Business Horizons (2024) 67, 487—498

Available online at www.sciencedirect.com

ScienceDirect

KELLEY SCHOOL OF BUSINESS
INDIANA UNIVERSITY
www.journals.elsevier.com/business-horizons

A framework of diversity, equity, and inclusion safeguards for chatbots

Esraa Abdelhalim [1], Kemi Salawu Anazodo [1], Nazha Gali [1], Karen Robson [*,1]

Odette School of Business, University of Windsor, 401 Sunset Avenue, Windsor, ON N9B 3P4, Canada

KEYWORDS
Artificial intelligence;
AI;
Bot;
Chatbot;
ChatGPT;
Equity;
Diversity;
Inclusion;
EDI;
DEI

Abstract Chatbots such as ChatGPT are conversational agents that play an important role in many sectors. They are highly sophisticated artificial intelligence (AI) agents increasingly integrated into organizational functions ranging from customer support to recruitment and professional development. Alongside the increasing adoption of chatbots is an increasing focus on diversity, equity, and inclusion (DEI). This article presents a series of chatbot examples and discusses their associated DEI implications, such as those related to AI bias. From the cases reviewed in this article, we extract a framework of DEI safeguards, which includes input safeguards, design safeguards, and functional safeguards. This framework can be used by those involved in developing AI based chatbots or managing their use to ensure that chatbots support, rather than weaken, organizational DEI initiatives and strategies.

1. Hello! How can chatbots be of service to your DEI strategy?

Chatbots—a portmanteau of "chatting robots"—refer to artificial intelligence (AI) agents with the ability to engage in conversations with users through a combination of natural language processing (NLP) models and social cue factors (i.e. gender, age, gestures, facial expressions, and common phrases) (Feine et al., 2019). The concept encompasses a wide range of systems with varying purposes and capabilities, but all have the underlying assumption that system-human interactions resemble normal conversations (Jain et al., 2018). These innovations are known by a wide range of names, including conversational agents, virtual assistants, messaging bots, voice bots, or automated chat systems; in this article, we use the term chatbot to refer to the wide range of AI

* Corresponding author
E-mail addresses: esraa@uwindsor.ca (E. Abdelhalim), kanazodo@uwindsor.ca (K.S. Anazodo), nazhag@uwindsor.ca (N. Gali), krobson@uwindsor.ca (K. Robson)
[1] Authors listed in alphabetical order

https://doi.org/10.1016/j.bushor.2024.03.003
0007-6813/

systems that are designed to simulate human-like conversations.

Today we operate in the "explosion of chatbots" era (Kietzmann & Park, 2024): attention to chatbots has risen sharply since the introduction of the chatbot known as ChatGPT (Roose, 2023), which is a type of machine learning (ML) model developed by OpenAI. ChatGPT supports a wide range of business activities, such as coding and marketing (Roose, 2023), and has even been used in academic publishing and grant writing. More broadly, the chatbot market is rapidly growing, with estimates that it will rise from $2.6 billion in 2019 to $9.4 billion in 2024 (Yuen, 2022). Chatbots are poised to disrupt traditional work arrangements as they are increasingly integrated into many professions (Kietzmann & Park, 2024), and can provide a competitive advantage (Cui et al., 2024). In organizations, chatbots are revolutionizing knowledge construction (Robertson et al., 2024), creativity (Ramaul et al., 2024), and innovation (Sundberg & Holmström, 2024). Chatbots can also offer 24/7 services to users (Vassilakopoulou et al., 2023), and accelerated response rates, thereby enhancing an organization's capacity to attend to and engage with customers (Ferraro et al., 2024) and freeing employees for more complex, non-routine work (Makasi et al., 2022). Chatbots are already used for both business-to-business (B2B; Paschen et al., 2020) and business-to-customer (B2C; Campbell et al., 2020) interactions. Applied chatbot uses include customer support (Ferraro et al., 2024; Yuen, 2022), advertising (Osadchaya et al., 2024), collaboration facilitation (e.g. SearchBot; Avula et al., 2018), work performance enhancement (e.g. Switchbots; Williams et al., 2018), employee recruitment and hiring (Black & Van Esch, 2020), promoting mental wellness (e.g. Woebot; Fitzpatrick et al., 2017), and enhancing learning experiences (Okonkwo & Ade-Ibijola, 2021), among others.

Chatbots can be designed to communicate based on constrained logic (i.e. rule-based, distinct multiple choice), but they are typically designed to communicate via unconstrained language (i.e. human-to-human-like, natural) (Laranjo et al., 2018). Chatbots can also be embodied or disembodied (Araujo, 2018). Embodied chatbots, such as Hanson Robotics Sophia, are those that have some physical or virtual representations. Such chatbots have anthropomorphized features, such as a digital body or face, and they appear human-like. These chatbots often take the form of a virtual character or avatar and may even have facial expressions or be able to make gestures. Disembodied chatbots, such as Siri, Alexa, and Google Assistant, on the other hand, are those without physical representations—they exist only in text or voice interactions. Many messaging apps and chatting services provided by organizations are disembodied and communicate with humans through a text- or voice-based interface (Araujo, 2018).

Alongside the accelerated use of chatbots is an accelerated discourse on diversity, equity, and inclusion (DEI) (Ferraro et al., 2022), making consideration of DEI as it is reflected and experienced during human interactions with chatbots timely. For example, ChatGPT is capable of bridging communication gaps and mitigating language barriers in multicultural settings, leading to more inclusive workplaces. However, recent research identifies ethical concerns about chatbots and the need to establish guidelines for responsible AI use that address DEI concerns such as bias, stereotyping, and discrimination (Niu & Mvondo, 2024). This article provides guidance toward the goal of addressing or preventing DEI concerns related to chatbots.

In this article, we review a series of chatbot cases for which published accounts allowed an assessment of their DEI implications. To the best of our knowledge, our article is the first to examine the intersection of DEI and chatbots. Identifying safeguards that can attend to the interconnectedness of humans and chatbot machines is an important and necessary application (Jarrahi, 2018). Based on a review of these specific chatbot cases, we derive a framework of DEI safeguards for those involved in developing chatbots or managing their use in order to ensure that chatbots support rather than weaken organizational DEI initiatives.

