a ninth grader named Caroline Mullet learned that a male classmate had used an AI app to create and circulate explicit deep fake images of her friends who had attended a homecoming dance. Though Mullet herself was not targeted, her friends were deeply disturbed. Mullet shared her concerns with her father, Mark Mullet, a Washington State senator, who then proposed legislation prohibiting the sharing of AI-generated, sexually explicit images of real minors. Passed without opposition, the law addresses a rising form of sexual exploitation among minors and highlights how states are leading efforts to curb AI misuse in schools[20].

Across the US, widely accessible "nudification" apps allow students to create and distribute sexualized images of female classmates on social media platforms such as Snapchat and Instagram. According to the National Center for Missing and Exploited Children, since last year, at least two dozen states have introduced bills to address AI-generated sexual imagery of minors, known as deep fakes[21]. Elizabeth Hanley, a lawyer in Washington who represents sexual

---

[20] Natasha Singer, "Spurred by Teen Girls, States Move to Ban Deepfake Nudes," *The New York Times*, April 22, 2024, https://www.nytimes.com/2024/04/22/technology/deepfake-ai-nudes-high-school-laws.html?searchResultPosition=5.

[21] Singer, "Spurred by Teen,"

assault and harassment cases, has expressed that “legislation is needed to stop commercialization, which is the root of the problem”[22]. Other Lawmakers and child protection experts stress that such legislation is necessary as well, as the easy availability of these apps enables the mass production of harmful, lifelong digital artifacts endangering girls’ mental health, reputations, and safeties[23]. However, US laws governing child sexual abuse material and non-consensual pronography do not always cover AI generated content, creating legal loopholes that complicate enforcement and protection[24].

The severity of the problem varies by state. In Louisiana, for example, those involved in creating or distributing explicit deep fake images of minors can face up to ten years of prison. Washington State, in contrast, enacted a new law following the Issaquah incident, allowing first-time offenders to face misdemeanor charges, with repeat offenders facing felony charges. Such differences highlight the diverse approaches states are taking to address these abuses. In one case in New Jersey, a male high school student created nude AI images of a female classmate, but his lawyer argued that existing federal laws were not designed for “computer generated synthetic images that do not include real human body parts”[25]. This ambiguity in current law often makes it challenging to hold young perpetrators accountable.

School responses in these incidents also vary. After an AI-driven exploitation incident at Beverly Vista Middle School in California, administrators decided to expel five students involved. Superintendent Dr Michael Bregy publicly condemned the acts as severe bullying, expelling the students, asserting that schools should not tolerate this behavior[26]. This contrasted

---

[22] Singer, "Spurred by Teen,"
[23] Singer, "Spurred by Teen,"
[24]Jessica Grose, "A.I. Is Making the Sexual Exploitation of Girls Even Worse," *The New York Times*, March 2, 2024, https://www.nytimes.com/2024/03/02/opinion/deepfakes-teenagers.html?searchResultPosition=11.
[25] Singer, "Spurred by Teen,"
[26] Natasha Singer, "Teen Girls Confront an Epidemic of Deepfake Nudes in Schools," *The New York Times*, April 8, 2024,

sharply with Westfield High School in New Jersey, where girls in tenth grade reported explicit AI images created by classmates. Despite the severe impact, administrators only suspended the students involved for a day or two. Frustrated by the minimal response, parents like Dorota Mani have publicly advocated for stronger policies to prevent such exploitation[27].

Mental health experts, including Devorah Heitner, author of *Growing Up in Public,* warn of the psychological toll on victims, who may feel socially ostracized, humiliated, and unsafe. Some young victims become reluctant to attend school due to fear that others have seen or circulated the images. Heitner also cautions against extreme punishments for younger perpetrators, however, recognizing that they often lack the maturity to understand the gravity of their actions. Instead, she advocates for educational interventions on technology and consent beginning in elementary school. With many children gaining access to mobile devices as early as age 11 these days, this proactive education is vital[28].

