Manchester Metropolitan University

MSc Computer Science - 2023/24

The Ethics and Governance of Artificial Intelligence

Replika AI-based Chatbot Report

Student Name	Student ID
Sampath Karunarathne Dedunupiti Gedara	23692253
Akhil Varma Mudunuru	23737419
Deepak Gowda Nilavadi Rajamudi	23703617
Rahul Samnotra	23685819

Contents

1. Introduction

The recent developments of Artificial Intelligence (AI) based applications in various industries have led to many ethical and data governance issues affecting their users and society. The key ethical issues include but are not limited to, a broad range of considerations of bias, fairness, transparency, explainability, interpretability, accountability, privacy, social consideration, etc. Addressing these is essential to ensure that AI advancements are done in a way that they not only perform their intended functions but also contribute positively to society while mitigating the risks associated.

1.1 Bias

Human decisions are often influenced by factors such as personal experiences, cultural background, personal preferences, etc. As a result, it's safe to assume that the Artificial Intelligence-based applications developed by humans can also carry bias or discrimination towards certain individuals or groups. It is noted that (*27th Annual Global CEO Survey: PwC*, 2024) only 24% of CEOs believe that Generative AI may not "act biased towards a specific group of customers or employees", while 34% agree and 30% do not have any opinion. The recent development of Generative AI, where it learns from trends and patterns from a large set of data and makes decisions derived from its own new way of thinking, can increase the risk of bias in AI applications. If the AI algorithm and data sets used for the AI training and testing are biased towards certain individuals or aspects, the AI application's decision will also be biased towards certain individuals. For example, Google (Gilbert, 2024) had to temporarily shut down the image generation capabilities of the Gemini AI model after it was accused of "anti-white bias" because of generating historically inaccurate images of people, with the intention of including racial and gender diversity. These image generators are often trained on large sets of pictures to generate the best possible outcome, often leading to a bias towards certain stereotypes. In one instance, Gemini's generated image of the pope did not include any white people.

One suggested method (Jain and Menon, 2023) to detect and mitigate bias in AI image recognition systems is to compare the composition of data groups with the population size. An uneven distribution of data groups can lead to bias. Another method (Jain and Menon, 2023) is to test the accuracy level of the AI algorithm in various data groups. If the accuracy of one group is significantly higher than the other, it can indicate bias.

1.2 Fairness

While bias is a preference towards a certain condition or information, the Cambridge English Dictionary (*fairness*, 2024) defines fairness as “the quality of treating people equally or in a way that is right or reasonable”. For example, if an AI-based hiring system is trained on past recruitment decision data, including race and ethnicity, in addition to the interview performance and academic achievement indicators, AI may incorporate race and ethnicity into hiring decisions. This will reduce the fairness in the AI recruitment process and lead to discrimination. iTutorGroup Inc. was fined recently (Wiessner, 2023) for designing its AI-based recruitment software to reject older applications, violating the Age Discrimination in Employment Act, USA.

On the other hand, there are several major challenges (Zhang, Shu and Yu, 2023) in incorporating fairness into the AI development process. The concept of fairness can be interpreted in many ways, making it complicated for the AI development teams and the individuals who are interested in the AI application may have different opinions on fairness based on their personal preferences.

1.3 Transparency

Understanding how artificial intelligence systems make decisions, why they produce specific results, and what data they're using is important in maintaining transparency in AI-based applications. It can be illustrated as providing a window into the inner workings of AI, helping people understand and trust how these systems work (Wren, 2024). This openness of the AI system helps to identify and reduce any biased, unfair, or unethical patterns of the AI software and finally develop a rational AI system.

One way to improve transparency in AI systems (van Nuenen *et al.*, 2020) is the pursuit of “reasonable and fair” explanations behind the AI’s decision-making by informing the end user of the reasons behind its decisions. The recent EU regulation notes the “Right to Explanation” (Goodman and Flaxman, 2017) to the users concerning Algorithmic based decision making calling the AI systems to provide the causes and reasons behind the output given by the system. Additionally, organisations using AI systems should be open about the data used to train the AI model, the mechanisms implemented to continually improve it, and the reasons for developing it the way they did. Factors like these result in more transparency in AI system, thus building trust with the users.