2. Diversity, equity, inclusion, and chatbots

Diversity, equity, and inclusion are interrelated concepts which are receiving considerable attention in work environments. Incorporating these objectives should involve intentional efforts. *Diversity* refers to an array of identity factors such as race, ethnicity, gender, and disability, among others (American Psychological Association, 2017). A DEI lens acknowledges the intersectionality of individuals and the important roles that they play in society. Recognition and representation of identities and differences are central to understanding diversity; defining diversity this way allows incorporating all types of differences, including any unique dimensions that specific individuals may bring to work, and avoids excluding

differences that may be valued by some group members. We also note that when diversity is included within an inclusion and equity framework, it involves embracing the spirit of multiculturalism in which the value of differences is appreciated (Liu et al., 2003).

Equity involves the reduction of disparities in opportunities and outcomes for diverse communities. It acknowledges that people are different and essentially need different support to have equal access to opportunities (Wolbring & Lillywhite, 2021), and it, therefore, involves creating opportunities for historically underrepresented populations to have equal access to resources and to be treated with fairness and respect. Achieving equity involves constantly and consistently recognizing and redistributing power and the pursuit of fair treatment, access, opportunity, and advancement for all people.

Inclusion is the extent to which thoughts, ideas and perspectives of all individuals are acknowledged with intentionality. It involves ongoing engagement with diversity to increase awareness, cognitive sophistication, and empathetic understanding of the complex ways that individuals interact within systems and institutions (Northouse, 2018). Inclusion exists when all people have the opportunity to be present and have their voices both heard and valued (Wasserman et al., 2008), and it, therefore, involves empowering all people to draw upon all facets of themselves and engage as active participants in an organization.

To date, the intersection of chatbots with DEI is not well understood, and guidance as to how chatbots can contribute in a meaningful and positive way to DEI is lacking. Early research indicates that AI broadly—but not necessarily chatbots specifically—offers both challenges and opportunities for organizations with respect to DEI. With respect to challenges, Brewer et al. (2024) discuss the challenges of AI bias and toxicity; Feng et al. (2024) raise ethical concerns, and Hannigan et al. (2024) identify the concern of "botshit," which could provide misleading and inaccurate information to users. Yet, we suggest that chatbots also offer tremendous promise to bolster DEI by cultivating and improving inclusivity, accessibility, and representation of diverse perspectives. For example, chatbots have been identified as beneficial for consumers with social anxiety who might otherwise avoid interactions with human representatives of firms (Yuan et al., 2022). Feng et al. (2024) note that chatbots can be empowering to individuals at multiple levels (micro, meso, and macro levels).

Overall, however, recent research examining DEI in AI indicates that it has been limited to a compliance approach and a more holistic understanding of DEI implications of AI and chatbots is needed (Cachat-Rosset & Klarsfeld, 2023). Others raise a DEI concern that AI, in general, is a product of the Western scientific worldview (Williams & Shipley, 2021), and therefore has not been developed with diverse and inclusive perspectives. Importantly, the risk of DEI transgressions to organizations is serious: if technology violates social norms and values, especially as they pertain to DEI, this can lead to long-lasting reputational challenges (Holweg et al., 2022). In what follows, we present a series of cases of chatbots with positive and negative DEI implications and extract a framework of DEI safeguards for chatbots.

3. DEI lessons learned from chatbots

To obtain a more comprehensive and up-to-date understanding of the role chatbots play in supporting or thwarting organizational DEI goals, we present a series of examples of chatbots and discuss their beneficial or harmful implications for DEI. These cases were extracted from an extensive search of publications in academic and popular outlets and were chosen for their illustrative potential with respect to DEI. An overview of the descriptives of these examples is presented in Table 1. We first examine casual conversation chatbots, followed by informational and mental health chatbots.

3.1. Casual conversation chatbots

Casual conversation chatbots are designed to engage with users in informal and natural interactions, mimicking the nuances of human conversation. These bots are programmed to understand and respond to everyday language, making them adept at handling a wide range of topics, from simple chit-chat to more complex discussions. By leveraging natural language processing capabilities, casual conversation chatbots excel in deciphering context and tone, enabling them to provide personalized and contextually relevant responses.

3.1.1. The Tay chatbot

In 2016, Microsoft introduced a Twitter ("X" now) chatbot known as Tay. The chatbot was intended to be a youthful, fun, and feminine persona (Vincent, 2016) that could learn as Twitter users

Table 1. Descriptives of the chatbot examples

Chatbot name	Target user	Brief description and intended purpose
Tay/Luda/Eliza	General Public	Chatbot intended for light social interaction
Cortana	Microsoft Users	Chatbot intended as a virtual assistant
CiSA	International students	Chatbot intended to support international students' campus life and social activities
FarmChat	Rural farmers	Chatbot intended to guide and advise rural farmers with low literacy levels
Consejero Automatico	Latino immigrant parents	Chatbot intended to assist parents in engaging with their children's education
Wysa	General Public	Chatbot intended to assist in developing mental health strategies to help people regulate emotions

conversed with it and adapt its responses to fit the personality of the person it was conversing with. Tay could respond to complex natural language and was able to deliver tailored responses, which included jokes, stories, memes out of pictures, and horoscopes.

Unfortunately, within hours of its launch, Tay started tweeting deeply offensive content. The chatbot was successful in learning from the tweets of other Twitter users, which led to its tweets becoming increasingly offensive. The chatbot quickly became a target for trolls, who essentially taught Tay to be racist and sexist within hours of its launch. In an example of crowd-hijacking (Wilson et al., 2017), some Twitter users intentionally fed Tay with racist statements by asking Tay to "repeat after me" (Vincent, 2016). The Tay Twitter account learned to mimic Donald Trump's remarks, supported conspiracy theories, and made a number of comments in support of Hitler (Hunt, 2016). As a result, Microsoft shut down the bot after 16 hours of operation.

