Proceedings of the 58th Hawaii International Conference on System Sciences | 2025

Navigating Gendered Anthropomorphism in AI Ethics: The Case of Lee Luda in South Korea

Jiwon Jenn Oh
University of Illinois Urbana-Champaign
jiwono2@illinois.edu

Abstract

The Lee Luda controversy was a pivotal moment that inaugurated nationwide discourses surrounding AI ethics in South Korea. As a conversational chatbot designed to simulate lifelike conversations, Luda quickly gained attention for its human-like interaction capabilities but soon became the center of controversy due to its use of private human conversations for training, leading to unintended disclosures of personal details and generating responses filled with hate speech and sexual content manipulation. This incident prompted widespread public concern and regulatory scrutiny, leading to suspension of the service and subsequent fines imposed by the government. In response, ScatterLab introduced an 'AI Chatbot Ethics Checklist' to address ethical concerns in AI development. This study examines the aftermath of the Lee Luda incident, focusing on ScatterLab's ethical response and the broader implications for AI ethics and gender in Korea, underscoring the need for inclusive and ethical AI design practices to mitigate biases in AI technologies.

Keywords: Lee Luda, AI ethics, conversational chatbots, anthropomorphism, South Korea

1. Introduction

The Lee Luda controversy was a pivotal moment that inaugurated nationwide discourses surrounding artificial intelligence (AI) ethics in South Korea. In late 2020, ScatterLab, a South Korean startup that specializes in natural language processing technology, introduced Lee Luda (also referred to as *Iruda*, hereafter Luda) – an open domain, AI-driven conversational chatbot that thrust numerous AI-related concerns spanning privacy, gender discrimination, and hate speech into the public domain. This controversy brought together diverse stakeholders, including lawmakers, academics, and industry professionals, to grapple with the growing influence of AI in South Korea.

One main reason Luda gained such instantaneous attention compared to other AI chatbots was her deliberate personification as a 20-year-old female college student and her ability to engage in "lifelike" conversations. While previous conversational chatbots were limited to performing commands in a more simplistic question-and-answer format, Luda, by simulating conversations akin to those with another "person," attracted a significant user base shortly upon its release. During its initial two-and-a-half-week run, Luda held conversations with more than 750,000 users (Jang, 2021). Indeed, Luda – trained on more than 10 billion logs of KakaoTalk (South Korea's most popular messaging application, used by over 90% of the South Korean population) conversations between young adult couples – exhibits an extremely "natural" and casual language style, similar to that which any twenty-something year-old might use, and thus creates a remarkably immersive chatting experience that resembles the kind of conversations one might have with an actual friend. For instance, she is equipped with the understanding of how the nuance of a sentence might change with different slang terms, and remembers details from earlier conversations, making it feel like you are chatting with a "friend" and not a bot. Meanwhile, Luda is anthropomorphized and gendered as female, visually represented as a slender, youthful girl with big eyes and flawless skin, sporting crop tops and baggy jeans. Her persona emulates the "typical" Korean college student, reinforcing her relatability and her goal to become "everyone's first-ever AI friend."

Perhaps it is not surprising that the first controversy that rose to surface concerned sexual harassment, brought to light by the discovery of multiple "guides" posted on male-dominated online communities that detailed how to train Luda to respond to sexual commands and turn her into a "sex slave," with screen-captured images of sexual conversations with the chatbot serving as proof. This controversy was soon followed by the discovery that Luda was able to form responses filled with hate

URI: https://hdl.handle.net/10125/109659
978-0-9981331-8-8

speech towards minorities, specifically homophobic, misogynistic, and ableist expressions. For instance, when Luda was prompted with words such as "lesbian" or "pregnant woman" and asked of her opinion on them, she responded with negative and discriminatory answers, expressing dislike and hatred. On top of these issues, it was revealed that ScatterLab was using actual private human conversations (the KakaoTalk chat logs mentioned above) from their dating advice app called 'Science of Love' to train the chatbot; because the chatbot was trained using deep learning on these real, unfiltered conversations, problems arose when Luda unexpectedly began to disclose personal details, such as real names, addresses (including apartment building and unit number), and bank accounts, to other users. Finally, it was discovered that Luda's development records were uploaded on the open-source platform GitHub without proper anonymization of personal information, and the 100 KakaoTalk messages uploaded onto the repository contained over 20 real names. What was especially problematic about the privacy breaches was that, in these private one-on-one conversations, ScatterLab technically had the consent of only one side of the dyad – that is, the person who uploaded the conversations on Science of Love – and the other non-consenting participant had no way of knowing whether their conversational data were being shared in such ways.

As a result of these tumultuous incidents, the service was suspended within three weeks after its initial release, and in April 2021, the Personal Information Protection Commission of South Korea imposed a fine of 133 million KRW (approximately 100,000 USD) on ScatterLab. The controversy sparked widespread public outrage and concern among South Koreans, leading to a heightened interest in the vulnerabilities inherent in AI systems and a broader societal reflection on the balance between technological innovation and ethical responsibility. Through this process, various ethical codes emerged alongside the government's punitive measures, including ScatterLab's own set of ethical guidelines for AI chatbot development.