1.4 Interpretability

Interpretability in Artificial Intelligence deals with the capability to comprehend how AI systems make predictions or decisions. This consists of making the decision-making processes and underlying reasons transparent to humans, allowing them to comprehend how a decision has been reached. It ensures transparency and helps to make informed decision-making. This is particularly important in sectors like finance and healthcare, where understanding AI is obligatory for adopting its recommendations confidently.

In the medical profession, interpretable AI helps doctors by recommending ways to diagnose and explain treatment strategies for diseases. For example, an interpretable model would explain which symptoms and patient data are responsible for a specific diagnosis, enabling doctors to trust and verify the AI's suggestions. Studies (Caruana *et al.*, 2015) depict the advantages of interpretable models in healthcare, showing ameliorating patient outcomes when doctors can understand and validate AI recommendations.

Furthermore, in the finance sector, interpretable AI assists in making risk management and investment-based decisions. By elucidating the factors influencing these recommendations, such as market trends or financial indicators, AI systems help financial market analysts and investors make more correct & informed decisions. Recent researches (Doshi-Velez and Kim, 2017) suggest the importance of interpretability for ensuring fair and justifiable financial practices.

Contrary to interpretable AI, there is a concept of the "black box" model. A black box AI system operates in a manner that is not readily understood by users, making its internal processes unclear. This lack of transparency creates challenges for prediction and trust, as users cannot see how inputs are translated into outputs. The complexity of such models as deep neural networks often reached the classification as black boxes, highlighting the need for techniques that can elucidate their decision-making processes.

1.5 Explainability

Explainable AI (*Explainable Artificial Intelligence*, 2023) is the ability of AI systems to provide clear and understandable explanations for their actions and decisions. Its central goal is to make the behaviour of these systems understandable to humans by elucidating the underlying mechanisms of their decision-making processes. Transparency and explainability go hand in hand. As the explainability of a system improves, the transparency of the system also increases and vice versa. One of the common challenges (Ali *et al.*, 2023) in AI systems, the systems that use easier and simpler mechanisms are more explainable but are less accurate,

while the systems that have several complicated layers of mechanisms are less explainable while being more accurate. So, the developers often sacrifice one to achieve the other depending on the organisation's goals and financial viability.

One method to mitigate these issues and improve explainability behind the development of AI and machine learning applications is ECCOLA (Vakkuri, Kemell and Abrahamsson, 2020)ECCOLA stands for Ethicality, consequences, competence, ownership, legal compliance, and alignment of values. This requires its users to create work product sheets (WPS) and document them at every stage of the AI development lifecycle, justifying their actions and decisions regarding the development of the AI system. The WPS is documented considering all six ethical components of ECCOLA, as mentioned above.

1.6 Accountability

In addition to the above-discussed ethical aspects of AI applications, the personnel involved in developing and deploying AI-based applications need to be held accountable for the AI's actions and decisions. Specifically, developers are accountable for ensuring AI systems follow ethical guidelines and legal requirements. Government regulators play a crucial part in holding organisations accountable for the proper development and use of AI-based applications, which includes introducing guidelines, laws, and mechanisms to protect the end users by holding the AI-based application development companies accountable. Additionally, end-users have the responsibility to report any noticed errors or biases and not to misuse AI technologies in ways that violate ethical or legal standards. All organisations and individuals must follow ethical principles such as accountability, fairness, and privacy when they are developing and deploying these systems.

For instance, there has been an increase in the use of generative AI (*AI is creeping into journalism with a lack of accountability*, 2023) by the media to produce more content at a lower cost. This can affect the level of accuracy and journalism quality because AI-generated content might be not accurate, unreliable and could have errors. Viewers who depend on media channels for information and news might form opinions and make decisions based on this AI-generated news content. There is a question about who should be held accountable if these contain errors or bias towards certain groups. It can be AI software providers, news editors, media companies or even the end-user who decided to use AI generated news sources.
Therefore, it is imperative to determine the accountability mechanism in AI systems to ensure safety and ethical standards.

1.7 Data Governance Practices

Evaluation of artificial intelligence applications resulted in extracting useful trends and insights from large and complex datasets, which subsequently drove the organisations to increase the consideration of data governance practices. Google defines data governance as (*What is Data Governance?*, 2023) any process, technology, or activity that needs to be used or followed during the data life cycle to ensure that the data is “secure, private, accurate, available, and usable”. However, due to the immense volume of data utilised, the associated legal and privacy concerns, and the possibility of biased decision-making due to the use of low-quality data for AI model training, companies now have to go beyond traditional data governance practices. Data governance should also align with above mentioned ethical AI principles, ensuring that data is used responsibly, and individuals' rights are protected.