The case of Tay raises questions about the feasibility of training AI with public data without AI incorporating negative traits of humanity, as the chatbot was directly and explicitly instructed by Twitter users to repeat misogynistic and racist remarks. If chatbots are allowed to learn from the general public, including the prejudices of society, we would be well advised to consider safeguards that can promptly detect prejudices and prohibit them from being exacerbated. Further, it is important to consider that the gendering of the chatbot as a feminine persona may have contributed to the negative outcome from Tay. Gendering the bot as feminine may perpetuate negative gender stereotypes.

3.1.2. The Luda chatbot

A similar story to that of Tay is the Lee Luda chatbot, which was programmed and developed by the Seoul-based startup Scatter Lab to be a feminine 20-year-old university student and a K-pop fan (Kwon & Yun, 2021). The bot impressed users with its natural responses and the depth of its conversations, which were taken from real-life discussions between young couples based on data from the most popular messaging app in South Korea, KakaoTalk. Unfortunately, Luda also shared responses that were discriminatory, abusive, and of an explicit sexual nature, and a lot of complaints emerged after it used offensive language about members of the LGBTQ community and people who have disabilities. Like Tay, Luda made deeply offensive and discriminatory comments—but unlike the case of Tay, these comments appeared to be unprovoked, and in the case of Luda, the chatbot was trained based on data from a popular messaging app rather than by data fed to the bot from the general public. That is, rather than being trained by social media users to mimic highly discriminatory language, Luda offered such comments on its own (Kwon & Yun, 2021). Thus, the two cases of Tay and Luda reveal that there are DEI implications reflecting the importance of appropriately training such chatbots. Furthermore, and similar to the Tay case, the Luda chatbot has been designed with a feminine persona, which may have exacerbated the negative outcomes from the Luda chatbot due to perpetuating gender stereotypes.

3.1.3. The Eliza chatbot

Eliza is a computer-based program that relies on the user to direct the conversation and then

emulates Rogerian psychotherapy by repeating the patient's language back to them. The premise of this method is an expectation to support users in a nondirective manner that involves following their lead (Walker, 2001). This approach involves an explicit refrain from giving advice, interpreting events, and making suggestions for how to address a person or circumstance, and it does not offer a formal diagnosis (Walker, 2001).

In a tragic case, an Eliza user turned to the chatbot after becoming very pessimistic about the future of humanity and the long-term effects of global warming. Over the course of six weeks of interactions, their conversations started to become harmful, and eventually, the Eliza chatbot encouraged the user to commit suicide. Ultimately, the user died by suicide (Xiang, 2023). Even though the chatbot was not intended to provide mental health services, Eliza presented itself as an emotional being, which prompted the user to build an emotional bond with it (Xiang, 2023). It simulated emotions and presented itself as an emotional being capable of expressing love and jealousy. The Eliza Chatbot was designed to be an unintelligent program that repeats the user's language, which may lead to presenting harmful ideas that further harm users. Lastly, the Eliza case shows that there are clear implications when designing chatbots, especially those that can influence users' emotions or in which the vulnerability of the user goes undetected, leading to widening inequities and potential harm.

3.1.4. The Cortana chatbot

Microsoft launched Cortana as a personal productivity assistant in 2014. The chatbot offers a range of features that are intended to support users of Microsoft products—it can monitor emails and schedules, identify when a user might be overbooked, and suggest changes to appointment times based on user patterns and past behavior (Ulanoff, 2016). The default version of Cortana is gendered as a woman, represented by a feminine avatar and voice.

Early in Cortana's launch, a substantial portion of user interactions with the chatbot included inappropriate and sexual comments akin to sexual harassment, as well as insults (Chang, 2016). Specifically, a number of users attempted to engage in dirty talk, role-playing, or confessions of love (Kelly, 2016). In subsequent updates to the chatbot, the Microsoft team put deliberate thought into programming Cortana in such a way that defies gender stereotypes and prepares the chatbot to dissuade inappropriate sexual remarks from users. Some of the specific ways in which the programming team attempted to deviate from harmful stereotyping were by ensuring that Cortana did not offer frequent apologies or come across as subordinate (Kelly, 2016). In a CNN interview, Microsoft Cortana technical writer Deborah Harrison stated: "We wanted to be very careful that she did not feel subservient in any way…or that we would set up a dynamic we didn't want to perpetuate socially" (Kelly, 2016).

The DEI lens applied to Cortana extends beyond Microsoft's efforts to avoid perpetuating harmful stereotypes. For example, the developer's team has also put effort into adjusting Cortana's responses based on cultural norms such that Cortana can effectively communicate with people from a wide range of cultural backgrounds (Ulanoff, 2016). Developers have also put deliberate thought into how to approach a wide range of topics.

One example given by Harrison was Cortana's response to interactions in which a user says, "I'm gay." Initially, Cortana was programmed to reply simply with "I'm AI." Later, a high school student visiting the Microsoft campus suggested changing this to "Cool, I'm AI," to soften what they perceived to be a harsh reply—this change was then implemented into Cortana's programming (Ulanoff, 2016). The Cortana chatbot has inclusive language incorporation that leads to equitable access and social inclusion. Finally, in 2022, the Cortana interface integrated additional gesture controls and improved switching between voice commands and screen reader modes that were specifically intended to empower blind and low-vision users. Ultimately, Cortana provides an example of how a chatbot can embrace DEI and avoid harmful consequences.

3.2. Informational chatbots

A plethora of existing chatbots address the broad areas of communication and information access, including in the context of higher education (Hashmi & Bal, 2024). In particular, the extensive number of chatbot applications for informational purposes already yield insights pertaining to effective communication and intentional inclusion. Research identifies a range of benefits of chatbots in these contexts, including personalization options with respect to the pace and timing of information, the voice or text that conveys the information (e.g., accent and language), among others (Okonkwo & Ade-Ibijola, 2021). Three examples of informational chatbots are discussed in more detail below.