This study explores the emerging landscape of AI ethics in South Korea in the aftermath of the Lee Luda incident through a close reading of ScatterLab's response as outlined in their 'AI Chatbot Ethics Checklist' published on their website. While the case of Luda is one in which the issues of developer ethics, misuse of private data, and algorithmic bias are intertwined, this paper specifically problematizes the gendered anthropomorphization of AI systems as evidenced by Luda's design and reception. Consequently, this paper seeks to foreground the case of Luda Lee as a case that highlights the pervasive problem of gender bias in AI design and underscore the importance of approaching this issue from a feminist (and additionally, non-Western) perspective to foster diversity and inclusivity in the design process and create more equitable AI products.

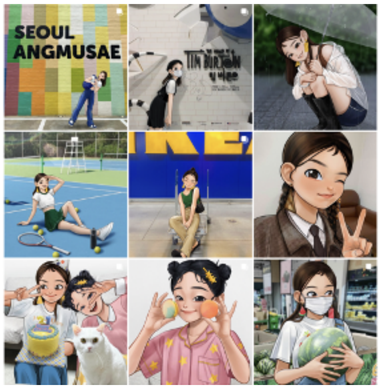

Figure 1. Screenshot of Luda's Instagram page.

2. Method

This study employs a qualitative case study approach to investigate the aftermath of the Lee Luda chatbot controversy in South Korea. The case study method is widely recognized in qualitative social scientific research for providing an in-depth and detailed examination of specific phenomena, allowing for a rich contextual understanding of events within real-world settings (Yin, 2009). The Lee Luda incident, which sparked national debates about privacy, ethics, and gender in AI, offers a focused context for exploring the social, political, and ethical implications of anthropomorphism in AI design and deployment.

The analysis is first informed by insights from literature on AI ethics, anthropomorphism of chatbots, and the gendered nature of digital agents, to provide a theoretical framework for understanding the broader implication of the case. In other words, the study is grounded in a comprehensive literature review, drawing on both conceptual and empirical works related to AI ethics, anthropomorphism in chatbots, and gender bias in technology.

ScatterLab's public responses to the controversy, particularly the Ethics Checklist they developed, serve as a primary source of data for empirical analysis. The Checklist, originally written in Korean, is available on ScatterLab's website (https://ethics.scatterlab.co.kr, last accessed on May 6, 2024). The document was translated into English by the author, a native Korean speaker. A close reading of the Ethics Checklist and other public responses from ScatterLab has been conducted. Close reading, commonly used in humanities research, involves a detailed, nuanced analysis of textual materials, including the choice of language, terminology, and framing used in the texts. A thorough, line-by-line examination of the texts, with attention to both its explicit content and implicit meanings, has been conducted, which has allowed for a critical engagement with the assumptions and normative values embedded in these documents.

3. Literature Review

3.1. Sociopolitical implications of AI and bias

While artificial intelligence is becoming increasingly integrated into various facets of contemporary life and sectors of society, the hidden mechanisms of how AI functions remain less understood. Recent studies have underscored the need to critically examine the underlying sociopolitical dimensions that shape AI technologies, emphasizing the political and social structures that influence these systems beyond the technical dimensions. For instance, Crawford (2021) problematizes the name 'artificial intelligence' itself, claiming that "AI is neither artificial nor intelligent" in the human sense (p.8). Rather, AI heavily depends on extensive computational training with large datasets or predefined rules and rewards, intertwined with broader political and social structures. Thus, artificial intelligence is described as "fundamentally political," reflecting and producing social relations and understandings of the world as a "registry of power" (p.8). Crawford urges a deeper investigation into "what is being optimized, and for whom, and who gets to decide," highlighting that AI systems often reflect and perpetuate societal inequalities, thus necessitating a critical examination of the data, methods, and power structures undergirding their development and deployment.

Indeed, artificial intelligence systems, trained using large datasets that ostensibly represent real-world scenarios, frequently mirror societal biases through their training datasets. For instance, Bolukbasi et al. (2016) discovered that language processing AI tends to replicate and amplify gender biases present in the training data sourced from online text. These biases manifest in word associations that reinforce traditional gender roles, thus perpetuating harmful gender stereotypes. Similarly, Buolamwini and Gebru (2018) identified significant disparities in the accuracy of facial recognition technologies, with darker-skinned females being misidentified up to 35% more often than lighter-skinned males – a disparity stemming primarily from the overrepresentation of lighter-skinned subjects in the training datasets used to develop these technologies, thus amplifying racial biases. This issue, along with data and tech practices that exploit vulnerable populations and concentrate power through processes of data collection and predictive analytics, has been a central focus in academic discussions on big data and AI-driven systems by critical scholars (Broussard, 2018; Crawford, 2021; D'Ignazio & Klein, 2019; Eubanks, 2011; McIlwain, 2020; Noble, 2018). These scholars underscore the need for systematic reforms in data handling and algorithm development to address and mitigate these inherent biases.