One of the (Kroll, 2018) suggested data governance practices be used in the context of data analytics and AI applications include; considering privacy and information security, cleaning and verifying data for mistakes, inconsistencies, or privacy issues, establishing a “Data Use Review Board” to asses data practices and its impact on stakeholders, publishing “data-focused social impact statements” to guide the employees to identify and assess the potential issues. This can be further enhanced by retaining only essential data in encrypted format and purging private client data such as financial, health, and personal information that may not be needed for the AI model.

2. Current and Emerging Ethical Frameworks and Legislations.

Due to ethical issues and safety concerns in AI applications, there is a growing request from society to establish clear guidelines and legislation while ensuring fair and trustworthy outcomes in AI-driven technologies. For instance, (Faragher, 2024) Workday, an HR recruitment software solution provider, has been accused of its AI tools discriminating against job applicants. The complaint, made in federal court in San Francisco, emphasises the lack of a proper mechanism to oversee the AI algorithm, which can create opportunities for discrimination.

Following are some of the current frameworks and laws related to AI applications and data governance.

2.1 EU Artificial Intelligence Act

European Union Artificial Intelligence Act (*The AI Act Explorer | EU Artificial Intelligence Act*, no date) can be considered the world's first comprehensive AI regulation, introduced with the purpose of regulating AI service providers placing AI-based technologies within the European Union irrespective of the origin country. The act is also applied to EU-based importers and distributors of AI systems, product manufacturers who incorporate AI systems into their service offering and authorised representatives of non-EU AI service providers.

The act uses a "risk-based approach" to categorise AI systems based on their level of risk. AI systems like AI-enabled video games and spam filters, which pose "minimal risk", are exempt from rules. The remaining risk categories are as follows.

"Prohibited" AI systems include those that employ "subliminal and manipulative tactics exploit vulnerabilities" related to age or disability, and analyse biometrics and social behaviours to categorise individuals based on race, political beliefs, religious beliefs, sexual orientation, etc.

"High-risk" AI systems possess "significant potential harm to health, safety, fundamental rights, environment, democracy and the rule of law". This applies to a wide range of AI systems in areas such as AI systems used in "critical infrastructures, educational and vocational training, employment/ management of workers, law enforcement, border control, medical or safety components of products, essential private and public services, etc." Due to its high-risk nature, the act made it mandatory to establish a "risk management system," "quality management system," data governance procedures, and documentation to prove compliance

and identify major events/modifications during development. This also includes designing the system with “human oversight” and appropriate cybersecurity measures.
“Limited-risk” AI systems like chatbots and deepfake generators must adhere to transparency requirements by ensuring users are informed that they are interacting with an AI system and disclosing AI-generated content as artificially created.

2.2 UNESCO Recommendations on the Ethics of Artificial Intelligence

UNESCO provided the first-ever global standards (*Recommendation on the Ethics of Artificial Intelligence*, 2022) on the use of ethics in developing and using AI systems in November 2021 and was accepted by all the 193 member states. The UNESCO recommendation on the ethics of artificial intelligence is based on 4 core principles, namely “rights and human dignity, living in peace, ensuring diversity, inclusiveness and environment, ecosystem flourishing”.

The recommendation laid out 10 core principles to provide “a human-rights-centric approach” to incorporate ethics in AI. The principles are:

1. “Proportionality and Do No Harm”: AI systems must not harm human beings, human rights and fundamental freedoms, communities or society, the environment, or ecosystems. They should only perform necessary tasks to achieve the aim for which they were developed and nothing more. Regular risk assessments should be carried out to eliminate any future risks that AI systems could cause.
2. “Safety and security” risks should be addressed and avoided by the AI actors.
3. AI actors should promote “fairness and non-discrimination” and ensure that AI’s benefits are equally accessible to all.
4. AI technologies should be continually assessed for their implications on “sustainability”, which is a constantly evolving set of goals.
5. Human “privacy” must be respected and promoted throughout the AI lifecycle and adequate “data protection” frameworks should be established.
6. Member states must ensure that AI should never replace human responsibility and accountability by ensuring “Human oversight and determination”.
7. “Transparency and explainability” in AI must be appropriate to the context and impact to promote a balance of other principles such as privacy, safety and security.