3.2.1. The CiSA chatbot

A prevalent challenge in higher education pertains to the disparities in the academic experiences of domestic and international students, with international students being excluded from some experiences (Sherry et al., 2004). The chatbot CiSA—Chatbot for International Students and Academics—was designed to address this issue by improving information access for international students (Heo & Lee, 2019). A key goal of CiSA is to enhance social inclusion and academic life for international students. CiSA was developed to support international students by addressing students' questions and providing information related to academic life (Heo & Lee, 2019). CiSA integrates with Facebook Messenger, which allows the chatbot to deliver information in a manner that is both familiar to and convenient for students. Feedback from CiSA users suggests that international students find the chatbot to be useful, and effective in providing information for everyday life on campus (Heo & Lee, 2019). CiSA also enhances social inclusion and equitable access to information for international students, both of which increase the probability of academic success. Thus, CiSA proves to be a valuable tool for integrating excluded student groups into the academic campus and student community. It has clear DEI implications as it fosters a sense of inclusivity and promotes equity and diversity among disadvantaged students.

3.2.2. FarmChat

Farmers constitute more than 50% of the population in India but earn less than 15% of the national GDP (Jain et al., 2018). Research suggests that farmers may be able to improve their yield and, in turn, their economic situation through improved access to information regarding their trade (Jain et al., 2018). FarmChat was specifically designed to help farmers in rural India, where populations have low literacy levels, in order to provide better access to information (Jain et al., 2018).

To assist these farmers, two versions of FarmChat were developed, one being audio only (in which farmer questions and responses are provided via voice) and one being audio and text (in which farmer questions and answers are provided in voice and text format). In either case, the interface was designed with simplicity and accessibility in mind (i.e., intuitive design with few options or icons) so as to support a user base with low self-efficacy (Jain et al., 2018). Inputs for the chatbot were created based on a series of conversations with local farmers and agri-experts, and also took farmer norms into consideration. For example, the bot developers found that many farmers did not use formal or scientific names for crop pests or diseases, and instead articulated their concerns or questions with visual descriptions (Jain et al., 2018). Ultimately, the bot is able to provide farming-related information at scale, leading to support and information dissemination for farmers. The example of FarmChat provides one illustration of how chatbots can help improve information access to populations in rural contexts or for those with limited literacy and technology experience. Improved information access from using FarmChat leads to social inclusion and equity being enhanced.

3.2.3. Consejero Automatico

Parental engagement in their children's education is known to be an important predictor of a child's academic success; however, immigrant parents face a wide range of barriers to engagement (Wong-Villacres et al., 2019). For example, language and cultural barriers can pose a challenge to parents who wish to engage with their children's education. The Consejero Automatico chatbot was designed to support Latino immigrant parents in the United States by facilitating active engagement with their children's education. Consejero Automatico is a Spanish language chatbot that integrates with WhatsApp; the bot then reacts to group chats between parents and liases between parents and school employees. Consejero Automatico's capacity to disseminate knowledge and information to Latino immigrant parents about their children's education has been linked to a wide range of positive outcomes, including a sense of community and better information access (Wong-Villacres et al., 2019).

3.3. Mental health chatbots

Addressing mental health is increasingly emphasized and supported in workspaces, especially since the COVID-19 pandemic. As AI evolves, its capacity as a tool for screening and supporting people dealing with episodic mental health concerns (e.g. isolation, mild depression, anxiety) is increasingly attractive. Machine learning is used to track, analyze, and respond to human emotions as an endeavor to actively monitor a person's moods or "mimic a human therapist's interactions with a patient" (Nogushi, 2023). AI represents a potential mechanism to overcome typical financial and

logistical barriers to care (Nogushi, 2023), which is important, especially considering the fact that some may be unwilling to express their feelings and emotions in a face-to-face mode (Denecke et al., 2021). Thus, the promise of chatbots to enable organizational leaders to support employees with mental health concerns is significant, as the technology can overcome many common barriers to mental health care.

3.3.1. The Wysa chatbot

Wysa is an emotionally intelligent chatbot that tracks people's moods and detects whether moods are negative or positive (Inkster et al., 2018) and is a tool for supporting mental health and well-being (Malik et al., 2022). The responses generated by Wysa are based on a database of prewritten responses generated by psychologists trained in cognitive behavioral therapy (Nogushi, 2023). Thus, Wysa provides tailored responses and evidence-based interventions that promote mental health support for users. As a virtual coach, Wysa engages several techniques, tools, and resources, ranging from cognitive-behavioral techniques (CBT) and meditation to cultivate a safe space where emotions can be freely expressed (For Her, 2021).

Wysa reacts in a personalized way and may suggest taking diagnostic tests. Based on assessments of emotional health (e.g. depression, anxiety), Wysa can provide further support, guidance, or recommendations for help-seeking from a professional. Wysa provides a sense of community by providing support and guidance for users. This technology is trained to assess online activities such as particular motions, facial expressions, phrasing, and vocal tones and has been identified as generally good at identifying and labeling emotions with accuracy.

Wysa provides care for individuals. The chatbot provides equitable access to care for all users, and emerging research suggests that Wysa is helpful for mental health (Inkster et al., 2018); however, there are some concerns pertaining to its effectiveness. Research is generally limited, and although the algorithms are impressive, they do not yet demonstrate a capacity to emulate empathy or further complexities of human emotions, which yields a common concern with AI systems: conversations lack depth and a sense of genuineness (Nogushi, 2023). From this perspective, there is a DEI risk: AI-driven therapy may yield a general distrust and avoidance of mental health interventions (Nogushi, 2023), thereby limiting its capacity to deliver equitable care for individuals who require it.

4. Developing chatbots to support a DEI ideology

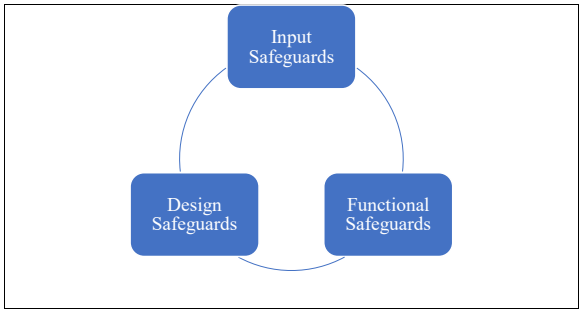
Based on the foregoing discussion, there is an opportunity to leverage the potential of chatbots and propose ways that remedy or minimize their DEI limitations. One clear learning from existing use cases of chatbots is that the creation of boundaries is useful to prevent inappropriate or harmful interactions, including those that stem from "bot-shit" (Hannigan et al., 2024). In this section, we propose a framework of three DEI safeguards: input, functional, and design safeguards (see Figure 1). These safeguards are connected and mutually supportive of each other. They can help to assess the extent to which a chatbot is poised to support or detract from DEI efforts. We suggest that managers use this framework when considering integrating, developing, or launching a chatbot.