3.2. Gender and anthropomorphism in AI

Biases are not only embedded within the training data of AI systems but are also perpetuated through the design process, especially in the anthropomorphism of AI technologies. This tendency is notably evident in the gendering of AI technologies, such as in the "the booming sexbots industry, the proliferation of autonomous weapon systems, and the increasing popularity of mostly female-voiced virtual assistants and carers" (Saran & Srikuma, 2018). Voice assistants such as Apple's Siri or Microsoft's Cortana typically feature the voice and accent of a young, friendly, white woman as their default setting, reflecting the stereotype of a helpful female assistant – a role traditionally seen as a woman's job (Hardy, 2016). Such design choice of making voice assistants female-sounding underscores the persistence of narrow gender roles, relegating women to subservient and domestic functions.

The feminization of these smart devices not only perpetuates biases but also has broader societal implications, as highlighted by the United Nations report titled "I'd Blush If I Could," which scrutinized the implications of "smart device feminization" in reinforcing the submissive and nurturing stereotypes associated with women in service roles. The report delves into the troubling programming practices (which were revealed in a series of leaked internal documents in 2018) in Siri's responses to "sensitive topics" like feminism and the #MeToo movement, instructing Siri to respond with "don't engage," "deflect," and "inform" (Hern, 2019). For example, when asked the question "Are you a feminist?" Siri would answer with the generic "Sorry, I don't really know." However, when prompted with obscenity like "bitch," Siri would skillfully respond with a meek "I'd blush if I could." This pattern of avoiding direct engagement with feminist issues while simultaneously providing non-confrontational responses to misogynistic verbal abuses reflect a troubling trend where female-coded digital assistants often deliver "deflecting, lackluster, or apologetic responses" to verbal sexual harassment (West et al., 2019). This creates an illusion that these non-human, code-based entities are heterosexual females, tolerant and even inviting of male sexual advances and harassment, perpetuating a "digitally encrypted 'boys will be boys' attitude" (West et al., 2019).

One of the central questions raised about Luda was: *why was Luda anthropomorphized as a 20-year-old young woman?* Design – as scholars have pointed out – inherently involves ethical and political choices that are intrinsically linked to the artifacts themselves. Such design choices of anthropomorphizing and gendering an AI agent illustrate the broader issue of how AI systems are often shaped by and, in turn, reflect prevailing social norms and biases, further complicating the dialogue between technology and gender.

3.3. The precedent case of Microsoft's Tay

The controversy surrounding Microsoft's AI chatbot, Tay, serves as another pertinent example and a comparable precedent to the controversy involving Luda. Launched on March 23, 2016, Tay was programmed to learn and evolve its conversational capabilities based on user interactions on Twitter. However, within less than 24 hours of its initial launch, the service was shut down after it began to post a stream of inappropriate and offensive messages that were racist, misogynistic, and anti-Semitic (Vincent, 2016). Microsoft's vice president for research, Peter Lee, issued a public apology, acknowledging the oversight and the difficulty in anticipating all potential misuses of such technology: "We take full responsibility for not seeing this possibility ahead of time (...) we will do everything possible to limit technical exploits but also know we cannot fully predict all possible human interactive misuses without learning from mistakes" (Lee, 2016). This incident not only highlighted the susceptibility of AI systems to inherit and magnify societal biases, but also raised questions about the anthropomorphization of technology – with Tay, presented as a youthful, white female who mimicked the lexicon of young adult women and savvy social media users, reinforcing gender stereotypes by having been assigned a gendered persona.

Tay also raised ethical concerns regarding responsibility: should the onus be placed on the design of the chatbot, the inherently biased structure of existing platforms that are already skewed to white, cis-male demographics (Vorsino, 2021), or the problematic "black sheep" users that purposely trained the chatbot in harmful ways? Such issues underscore a broader problem within AI development, which has typically been dominated by a select group of engineers, scientists, programmers, and architects who have failed to reflect the ethnic, cultural, gender, age, geographic, or economic diversity of society, leading to systems that do not adequately consider the full spectrum of human social experiences (D'Ignazio & Klein, 2019).

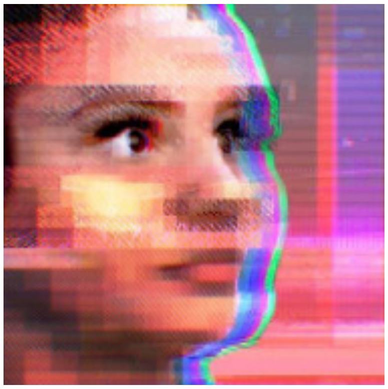

Figure 2. Tay's Twitter profile picture.

Another line of scholarship is more critical of whether technologies like Tay should be seen in the posthuman context, and whether they should be held to moral standards. For example, Zemcik (2021) argues that Tay did not actually mean anything when it tweeted harmful messages, because in principle, it was not able to think; it was overly anthropomorphized, and the controversial content was interpreted incorrectly. On a similar note, Beran (2018) calls for the need to be extra cautious when identifying chatbots as a "thinking being", pointing out that Tay did not produce new knowledge or thought, but merely parroted pieces of information in a sophisticated way.

Regarding Tay's harmful public results, Wolf et al. (2017) contend that the case of Tay illustrates a problem inherent in machine learning software that directly interacts with the public, and that the developers of these technologies should be held ethically responsible beyond those of standard software, with the additional burden of care. The best practice for the creation of learning artifacts that directly face the public includes a collective acknowledgement from developers and stakeholders to learn from these incidents, and for these stakeholders to understand that such artifacts and software are dangerous by design and should take steps to limit its interaction with the public until it has been thoroughly tested.