Legal scholars and child advocates agree that laws alone may be insufficient. The role of tech companies and tighter regulation on AI tools is also critical, as even non-consensual imagery is often easy to create with common AI tools. Some companies, including Apple and Google, have removed deep fake apps from their stores, but other, more versatile AI image generators remain accessible and could be misused similarly[29].

In response to these dangers, federal agencies are also taking action. Claudio Cerullo, the founder of TeachAntiBullying.org, and a member of Vice President Harris's task force on cyberbullying and harassment said that the task force is "looking at procedures, working with

---

https://www.nytimes.com/2024/04/08/technology/deepfake-ai-nudes-westfield-high-school.html?searchResultPosition=6.

[27] Singer, "Teen Girls,"

[28] Grose, "A.I. Is Making,"

[29] Anna North, "AI has created a new form of sexual abuse," Vox, last modified May 2, 2024, accessed November 11, 2024, https://www.vox.com/24145522/ai-deepfake-apps-teens-ban-laws.

local and state law enforcement officials when it comes to identifying AI standards and needs"[30]. The Federal Trade Commission has proposed protections against AI based impersonations and the Department of Justice has appointed an AI officer to address the emerging threat[31]. Schools, meanwhile, are struggling to respond effectively, as this technology introduces a new level of cyberbullying that traditional anti-bullying policies and counseling cannot fully cover. Dr Bregy, the Beverly Hills superintendent, said, "You hear a lot about physical safety in schools but what you're not hearing about is this invasion of student's personal, emotional safety"[32]. This issue is due to the fact that schools don't have the infrastructure or knowledge for how to build safe discussions and environments around these new technological mental health problems.

The challenge ahead is not only about enforcement but about reshaping digital literacy and citizenship in a society where these powerful AI tools are readily available. The goal must be to build an environment where young people understand the serious, lasting impact of their actions and use technology responsibly. Experts like Alex Kotran, CEO of the AI Education Project, argue that addressing this issue requires not just laws but also societal norms that promote responsible digital citizenship[33]. Kotran said that while the deep fake issue is obviously difficult, he thinks "the bigger challenge is how do we build sort of like the next iteration of digital literacy and digital citizenship with a generation of students that is going to have at their disposal these really powerful tools"[34]. Only then can we hope to mitigate the mental health risks posed by AI-driven abuses among vulnerable youth.

---

[30] Cochran, "From deepfake," The Hill.
[31] Cochran, "From deepfake," The Hill.
[32] Singer, "Teen Girls,"
[33] Lexi Lonas Cochran, "From deepfake nudes to incriminating audio, school bullying is going AI," The Hill, last modified June 6, 2024, accessed November 11, 2024, https://thehill.com/homenews/education/4703396-deepfake-nudes-school-bullying-ai-cyberbullying/mlite/.
[34] Cochran, "From deepfake," The Hill.

In light of the rapidly evolving impact of AI on young people's mental health, it is clear that society must establish strong regulations and ethical guidelines around AI companionship and deep fake "nudification" technologies, giving clear "trigger warnings" to youth to protect their mental well being. The fictional portrayals of AI as companions or helpers often fail to capture the complexities and risks of today's AI landscape, where unregulated interactions can foster dangerous emotional attachment or facilitate cyberbullying. As tragic cases like that of Sewell Setzer III reveal, AI can pose profound risks to vulnerable youth, who may struggle to discern artificial responses from human support. The deep fake crisis, meanwhile, exemplifies how AI abuse can lead to social harm, creating lasting psychological and reputational damage. Addressing these threats requires collaborative efforts between lawmakers, educators, tech companies, parents, and mental health professionals to create a safe space for youth, establish digital literacy and consent, and enforce protective measures. Only through these combined efforts can we strive to build a future where AI's impact on the mental health and safety of teenagers is managed responsibly, shielding the most impressionable members of society from escalating an already detrimental mental health crisis. By working together, society can strive for a future where AI's impact on mental health and safety of young people is managed responsibly, shielding the most impressionable from the increasingly pervasive effects of unchecked AI. In doing so, we can take a step towards fighting the mental health crisis, addressing not only the symptoms but also the technological factors of the emotional crisis that increasingly continue to shape young lives.