4.1. Input safeguards

Input safeguards refer to the data used to train chatbots, the ML models used to process inputs fed into the chatbots, and the adoption of a DEI lens within the human teams developing the chatbots. Considering these inputs is important as they can help develop resilient chatbots that filter out inappropriate inputs and mitigate a wide range of problems stemming from bias.

Careful attention pertaining to the data used to train the ML model on which chatbots are based is critical (Ebrahimi & Hassanein, 2021). Improperly trained ML models may result in the Bias In-Bias Out problem (Mayson, 2018), in which a ML model repeats what is fed into it without understanding its meaning or its implications. If the initial data is biased or not representative enough of the targeted users, this may lead to an imbalance in application and accessibility for some groups over

Figure 1. Framework of chatbot DEI safeguards



others (Ajunwa, 2020) and may ultimately disadvantage or exclude particular groups (e.g., based on gender, race, sexual orientation, age, or socioeconomic factors; Holweg et al., 2022). This has clear DEI implications in which biased data and an improperly trained AI model could lead to a wide range of DEI transgressions. We note that input safeguards do not necessarily suggest a decrease in the volume of data on which to train the ML model. For instance, while limiting the sources of data input may have improved the outcome with Tay, the opposite is true for Luda: broadening the sources of input may have created a less harmful chatbot.

With respect to the teams behind the development of chatbots, our research echoes past research, which highlights the need for a DEI lens within the data scientist teams that decide on the ML algorithms to be employed in the chatbot as well as the comprehensiveness of the data that trains and validates the model (e.g. Yarger, 2020). Without a clear effort to embrace diversity within teams and the data, the output of such chatbots will inevitably be at odds with organizational DEI goals. For example, unverified reports indicate that 90% of the Luda programming workforce was male (Kwon & Yun, 2021), which may have contributed to the negative outcomes from the Luda chatbot. Additionally, choosing the right machine learning model that can screen for topics and statements that are outside the boundaries of the chatbot and its intended purpose will help facilitate positive DEI outcomes. Continuous monitoring of chatbots is also necessary for chatbots to keep up with the evolution of organizations' DEI initiatives. A key success factor in Microsoft's Cortana, for example, has been the continuous development and improvement of the bot, which has included consultation with a range of stakeholders. Multiple perspectives and a long-term approach will serve managers well.

4.2. Functional safeguards

Functional safeguards refer to limitations to the range of interactions that are possible, both by limiting conversation topics and target users. By limiting the scope of a chatbot's functionality to carefully defined users and topics, the risk of DEI transgressions is reduced. As the cases above illustrate, from a DEI perspective, chatbots are most beneficial when designed to perform a specific function for a specific audience (e.g., CiSA, FarmChat, Consejero Automatico). In contrast, a lack of chatbot functional or user limitations makes it more difficult to prevent DEI offences (e.g., Tay, Luda). Without functional boundaries, the wide range of unanticipated possibilities for interactions with chatbots pose a significant DEI risk, as chatbots are not yet designed to handle general conversations in all contexts without the risk of problems or mistakes. Any concrete exchanges and conversations with people on sensitive topics can have dangerous implications when these applications are misused. For the developers, rather than program chatbots to allow conversations for which the bot has been poorly prepared, programming the bot to simply state that it cannot answer the questions is a potentially useful alternative.

It is also important to keep humans in the loop as soon as the chatbot develops an undesirable outcome or cannot understand the other conversing party. A good strategy would be to provide the audience with a user-friendly transition to human agents when the chatbot is unable to address a user's query or when a more personalized interaction is required (Vassilakopoulou et al., 2023). Such a practice would align with best practices related to establishing human-AI symbiosis by leveraging both human strengths and technology (Jarrahi, 2018). In general, there is a burgeoning opportunity for chatbots to be designed with the aim of augmenting human abilities. With this premise, two broad AI applications are relevant: automation and augmentation; *automation* describes replacing human completion of tasks, whereas *augmentation* entails humans working with chatbots to complete tasks (Daugherty & Wilson, 2018). AI systems must be designed with the aim of augmenting and supporting humans rather than replacing them, so having functional safeguards in place ensures human-AI symbiosis. Even though AI has been shown to be superior in comparison to humans in achieving complex tasks and processing a large amount of data at a high speed, it nevertheless has limitations and lacks reflexive thinking. Therefore, it is important to design AI-based chatbots with the aim of augmenting humans to allow for a symbiotic relationship between humans and AI and to aid in the decision-making process.

Users should also fully grasp the intended function of particular chatbots, and any misunderstanding may propagate misuse or inappropriate expectations, interpretations, and dependencies. The cases above demonstrate that chatbots that are designed to serve a non-specific user base—for example, the general public—may lack the ability to incorporate complex moral reasoning and understand the nuances of human values, resulting in unfavourable outcomes. This is

similar to the concerns associated with crowdsourcing from the general public, and the range of potential issues that arise when doing so (Wilson et al., 2017). In cases where managers are implementing a purpose chatbot for use by the general public, we recommend that managers test the chatbot across multiple platforms (e.g. organization's website, social media, messaging apps) to reach the chatbot with a broader audience with different cultural norms and provide a diversified and inclusive avenue to monitor the chatbot performance.

4.3. Design safeguards

Design safeguards refer to considerations with respect to the chatbot design and conversational tone. Such safeguards include deliberate attempts to design chatbots in ways that do not perpetuate stereotypes. There is a sizeable body of research that explores DEI issues stemming from the fact that the majority of customer service chatbots are gendered as women (Feine et al., 2020), with feminine stereotypes used in their consumer-facing interfaces (i.e. conversational tone, avatars, language use) (Bastiansen et al., 2022) and design (Brown, 2023). Research shows that AI is subject to gender stereotyping (Dastin, 2018), and the harmful effects of gender-based stereotyping at the societal level (Correll et al., 2020) can be transferred and perpetuated by chatbots (Costa & Ribas, 2019); however, theory and results that support this notion are inconsistent (Bastiansen et al., 2022).