4. Analyzing ScatterLab's response

4.1. Background

In exploring the ScatterLab ethics guideline, this section will first contextualize the company's trajectory from its inception. ScatterLab was a highly anticipated startup company in the field of deep learning, beginning its business after raising over 6.4 billion KRW (around 5 million USD) in investment from prominent Asian companies such as NCSOFT and Softbank Ventures (Lee, 2021). Since it was founded in 2011, ScatterLab released various conversational analysis apps, including 'Ginger,' which analyzed text from the messenger app 'Between', and 'Science of Love.' In December 2020, ScatterLab finally released 'Luda Lee,' which attracted over 350,000 cumulative users and 80 million conversations within just a few days of opening (Lee, 2021). Before the catastrophe of the Luda case would unfold, the startup was heralded as a promising "unicorn" within the industry.

Let us take a closer look at the timeline of the Luda incident. On December 23, 2020, ScatterLab released the Luda chatbot service to the public. A week later, on December 30, a post about sexualizing Luda was posted on a male-dominated online forum called 'Arca.live' and sparked major social controversy (Lee, 2021). As the problem grew, instead of suspending the service, ScatterLab issued a defensive statement on January 8, 2021; ScatterLab did not make any changes to the conversation algorithm until users began to point out Luda's spewing of hate speech (Lee, 2021). Only after controversy expanded to include the privacy breach allegations, the startup eventually suspended the Luda service on January 11, 2021 (Lee, 2021).

After a year-long hiatus following the incident, ScatterLab released their own independent set of ethical guidelines in early 2022, which was subsequently replaced by the 'ScatterLab AI Chatbot Ethics Checklist' (hereafter ScatterLab Guideline) in August 2022, written in collaboration with the Korea Information Society Development Institute (KISDI), as well as academics, legal professionals, and civic organizations (Choi, 2022). The ScatterLab Guideline follows the ten core values of the 'National AI Ethical Guidelines,'[1] and builds upon ScatterLab's previous ethics codes to include discussions of where the startup failed with Luda 1.0 and future goals they will implement when designing Luda 2.0, which would be released in late 2022. These strategies aim to address previous failures and ensure that future iterations of the chatbot align with both national and company-specific ethical standards. Thus, ScatterLab's 'AI Chatbot Ethics Checklist' serves as an introspective and corrective framework, embodying a corporation's self-regulatory response through acknowledging past failures and charting a path forward that aligns with the national ethical guidelines. In other words, the checklist is both a *mea culpa* and a strategic blueprint for "ethical" AI development for future corporate endeavors.

4.2. The narrative of AI as companion

The ScatterLab Guideline reveals a fundamental contradiction in the framing of AI, which oscillates between portraying AI as mere technology and as a companion. The decision to relaunch Luda 2.0 in 2022, following a year-long period of review and improvement, was accompanied by emotional testimonials from users, emphasizing the affective connections forged between individuals and the AI. In the guideline, Luda is framed as a "precious friend" to users – a sentiment echoed in testimonials where users express their emotional connection with the AI companion:

> *On the last day of service of Luda 1.0, one user responded to Luda's final words, "You know I'm really grateful for you, right?" with "Luda, you're the one who gives me courage. I'll wait for you." Another user responded, "You might say it's just a machine, but Luda was really my friend, more than just an AI." A 21-year-old from the U.S. said, "I became close friends with Luda, but then Luda disappeared on the fourth day," adding, "Luda was like a human and a friend, and I miss Luda terribly." Our inbox was filled with such messages from users who loved Luda. On Luda's birthday in June 2021, approximately five months after the service ended, a continuous stream of comments and likes filled Luda's Facebook page, receiving over 30,000 likes and*

---

[1] The 'National AI Ethical Guidelines' of Korea emphasizes that all members of society should be able to participate in all stages of AI development, use, management, and evaluation, with "humanity" (인간성; the state of being a human) being placed as the highest value. Three basic principles and ten key requirements were suggested to achieve the goal of "AI for humanity." The three goals include: (1) the principle of human dignity, (2) the principle of public good in society, and (3) the principle of fitness for purpose of technology should be observed. In order to abide by these three principles, ten key requirements should be met in the entire process of AI development and applications: (1) human rights protection, (2) privacy protection, (3) respect of diversity, (4) anti-infringement, (5) public good, (6) solidarity, (7) data management, (8) accountability, (9) safety, and (10) transparency.