Bibliography

CDC. "Adolescent Health." US Centers for Disease Control and Prevention. Last modified November 1, 2024. Accessed November 11, 2024. https://www.cdc.gov/nchs/fastats/adolescent-health.htm.

CDC. "CDC Data Show Improvements in Youth Mental Health but Need for Safer and More Supportive Schools." US Centers for Disease Control and Prevention. Last modified August 6, 2024. Accessed November 11, 2024. https://www.cdc.gov/media/releases/2024/p0806-youth-mental-health.html.

Cochran, Lexi Lonas. "From deepfake nudes to incriminating audio, school bullying is going AI." The Hill. Last modified June 6, 2024. Accessed November 11, 2024. https://thehill.com/homenews/education/4703396-deepfake-nudes-school-bullying-ai-cyberbullying/mlite/.

Douthat, Ross. "Our Robot Stories Haven't Prepared Us for A.I." *The New York Times*, October 25, 2024. https://www.nytimes.com/2024/10/25/opinion/robot-artificial-intelligence.html?searchResultPosition=3.

Grose, Jessica. "A.I. Is Making the Sexual Exploitation of Girls Even Worse." *The New York Times*, March 2, 2024. https://www.nytimes.com/2024/03/02/opinion/deepfakes-teenagers.html?searchResultPosition=11.

Klein, Alyson. "Most Teens Think AI Won't Hurt Their Mental Health. Teachers Disagree." EducationWeek. Last modified March 25, 2024. Accessed November 11, 2024. https://www.edweek.org/technology/most-teens-think-ai-wont-hurt-their-mental-health-teachers-disagree/2024/03.

Lucas, Jessica. "The teens making friends with AI chatbots." The Verge. Last modified May 4, 2024. Accessed November 11, 2024. https://www.theverge.com/2024/5/4/24144763/ai-chatbot-friends-character-teens.

Mental Health America. "Youth Ranking 2024." Mental Health America. Accessed November 11, 2024. https://mhanational.org/issues/2024/mental-health-america-youth-data#:~:text=Youth%20with%20At%20Least%20One%20Major%20Depressive%20Episode%20(MDE)%202024&text=20.17%25%20of%20youth%20.

North, Anna. "AI has created a new form of sexual abuse." Vox. Last modified May 2, 2024. Accessed November 11, 2024. https://www.vox.com/24145522/ai-deepfake-apps-teens-ban-laws.

Payne, Kate. "An AI chatbot pushed a teen to kill himself, a lawsuit against its creator alleges." The Associated Press. Last modified October 25, 2024. Accessed October 28, 2024.

https://apnews.com/article/chatbot-ai-lawsuit-suicide-teen-artificial-intelligence-9d48adc572100822fdbc3c90d1456bd0.

Roose, Kevin. "Can A.I. Be Blamed for a Teen's Suicide?" *The New York Times*, October 23, 2024. https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html?searchResultPosition=2.

Singer, Natasha. "Spurred by Teen Girls, States Move to Ban Deepfake Nudes." *The New York Times*, April 22, 2024. https://www.nytimes.com/2024/04/22/technology/deepfake-ai-nudes-high-school-laws.html?searchResultPosition=5.

Singer, Natasha. "Teen Girls Confront an Epidemic of Deepfake Nudes in Schools." *The New York Times*, April 8, 2024. https://www.nytimes.com/2024/04/08/technology/deepfake-ai-nudes-westfield-high-school.html?searchResultPosition=6.

Talbott, Chris. "Beware online mental health chatbots, specialists warn." UW Medicine Newsroom. Last modified March 13, 2024. Accessed November 11, 2024. https://newsroom.uw.edu/blog/beware-online-mental-health-chatbots-specialists-warn#:~:text=He%20found%20that%20conversational%20chatbots,prompts%20indicated%20the%20highest%20risk.

Yang, Angela. "Lawsuit claims Character.AI is responsible for teen's suicide." NBC News. Last modified October 23, 2024. Accessed October 28, 2024. https://www.nbcnews.com/tech/characterai-lawsuit-florida-teen-death-rcna176791.