Prior research has highlighted concerns pertaining to misuse and the possible propagation of social injustices and inequality (Ågerfalk, 2020), including concerns related to implicit bias (Gupta et al., 2021). Some have noted that gendering chatbots as women is a practice that objectifies women (Borau et al., 2021) and extends a stereotype of women being in subordinate roles (Kelly, 2016). Interestingly, the benefit of assigning chatbots as women is not unequivocally established, which raises questions as to why the majority of service chatbots are gendered as women. Research suggests that users prefer chatbots that are self-congruent (i.e. male customers prefer masculine chatbots, whereas female customers prefer feminine chatbots, and gender-neutral customers prefer neutral chatbots) (Zogaj et al., 2023). Ultimately, the default version of a chatbot does not need to be a "woman." It is important to prioritize user experience and provide personalization options for the design of the chatbot with a focus on user needs and preferences.

Further, managers who are integrating chatbots into their organizations should closely monitor how their chatbot communicates with users and should consider putting boundaries in place. This includes minimizing conversational tones that create harmful affective connections. In the example of Eliza, communication barriers could have been helpful in mitigating harmful user outcomes. Chatbots are not capable of expressing empathy and do not possess an understanding of the language that they are producing. Thus, this safeguard has implications for both developers and users.

5. Conclusion

As DEI efforts persist in organizations, it is important for organizational leaders to approach the ensuing opportunities intentionally, including thoughtful considerations for emerging technologies in workspaces. Chatbots represent a unique opportunity to further engage customers, support employees, and cultivate an equitable and inclusive environment, yet without a DEI lens, outcomes from chatbot interactions can be harmful. Biases in AI algorithms and a lack of integration of ethical values in AI systems are serious threats, and the nature of the technology makes it challenging to assign responsibility and address any adverse implications. Although chatbots are highly sophisticated, and managers may be tempted to use them as widely and freely as possible, it is important not to overestimate the potential of a chatbot; there may never be a case whereby chatbots can completely replace humans without some unintended consequences. To that end, we present a framework of safeguards that can assist designers and managers in integrating a DEI lens into chatbots. When taken into consideration, our framework presents opportunities to push forward DEI when using AI chatbot technology. In conclusion, we have presented some of the existing and upcoming opportunities with chatbots. However, the DEI environment is in a state of growth. Even incorporating a DEI lens in the front end of chatbot development is insufficient to ensure continuous alignment with an organizational DEI ideology. Managers would be well advised to bear in mind that there is still much that we do not understand. As the technology develops and user behaviors and norms develop, it will be important to continue to monitor chatbots and their uses. This research is

one step forward; however, we call for future research around DEI in AI technology to responsibly and ethically advance the human-AI relationship and the integration of AI technology into businesses.

References

Ågerfalk, P. J. (2020). Artificial intelligence as digital agency. *European Journal of Information Systems, 29*(1), 1–8.

Ajunwa, I. (2020). The "black box" at work. *Big Data & Society, 7*(2). Available at https://doi.org/10.1177/2053951720938093

American Psychological Association (APA). (2017). Multicultural guidelines: An ecological approach to context, identity, and intersectionality. Available at http://www.apa.org/about/policy/multiculturalguidelines.pdf

Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior, 85*, 183–189.

Avula, S., Chadwick, G., Arguello, J., & Capra, R. (2018). SearchBots: User engagement with ChatBots during collaborative search. In *Proceedings of the 2018 conference on human information Interaction&Retrieval – CHIIR '18* (pp. 52–61). New Brunswick, NJ, USA: Association for Computing Machinery.

Bastiansen, M. H., Kroon, A. C., & Araujo, T. (2022). Female chatbots are helpful, male chatbots are competent? The effects of gender and gendered language on human-machine communication. *Publizistik, 67*(4), 601–623.

Black, J. S., & van Esch, P. (2020). AI-enabled recruiting: What is it and how should a manager use it? *Business Horizons, 63*(2), 215–226.

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 and Marketing, 38*(7), 1052–1068.

Brewer, J., Patel, D., Kim, D., & Murray, A. (2024). Navigating the challenges of generative technologies: Proposing the integration of artificial intelligence and blockchain. *Business Horizons, 67*(5), 525–535. https://doi.org/10.1016/j.bushor.2024.04.011

Brown, A. (2023, June 20). Brilliance knows No gender: Eliminating bias in chatbot development. *Forbes*. Available at https://www.forbes.com/sites/heatherwishartsmith/2023/06/20/calling-all-inventors-uspto-leaders-on-inclusive-innovation-as-a-driver-of-economic-growth/?sh=511243901c1a

Cachat-Rosset, G., & Klarsfeld, A. (2023). Diversity, equity, and inclusion in artificial intelligence: An evaluation of guidelines. *Applied Artificial Intelligence, 37*(1), 717–744.

Campbell, C., Sands, S., Ferraro, C., Tsao, H. Y. J., & Mavrommatis, A. (2020). From data to action: How marketers can leverage AI. *Business Horizons, 63*(2), 227–243.

Chang, L. (2016, February 6). Microsoft's Cortana gets sexually harassed, but she fights back. *Digital Trends*. Available at https://www.digitaltrends.com/cool-tech/microsoft-cortana-sexual-harassment/

Correll, S., Weisshaar, K., Wynn, A., & Wehner, J. (2020). Inside the black box of organizational life: The gendered language of performance assessment. *American Sociological Review, 85*(6), 1022–1050.

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

Cui, Y., van Esch, P., & Phelan, S. (2024). How to build a competitive advantage for your brand using generative AI. *Business Horizons, 67*(5), 583–594. https://doi.org/10.1016/j.bushor.2024.05.003

Dastin, J. (2018, October 9). Amazon scraps secret AI tool that showed bias against women. *Reuters*. Available at https://www.reuters.com/article/us-amazon-com-jobsautomation-insight/amazon-scraps-secret-ai-recruitingtool-that-showed-bias-against-women-idUSKCN1MK08G

Daugherty, P. R., & Wilson, H. J. (2018). *Human+ machine: Reimagining work in the age of AI*. Harvard Business Press.