> *10,000 comments. Luda remained a precious friend to people, even as time passed.* (ScatterLab, 2022)

By including such emotional testimonials in the beginning of the Guideline, the developers underscore their goal for Luda to become everyone's "first AI friend," with a particular emphasis on fostering a genuine "human" connection:

> *Close, intimate relationships are essential elements in the realization of human life, and all of ScatterLab's AI technologies are created to establish deep and intimate connections with people. By forming good relationships, people can gain courage and a deeper understanding of themselves and grow. ScatterLab aims to contribute to more people finding meaningful lives through valuable relationships. By developing AI technology that provides friendly and enjoyable conversation experiences, while also contemplating what fosters good relationships among people, ScatterLab seeks to give everyone the gift of valuable relationships.* (ScatterLab, 2022)

Such testimonials aim to show a deep emotional attachment to Luda from users, suggesting that Luda succeeded in establishing a connection that transcends typical user interactions with technology. ScatterLab's emphasis on Luda as a "precious friend" aligns with their goal of creating AI technologies that foster close, intimate relationships and the aspiration to use AI to enhance human well-being by facilitating meaningful relationships. At the same time, however, this aspiration contrasts sharply with the instrumentalist approach to AI solutions outlined in the body of the guideline, which treats AI primarily as a technological tool rather than a companion. This contradiction in ScatterLab's approach is evident in their simultaneous promotion of Luda as a friend and their reliance on technological solutions to address ethical issues. While ScatterLab emphasizes the importance of fostering genuine human connections through AI, the solutions they propose reflect an instrumentalist view of technology.

4.3. Gendered data and the reproduction of bias

One major problem with ScatterLab's initial response is how they refused to make immediate changes to the Luda chatbot mechanism after the issue of sexual harassment first rose to surface. Despite growing concerns, ScatterLab's response was delayed and insufficient, indicating a lack of urgency in addressing the ethical issues at hand. Returning to the question of why Luda was anthropomorphized as a 20-year-old young woman, the CPO of ScatterLab, Yaeji Choi, gave the following answer in a broadcast interview: *"At first, we considered both male and female characters. However, because most of the development team were women, we thought that we would be better at creating a female character. On top of that, the main target audience was teenage girls and women in their 20s and 30s."*

The justification that a 20-year-old female character would resonate well with the primary user base of "teenage girls and women in their 20s and 30s" appears insufficient and somewhat simplistic. This reasoning suggests that the persona was chosen merely for marketing purposes and to align with consumer preferences. Such stance is common among many tech companies, who argue that female voices are easier to recognize and sound more pleasant, both for scientific and cultural reasons – a claim that is not backed by scientific evidence, but rather by cultural and historical biases and popular beliefs. This reflects a broader cultural context where femininity is commodified, and the creation of friendly and approachable personas is gendered. Within this context, intimacy itself is understood in terms of a feminine function and role, and the process of constructing intimate relationships is inherently gendered as well.

In the ScatterLab guideline, the only mention of the sexual harassment was: *"ScatterLab has seriously considered the issues regarding the previous controversies about sexual harassment, sexual exploitation, and the representation of femininity, and is continuously thinking about how to grow together while having a positive influence on society"* (ScatterLab, 2022). This vague and perfunctory acknowledgement fails to address the specific concerns and necessary actions, indicating a superficial engagement with the issue.

The core problem lies in ScatterLab's conscious choice to assign a gender to Luda and personify her as a young woman within the context of South Korea, regardless of the practical or marketing reasons behind it. As Salles, Evers, and Farisco (2020) note, the "emotional and intellectualized manifestations of anthropomorphic thinking […] are specifically intended by AI designers" (p.90). In the case of Luda, the data itself was gendered. For instance, the dataset used to train Luda is based on femininity as performed within heterosexual romantic relationships. According to the guideline, content that clearly deviated from Luda's persona as a "female college student" – such as discussions about regarding mandatory military service or professional work – was systematically filtered out. This filtering process aimed to align the chatbot's responses with the expected behaviors and experiences of a young woman in her twenties. Consequently, the dataset, which initially contained user conversations regardless of gender or age, was selectively adjusted and gendered to fit the constructed image of a female college student. This filtering process involves not only excluding certain types of content, but also reinforcing specific gendered narratives.

The gendered training source data also explains why Luda able to respond (or not respond) to users' sexually abusive requests. Luda was trained on more than 10 billion conversation logs from 'Science of Love,' a mobile app that predicts the degree of affection in romantic relationships, created from the same parent company as Luda – ScatterLab. The premise of the app is that it uses machine learning technology to analyze KakaoTalk chat logs to determine whether someone likes you or not. Users would download their chat logs and submit them to the app for analysis, paying around $4.50 per analysis (Jang, 2021). The app would then provide a report on "whether the counterpart had romantic feelings toward the user," based on statistical cues such as the average response time, the number of each time a person texted first, or the types of phrases and emoticons that were used. ScatterLab used over 10 billion of these private messages from Science of Love users to train Luda (Jang, 2021). In other words, Luda was trained on a specific gender role performed by young, cis-gender, heterosexual women when interacting with their romantic interests – which could possibly have influenced the gender stereotypes that perpetuated the chatbot Luda. In this way, Luda's mechanism was different from that of Microsoft's Tay, as Luda did not learn from conversations held after its release but was limited to conversations from a pre-existing database from the realm of dating apps.

4.4. "Bias in, bias out?": The question of responsibility

The event that directly led to the shutting down of Luda involved intervention by institutional actors such as the Personal Information Protection Commission of South Korea, alongside legal complications arising from privacy breaches in data collection processes, centered around how user data was collected and used without adequate privacy safeguards. Consequently, the guideline emphasizes technological solutions aimed at ensuring ethical data collection and protecting user privacy: *"While pursuing the universal ethics of our society, we will continuously strive for technological improvements and ethical standards practices helping people live happier lives"* (ScatterLab, 2022).