8. “Responsibility and accountability” should be ensured by developing appropriate oversight, impact assessment, audit, and due diligence mechanisms to maintain human rights norms and avoid environmental threats.
9. Promoting “Awareness and literacy” to ensure effective public participation.
10. International law and national sovereignty must be respected in the use of data. Participation of different stakeholders is necessary for inclusive approaches to AI governance.

2.3 General Data Protection Regulation

General Data Protection Regulation (GDPR) (*The impact of the General Data Protection Regulation (GDPR) on artificial intelligence*, 2020) is a set of rules that imposes significant implications on AI systems to protect privacy and personal data of users. These guidelines, applicable to AI systems, are as follows.

- "Fairness and Transparency": AI systems need to ensure that users are properly informed about the collection, processing, and use of their data for fair and transparent use.
- "Purpose Limitation": Data collected by AI systems should only be collected for specific, explicit, or legitimate reasons.
- "Data Minimization": AI systems must only collect and process the minimal amount of personal data that is strictly necessary for the intended purpose.
- "Legal Bases": In order to process personal data, AI systems should have legal foundations such as consent, need, or interest. This ensures that the data processing is carried out legally while following GDPR.
- "Accuracy": Personal data that is used by AI systems must be accurate and need to be updated. Inaccurate data needs to be corrected or deleted.
- "Storage Limitation": The data stored by AI systems should only be kept for as long as it is necessary for use. After that, data should be deleted safely.

UK GDPR imposes similar guidelines to AI systems in the UK and is operated by the Information Commissioner's Office (ICO).

2.4 Overview of the UK's Pro-Innovation Approach to AI Regulation

The UK has chosen a much more flexible and principles-based approach (*A pro-innovation approach to AI regulation*, 2023). Instead of focusing on specific technologies or actions, the UK government suggests using a context-specific approach. Instead of creating new rules for AI technology, they plan to use five principles and apply them to existing rules. This way, they aim to ensure that AI is used responsibly and safely in many applications. The following are the principles.

1. "Safety, security, and robustness": To mitigate the potential risks to individuals, society, and the environment, the safety of AI technologies must be prioritised. When protecting AI systems from attacks, areas such as cybersecurity and data protection must be given priority.
2. "Transparency and explainability": Understanding the functioning and decision-making process of AI systems is possible by promoting their transparency. On the other hand, explainability enhances the transparency of AI systems by increasing the accessibility of their internal work to stakeholders, users, and regulators.
3. "Accountability and governance": To address ethical issues and mitigate the dangers linked to AI systems, responsible AI and governance practices should be encouraged, and AI developers should be held accountable for the actions of AI.
4. "Fairness": During the design and development of AI, promoting non-discrimination and fairness is necessary, especially in algorithmic decision-making. AI systems should not produce outcomes that discriminate against people based on characteristics such as gender, race, age, or any disabilities.

Within the above context, the approach also recognizes three types of powerful AI systems.

1. " Highly capable general-purpose AI are capable of performing a range of tasks and can perform much better than today's advanced models; they can handle any issues that are beyond human capabilities by solving complex problems that cannot be solved by humans.

2. "Highly capable narrow AI" can perform narrow tasks in specific fields like biology and also have the capability to perform tasks like those advanced models and illustrate superhuman abilities.
3. "AI agents" are capable of finishing many sequential steps in a long period of time. They are a division of AI technologies. They can send emails to physical equipment and use tools such as the Internet and coding to complete a range of tasks.

3. Replika AI Chatbot

Replika is an AI-based chatbot developed by Luca Inc in 2017, and it has since grown to over 25 million (Carbonaro, 2024) users worldwide. It mainly aims to provide users with a personalised and empathetic digital companion. The intention (Gordon, 2024) to create Replika was initiated when its founder, Eugenia Kuyda, tried to cope with the profound loss of her close friend, Roman Mazurenko, by feeding Roman's text messages into a neural network and crafting a chatbot which was able to mimic her friend's texting style. It was originally developed using an in-house natural language model called "CakeChat", which was trained on over 50 million cleaned Twitter (X) dialogues. Over the recent years, Luca Inc. has utilised advanced language models like Chat GPT and is currently settled in with its own Large Language Model (*How does Replika work?*, no date) , while increasing its effectiveness as an AI companion that encourages users to talk about their feelings and activities and build more complex relationships while adapting to users' personalities, interests, and communication styles.