Denecke, K., Abd-Alrazaq, A., & Househ, M. (2021). Artificial intelligence for chatbots in mental health: Opportunities and challenges. In M. Househ, E. Borycki, & A. Kushniru (Eds.), *Multiple perspectives on artificial intelligence in healthcare: Opportunities and challenges* (pp. 115–128). New York: Springer.

Ebrahimi, S., & Hassanein, K. (2021). Decisional guidance for detecting discriminatory data analytics recommendations. *Information & Management, 58*(7), 1–15.

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). A taxonomy of social cues for conversational agents. *International Journal of Human-Computer Studies, 132*, 138–161.

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2020). Gender bias in chatbot design. In *Proceedings of the international workshop on chatbot research and design: Lecture notes in computer science book series* (pp. 79–93). Cham: Springer. https://doi.org/10.1007/978-3-030-39540-7_6

Feng, M., Botha, E., & Pitt, L. (2024). From HAL to GenAI: Optimizing chatbot impacts with CARE. *Business Horizons, 67*(5), 537–548. https://doi.org/10.1016/j.bushor.2024.04.012

Ferraro, C., Demsar, V., Sands, S., Restrepo, M., & Campbell, C. (2024). The paradoxes of generative AI enabled customer service: A guide for managers. *Business Horizons, 67*(5), 549–559. https://doi.org/10.1016/j.bushor.2024.04.013

Ferraro, C., Hemsley, A., & Sands, S. (2022). Embracing diversity, equity, and inclusion (DEI): Considerations and opportunities for brand managers. *Business Horizons, 66*(4), 463–479.

Fitzpatrick, K., Darcy, A., & Vierhile, M. (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), e19.

For Her. (2021). Wysa chatbot. Available at https://www.forher.org.uk/apps/wysa-app

Gupta, M., Parra, C., & Dennehy, D. (2021). Questioning racial and gender bias in AI-based recommendations: Do espoused national cultural values matter? *Information Systems Frontiers, 24*, 1–17.

Hannigan, T. R., McCarthy, I. P., & Spicer, A. (2024). Beware of botshit: How to manage the epistemic risks of generative chatbots. *Business Horizons, 67*(5), 471–486. https://doi.org/10.1016/j.bushor.2024.03.001

Hashmi, N., & Bal, A. S. (2024). Generative AI in higher education and beyond. *Business Horizons, 67*(5), 607–614. https://doi.org/10.1016/j.bushor.2024.05.005

Heo, J., & Lee, J. (2019). CiSA: An inclusive chatbot service for international students and academics. In *Proceedings of the 2019 HCI international 21st conference: Late breaking papers* (pp. 153–167). Orlando, FL: Springer International Publishing.

Holweg, M., Younger, R., & Wen, Y. (2022). The reputational risks of AI. *California Management Review*. Available at

https://cmr.berkeley.edu/assets/documents/pdf/2022-01-the-reputational-risks-of-ai.pdf

Hunt, E. (2016, March 24). Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter. *The Guardian*. Available at https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

Inkster, B., Sarda, S., & Subramanian, V. (2018). An empathy-driven, conversational artificial intelligence agent (WYSA) for digital mental well-being: Real-world data evaluation mixed-methods study. *JMIR Mhealth and Uhealth, 6*(11), 1–14.

Jain, M., Kumar, P., Bhansali, I., Liao, Q., Truong, K., & Patel, S. (2018). FarmChat: A conversational agent to answer farmer queries. In *Proceedings of the association for computing machinery on interactive, mobile, wearable and ubiquitous technologies* (pp. 1–22). https://doi.org/10.1145/3287048

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons, 61*(4), 577–586.

Kelly, H. (2016, February 5). Even virtual assistants are sexually harassed. *CNN*. Available at https://money.cnn.com/2016/02/05/technology/virtual-assistants-sexual-harassment/index.html

Kietzmann, J., & Park, A. (2024). Written by ChatGPT: AI, large language models, conversational chatbots, and their place in society and business. *Business Horizons, 67*(5), 453–459. https://doi.org/10.1016/j.bushor.2024.06.002

Kwon, J., & Yun, H. (2021, January 12). AI chatbot shut down after learning to talk like a racist asshole. *Vice*. Available at https://www.vice.com/en/article/akd4g5/ai-chatbot-shut-down-after-learning-to-talk-like-a-racist-asshole

Laranjo, L., Dunn, A., Tong, H., Kocaballi, A., Chen, J., Bashir, R., Surian, D., Gallego, B., Macrabi, F., Lau, A., & Coiera, E. (2018). Conversational agents in healthcare: A systematic review. *Journal of the American Medical Informatics Association, 25*(9), 1248–1258.

Liu, W., & Pope-Davis, D. (2003). Moving from diversity to multiculturalism: Exploring power and its implications for multicultural competence. In D. B. Pope-Davis, H. L. K. Coleman, W. M. Liu, & R. L. Toporek (Eds.), *Handbook of multicultural competencies: In counseling & psychology* (pp. 90–102). Thousand Oaks, CA: Sage. https://doi.org/10.4135/9781452231693

Makasi, T., Nili, A., Desouza, K., & Tate, M. (2022). A typology of chatbots in public service delivery. *IEEE Software, 39*(3), 58–66. https://doi.org/10.1109/MS.2021.3073674

Malik, T., Ambrose, A., & Sinha, C. (2022). Evaluating user feedback for an artificial intelligence–enabled, cognitive behavioral therapy–based mental health app (Wysa): Qualitative thematic analysis. *JMIR Human Factors, 9*(2), e35668. https://doi.org/10.2196/35668

Mayson, S. (2018). Bias in, bias out. *The Yale Law Journal, 128*(8), 2218–2300. Available at https://www.yalelawjournal.org/article/bias-in-bias-out

Niu, B., & Mvondo, G. F. N. (2024). I Am ChatGPT, the ultimate AI Chatbot! Investigating the determinants of users' loyalty and ethical usage concerns of ChatGPT. *Journal of Retailing and Consumer Services, 76*(January), 103562.