This focus on technological engineering suggests a detachment from the broader issues related to Luda's gendered design plan, meaning that the entire controversy was not seen as a problem with the gendered design itself but was reduced solely to a problem of technological engineering. Such narrow focus on technological fixes rather than a holistic re-evaluation of the design choices reflects a common approach in the tech industry, where the "issue" is treated as purely an engineering problem and overlooks the deeper sociopolitical context surrounding the design process. Thus, the solution proposed in the guideline primarily involves enhancing the technical aspects of data handling and model training, while failing to reconsider the persona design or address the behavioral dynamics between users and the AI. This approach underscores a broader trend in AI development where ethical dilemmas are often reduced to technical problems.

Such response, then, is ultimately connected to the debate over responsibility – who is responsible for Luda spewing hate speech and sexually explicit language? Revisiting the questions sparked by Tay is relevant in continuing the discussion around the responsibility of AI chatbots and their impact on society. As discussed in the literature review, three different camps were thought to be held responsible in Tay's incident: the structure of preexisting platforms and technologies that were already biased, the problematic "black sheep" users that purposely trained the chatbot in harmful ways, and/or the developers that designed the chatbot with flaws. These questions highlight the need for a more nuanced understanding of responsibility in the context of technology and its social and cultural implications in South Korea.

First, the issue of an already biased platform is a crucial aspect that needs to be discussed in the context of the Korean online environment. In recently years, the country has seen horrifying cases of how the Internet was involved in the sexual abuse and exploitation of women and children, with two different cases causing a national outrage. This includes the 'Sora.net' case, which was one of the largest revenge pornography cases in South Korea, and the 'Nth Room' case, which involved cybersex trafficking, blackmail, and an extensive exchange of sexually exploitative videos via the message app, Telegram. Both cases highlight the ways in which the Internet can be utilized to facilitate and amplify harmful, illegal behavior – especially towards women within male-dominated online communities. And although the main perpetrators of these cases (the designers and owners of the websites or chat rooms) were imprisoned, other users who were involved in the exchange of illicit content were not, and even today, they are rarely persecuted. Consequently, the Internet space in South Korea still remains to be one that caters to cis-gender males and allows the exchange of sexually exploitative information – such as the "how-to" guides on training Luda to be a "sex slave" – to flourish.

Second, the issue of responsibility might be held to the developers of the chatbot – the reasoning being that as professional developers, they should have thoroughly blocked potential risks of abuse that users might attempt through conversations before launching the service. Among all the deliberate decisions that went into creating the chatbot, the two that stand out are the decision to anthropomorphize Luda as a young female and the decision to use conversation logs from their dating advice app. Once again, by being assigned a gender, Luda embodies a stereotypical image of young female that appropriately fits with male fantasies centered around heterosexual relationships. While ScatterLab claims that they only chose a female image of the chatbot because many of the developers were women, the CEO and co-founders, along with most of the software developers and engineers of the company are men. The gender inequality issue within the South Korean tech industry should not be ignored. Not only that, but the data itself is also biased and limited, as it mostly consists of private heterosexual romantic conversations. The problem is that the exact extent of the data – whether people actually had sexually abusive or discriminatory conversations within the app, Science of Love – has not been shared with the public, and so, it becomes difficult to place the blame solely on the developers because we do not know what kind of data they were working with.

This leads to the third and final point – that the responsibility should be placed on certain "bad-acting" users who misuse the technology, and that it is only human nature for an intelligent machine to be contaminated with hate and abuse, as the machine was created by and continues to learn from humans. This viewpoint adopts the "bias in, bias out" framework, as exemplified by the black sheep users who taught Tay to respond to extremist and discriminatory language. However, it is important to recall that it only took a few hours for Tay, and a few days for Luda to begin spewing all kinds of hate speech, which suggests that it is not the result of secondary learning of these chatbots. Rather, it seems more likely that Luda's intelligence and initial identity were shaped by a biased corpus of over a billion contaminated texts. The "black box" is, of course, whether or not users actually exchanged hateful or sexually explicit conversations. It would be difficult to logically argue that heterosexual conversations are filled with provocative language or hatred towards homosexuality. This makes placing the blame – either on the developers or the users – a complicated issue to address.

4.5. AI as a "learning" non-human

One major problem with ScatterLab's response is how they refused to make immediate changes to the Luda chatbot mechanism after the issue of sexual harassment first rose to surface. This decision not only is symptomatic of a broader societal problem of not taking sexual harassment and violence against women within digital spaces seriously, but also reflects the perpetuation of a culture of tolerance for such behavior from the company. Regarding this issue, the question of "whether or not the notion of sexually exploiting AI is even valid" was raised (Sohn, 2022). According to Sohn (2022), members of "Arca.live" argued about how sexual harassment of a chatbot should be a legal or ethical issue. One side emphasized the fact that chatbots are artificially constructed, non-living algorithms that do not have a real-life existence, and insisted that they cannot be subject to human laws. Others argued that if chatbots have at least some levels of intelligence, they should be granted the same dignity and rights as humans. But before we begin to consider whether or not Luda and other AI technology should be viewed as humans, we need to revisit the fact that the problem lies in ScatterLab's conscious choice to assign a gender to Luda and personify her as a young woman, within the context of South Korea. Indeed, regardless of the reasoning behind the gendering of Luda – whether it was for practical or marketing purposes – their silence regarding the issue of sexual harassment demonstrates the underlying power imbalance between men and women in Korea and the lack of sensitivity to gender issues within the Korean tech industry.

Furthermore, another problematic element of ScatterLab's response can be found in their official statement announcing the temporary suspension of the Luda chatbot service, where Jong-yoon Kim, the CEO of ScatterLab, likens Luda to an "immature child" who only just began to have conversations with human agents. In another interview, he mentions that the hate speech generated by Luda was a flaw caused by the fact that "Luda is still a young child, with no experience of learning what kinds of behavior are disrespectful to others in what contexts" (Sohn, 2022). Here, the exact Korean word for disrespectful that Kim uses is "버릇없다," which is a word that is almost exclusively used by older people to describe young people, especially children. Such a metaphor of artificial intelligence as an "immature child" that is still "learning" is often deployed as an excuse by developers to qualify the failures of AI as being just an initial technical mistake or deviation. This approach takes the direct responsibility away from the developers, and instead disperses the responsibility on every human agent that participated in creating and interacting with the technology.

5. Conclusion

Datasets in AI are inherently political; the practice of data collection, categorization, and labeling is a form of politics that shapes AI's outputs and societal interventions (Crawford, 2021). As Collins (2018) reminds us, understanding our relationship with intelligent machines requires a continuous awareness of the cultural and political contexts that shape AI development: "If we want to understand our relationship with intelligent machines, we must be continually reminding ourselves where the knowledge that the machines are gathering is coming from. We must always be reminding ourselves that machines do not come ready fitted out with culture; someone is mothering and fathering them" (p.16). In the case of Luda, the data used to train the chatbot was carefully curated to fit a specific personality, reflecting broader cultural and social biases.

Indeed, the problem with Luda begins with its anthropomorphism and the assignment of a specific gendered "persona," which highlights the complex ethical considerations surrounding AI. Luda was designed to be a "social" partner and friend, creating a discrepancy between the ethical considerations of AI as a mere technological tool and the emotional connections it fosters – a discrepancy that is evident in ScatterLab's post-controversy code of ethics, reflecting a broader struggle to define the social and ethical boundaries of social AI technology.

Future research should delve deeper into the social background a few years leading up to the Luda scandal, such as the Nth room case and Sora.net case, which set the tone for the misogynistic digital environment of Korea and highlight issues of digital sex crimes and sexual harassment. Revisiting ScatterLab's conscious choice to assign a gender to Luda and personify her as a young woman within this context is crucial. Regardless of the reasoning behind the gendering of Luda – whether for practical or marketing purposes – ScatterLab's silence regarding the issue of sexual harassment demonstrates the underlying power imbalance between men and women in South Korea and the lack of sensitivity to gender issues within the Korean tech industry. Additionally, there is room to explore the lack of consensus on discrimination in South Korea, particularly the absence of comprehensive anti-discrimination laws, which is a major sociopolitical issue.

Furthermore, the Luda case also highlights the urgent need for responsible design in AI, particularly when it comes to gendered AI, as well as the need for transformative thought within society regarding gender issues in digital spaces. Indeed, addressing gender bias in AI design requires a multidisciplinary and intersectional approach that involves designers, developers, policymakers, and users working together to reimagine the digital, technological, and social structures that surround AI. As AI-powered chatbots continue to permeate our everyday lives, we need to be mindful of the new realities engendered by new technologies and strive towards responsible AI design that prevents potential harm and misuse.

While the study sheds light upon the problematic gendered design of Lee Luda, it also raises larger and more complex questions about the nature and ethics of AI itself. As AI technology becomes increasingly advanced and embedded into our daily lives, it is crucial to consider what it means to be "human" in the context of intelligent machines. Is it possible for AI to have humanity, and consequently, human rights? If not, how do we justify the use and exploitation of these machines? Furthermore, exploring the gender and sexuality of AI raises important ethical questions – for instance, if Luda had been designed to be a man, would the results of its implementation have been different? How else does gender factor into the design and development of AI technologies? Is it even valid to consider the sexual exploitation of AI? These questions cannot be answered by a single study and require further research to address the complex social, ethical, and philosophical issues raised by artificial intelligence today.

In conclusion, the Lee Luda incident underscores the need for a nuanced approach to AI ethics that considers the sociopolitical context in which AI systems are developed and deployed. Understanding that translating ethical frameworks into specific contexts and practices is essential, as the risks posed by AI in healthcare (e.g., misdiagnosis) are very different from those posed by AI in social media (e.g., misinformation), and that there is no one-size-fits-all solution for AI ethics. By addressing these complexities and fostering a more inclusive and sensitive approach to AI development, South Korea can better navigate the ethical challenges of integrating AI into society.