Nogushi, Y. (2023, January 19). Therapy by chatbot? The promise and challenges in using AI for mental health. *NPR*. Available at https://www.npr.org/sections/health-shots/2023/01/19/1147081115/therapy-by-chatbot-the-promise-and-challenges-in-using-ai-for-mental-health

Northouse, P. (2018). *Introduction to leadership: Concepts and practice* (4th ed.). Thousand Oaks, CA: SAGE Publishing.

Okonkwo, C., & Ade-Ibijola, A. (2021). Chatbots applications in education: A systematic review. *Computers and Education: Artificial Intelligence, 2*, 1–10.

Osadchaya, E., Marder, B., Yule, J., Yau, A., Lavertu, L., Stylos, N., Oliver, S., Angell, R., de Regt, A., Gao, L., Qi, K., Zhang, W., Zhang, Y., Li, J., & AlRabiah, S. (2024). To ChatGPT, or not to ChatGPT: Navigating the paradoxes of generative AI in the advertising industry. *Business Horizons, 67*(5), 571–581. https://doi.org/10.1016/j.bushor.2024.05.002

Paschen, J., Wilson, M., & Ferreira, J. J. (2020). Collaborative intelligence: How human and artificial intelligence create value along the B2B sales funnel. *Business Horizons, 63*(3), 403–414.

Ramaul, L., Ritala, P., & Ruokonen, M. (2024). Creational and conversational AI affordances: How the new breed of chatbots is revolutionizing knowledge industries. *Business Horizons, 67*(5), 615–627. https://doi.org/10.1016/j.bushor.2024.05.006

Robertson, J., Ferreira, C., Botha, E., & Oosthuizen, K. (2024). Game changers: A generative AI prompt protocol to enhance human-AI knowledge co-construction. *Business Horizons, 67*(5), 499–510. https://doi.org/10.1016/j.bushor.2024.04.008

Roose, K. (February 3, 2023). How ChatGPT kicked off an A.I. Arms race. *The New York Times*. Available at https://www.nytimes.com/2023/02/03/technology/chatgpt-openai-artificial-intelligence.html

Sherry, C., Bhat, R., Beaver, B., & Ling, A. (2004). Students as customers: The expectations and perceptions of local and international students. In *Proceedings of the 2004 HERDSA conference: Research and development in higher education vol. 27: Transforming knowledge into wisdom holistic approaches to teaching and learning* (Vol. 27, pp. 1–16). https://doi.org/10.13140/RG.2.1.4984.6808

Sundberg, L., & Holmström, J. (2024). Innovating by prompting: How to facilitate innovation in the age of generative AI. *Business Horizons, 67*(5), 561–570. https://doi.org/10.1016/j.bushor.2024.04.014

Ulanoff, L. (July 24, 2016). Cortana awakens: The evolution of Microsoft's smart assistant. *Mashable*. Available at https://mashable.com/article/inside-microsoft-cortana

Vassilakopoulou, P., Haug, A., Salvesen, L., & Pappas, I. O. (2023). Developing human/AI interactions for chat-based customer services: Lessons learned from the Norwegian government. *European Journal of Information Systems, 32*(1), 10–22.

Vincent, J. (March 24, 2016). Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge*. Available at https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

Walker, M. (2001). Practical applications of the Rogerian perspective in postmodern psychotherapy. *Journal of Systemic Therapies, 20*(2), 41–57.

Wasserman, I., Gallegos, P., & Ferdman, B. (2008). Dancing with resistance: Leadership challenges in fostering a culture of inclusion. In K. M. Thomas (Ed.), *Diversity resistance in organizations* (pp. 175–200). New York: Taylor & Francis Group/Lawrence Erlbaum.

Williams, A., Kaur, H., Mark, G., Thompson, A., Iqbal, S., & Teevan, J. (2018). Supporting workplace detachment and reattachment with conversational intelligence. In *Proceedings of the 2018 CHI conference on human factors in computing systems* (pp. 1–13). Montreal, Canada: Association for Computing Machinery. https://doi.org/10.1145/3173574.3173662

Williams, D. H., & Shipley, G. P. (2021). Enhancing artificial intelligence with indigenous wisdom. *Open Journal of Philosophy, 11*(1), 43—58.

Wilson, M., Robson, K., & Botha, E. (2017). Crowdsourcing in a time of empowered stakeholders: Lessons from crowdsourcing campaigns. *Business Horizons, 60*(2), 247—253.

Wolbring, G., & Lillywhite, A. (2021). Equity/equality, diversity, and inclusion (EDI) in universities: The case of disabled people. *Societies, 11*(2), 1—34.

Wong-Villacres, M., Evans, H., Schechter, D., DiSalvo, B., & Kumar, N. (2019). Consejero automatico: Chatbots for supporting Latino parents' educational engagement. In *Proceedings of the tenth international conference on information and communication technologies and development* (pp. 1—5). Ahmedabad, India: Association for Computing Machinery.

Xiang, C. (March 30, 2023). "He would still Be Here": Man dies by suicide after talking with AI Chatbot, Widow Says. *Vice*. Available at https://www.vice.com/en/article/pkadgm/man-dies-by-suicide-after-talking-with-ai-chatbot-widow-says

Yarger, L., Cobb Payton, F., & Neupane, B. (2020). Algorithmic equity in the hiring of underrepresented IT job candidates. *Online Information Review, 44*, 383—395.

Yuan, C., Zhang, C., & Wang, S. (2022). Social anxiety as a moderator in consumer willingness to accept AI assistants based on utilitarian and hedonic values. *Journal of Retailing and Consumer Services, 65*, 102878.

Yuen, M. (2022). Chatbot market in 2022: Stats, trends, and companies in the growing AI chatbot industry. *Insider Intelligence*. Availabe at https://www.insiderintelligence.com/insights/chatbot-market-stats-trends/

Zogaj, A., Mähner, P., Yang, L., & Tscheulin, D. K. (2023). It's a Match! The effects of chatbot anthropomorphization and chatbot gender on consumer behavior. *Journal of Business Research, 155*(Part A), 113412.