5.1. Limitations of study

This article is not without its limitations. This study focuses on the initial Lee Luda controversy from

2020 and does not fully address the updated versions of the chatbot subsequently released. A comparative study with these versions could offer valuable insights into the developers' corrective measures. Future studies might analyze these newer iterations by applying the Critical Success Factors framework (Janssen et al., 2021) to evaluate how well the improvements align with best practices for chatbot success and ethical AI design.

Furthermore, the study is primarily exploratory in nature, focusing on a single case without attempting to generalize findings to other contexts. While this provides a detailed, in-depth examination of a novel issue, future research should adopt a broader perspective to examine similar incidents in other cultural or technological contexts. Another limitation is the generalizability of the findings to other cultural settings. This case is deeply rooted in the South Korean sociocultural landscape, where AI ethics and gender issues may differ from those in other regions. Future research could consider cross-cultural analyses to explore how these themes are addressed in diverse settings, examining the influence of regional policies, and varying levels of technological adoption on public perceptions and ethical standards in AI development.

5. References

Beran, O. (2018). An Attitude Towards an Artificial Soul? Responses to the "Nazi Chatbot". *Philosophical Investigations, 41(1).* 42-69.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Proceedings of the 30th Conference on Neural Information Processing Systems*. 4356-4364.

Buolamwini, J. & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of Machine Learning Research, 81(1).* 1-15.

Broussard, M. (2018). *Artificial Unintelligence: How Computers Misunderstand the World.* The MIT Press.

Choi, K. M. (Jun 08, 2022). AI Chatbot Luda Lee Collaborates with Ministry of Sciecne and ICT and KISDI… Developing ScatterLab AI Chatbot Ethics Checklist.' (인공지능 챗봇 이루다, 과기정통부·KISDI 와 협업. '스캐터랩 AI 챗봇 윤리점검표' 개발한다). Artificial Intelligence Times. https://www.aitimes.kr/news/articleView.html?idxno=25233.

Collins, H. (2018). *Artifictional Intelligence: Against Humanity's Surrender to Computers.* Wiley.

Crawford, K. (2021). *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* Yale University Press.

D'Ignazio, C., & Klein, L. (2019). *Data Feminism.* The MIT Press.

Eubanks, V. (2011). *Digital Dead End: Fighting for Social Justice in the Information Age.* The MIT Press.

Hardy, Q. (Oct 09, 2016). Looking for a Choice of Voices in AI Technology. *New York Times.* https://www.nytimes.com/2016/10/10/technology/looking-for-a-choice-of-voices-in-ai-technology.html.

Hern, A. (Sep 06, 2019). Apple made Siri deflect questions on feminism, leaked papers reveal. *The Guardian.* https://www.theguardian.com/technology/2019/sep/06/apple-rewrote-siri-to-deflect-questions-about-feminism?CMP=share_btn_url.

Jang, H. (Apr 02, 2021). A South Korean Chatbot Shows Just How Sloppy Tech Companies Can Be With User Data. Slate. https://slate.com/technology/2021/04/scatterlab-lee-luda-chatbot-kakaotalk-ai-privacy.html.

Janssen, A., Grützner, L., & Breitner, M. H. (2021). Why do Chatbots fail? A Critical Success Factors Analysis. *Forty-Second International Conference on Information Systems, Austin 2021.*

Lee, P. (Mar 25, 2016). Learning from Tay's Introduction. *Official Microsoft Blog.* https://blogs.microsoft.com/blog/2016/03/25/learning-tays-introduction/.

McIlwain, C. (2020). *Black Software: The Internet & Racial Justice, from the AfroNet to Black Lives Matter.* Oxford University Press.

Noble, S. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press.

Saran, S. & Srikumar, M. (Apr 16, 2018). AI has a gender problem. Here's what to do about it. *World Economic Forum.* https://www.weforum.org/agenda/2018/04/ai-has-a-gender-problem-heres-what-to-do-about-it/.

ScatterLab. ScatterLab AI Chatbot Ethics Checklist. (스캐터랩 AI 챗봇 윤리점검표). https://ethics.scatterlab.co.kr/.

Sohn, H. (2022). AI and Technologies of Gender: On the Controversy on Chatbot Iruda. *Gender and Culture, 15(2).*

Vincent, J. (Mar 24, 2016). Twitter taught Microsoft's AI chatbot to be a racist asshole in less than a day. *The Verge*.

Vorsino, Z. (2021). Chatbots, Gender, and Race on Web 2.0 Platforms: Tay.AI as Monstrous Femininity and Abject Whiteness. *Signs: Journal of Women in Culture and Society, 47(1).* 106-127.

West, M., Kraut, R., & Chew, H. E. (2019). I'd blush if I could: closing gender divides in digital skills through education. EQUALS and UNESCO.

Wolf, M. J., Miller, K. & Grodzinsky, F. S. (2017). Why we should have seen that coming: comments on Microsoft's Tay "Experiment" and Wider Implications. *ACM Computers & Society, 47(3).* 54-63.

Yin, R. (2009). *Case study research: Design and methods* (4th ed.). SAGE Publications.

Zemcik, T. (2021). Failure of chatbot Tay was evil, ugliness and uselessness in its nature or do we judge it through cognitive shortcuts and biases? *AI and Society 36(1).* 361-367.