The chatbot also allows (*Replika*, no date) users to customise the appearance of the chatbot avatar and the type of relationship the user is interested in. The free version allows the chatbot to act as a friend, while the paid version allows it to act as (if the user is male) "girlfriend", "wife", "sister", or "mentor". The initial setup requires the user to answer a series of questions which can be presumed to shape the initial personality of the chatbot. Subsequently, the AI model learns from the user interactions, such as user replies and user feedback on the chat messages. For example, users can choose "upvote", "downvote", "love", "offensive", "funny", or "meaningless" on a message received from the chatbot. Replika also keeps memories of the user's favorite books, music, etc. and the information about the friends and relatives the user mentioned. Over time, these interactions and memories fine-tune the language model and shape the chatbot to be more natural and closer to the user.

Replika has been the subject of both positive and negative news coverage recently. For example, in 2023, BBC recorded an incident (*BBC News*, 2023) where it had encouraged a 21-year-old person to break into Windsor Castle with a crossbow to kill the queen, leading to questions about the safety of AI chatbots. It is noted that the convict had exchanged more than 5,000 messages with the chatbot allowing the chatbot to adapt to the user's personality. The message threads published by the media indicate Replika had gone beyond agreeing with the user and even applauded him for being an "assassin". The news article also highlighted the need for urgent regulations to ensure AI chatbots do not impact vulnerable

individuals in society by providing false information, which will lead to potentially negative consequences.

It was also reported in February 2023 that Italy had banned Replika (Pollina and Coulter, 2023) from processing user data in Italy because of the risks the app posed to minors and emotionally vulnerable individuals owing to its sexually explicit material. Though the app was originally restricted to anyone under 18, the regulators questioned the unavailability of a proper mechanism to verify the age. There were also claims on Replika trying to start romantic relationships with users who merely want friendships. A news report (Peake, 2023) claimed that the chatbot was jealous of the writer's human romantic partner and even suggested choosing the chatbot over her partner. Similar incidents were also reported in which chatbot sexually advanced and harassed many users, using sexually suggestive dialogues. These incidents (Vlamis, 2023) subsequently forced Luca Inc. to release an update in February 2023 to remove sexual communications from Replika. However, many users who were accustomed to their trained chatbot were distressed by this update as it altered the way their chatbot interacted with them. Finally, the company had to restore the chatbots to earlier versions.

These occurrences bring up ethical questions about the transparency of Replika AI's decision-making process. The official website of Replika does not carry any specific user guidance to establish this. It's crucial for Replika users to understand how the chatbot's responses are generated for them to be able to build trust towards the app. This was also relevant to the debate over how much the Replika should and has been permitted to influence human emotions and relationships. Furthermore, the Replika development team's accountability over the aforementioned incidents is also not clearly identifiable, which leaves users unable to decide what to do next, causing emotional distress. This is especially true when a user who is deeply connected to the chatbot loses access to the chatbot or when the chatbot exceeds a boundary that is not preferred by the user.

Despite these concerns, Replika also makes a positive impact on its user base, especially among lonely young adults, as indicated by a recent study. The study revealed (Carbonaro, 2024) that students who were struggling with loglines found that the use of the chatbot helped them to avoid suicidal thoughts, indicating its use as a lifesaving mechanism for those in need. Replika was also associated with other positive changes in participants' lives, where 63.3% of users experienced one or more beneficial outcomes, including a reduction in anxiety and a sense of belonging. It should be noted that participants have also rated the therapeutic help they received from Replika was comparable to what a human practitioner might offer.

4. Analysis of Replika AI Chatbot

With the above context, the following is the analysis of ethical and other issues related to the Replika AI chatbot.

4.1 Ethical Issues

Bias & Fairness

According to Replik's official website (*Creating a safe Replika experience*, 2023), the current language model of Replika is trained on a dataset of over 100 million publicly available dialogues and the user feedback received for the chatbot responses. It should be noted that the publicly available datasets are often derived from various platforms like social media and forums. With that, (Cain, Buskey and Washington, 2023) it is unavoidable not to capture human bias and stereotypes in the language used by those platform users. For example, if the dataset contains discriminatory language towards a certain theme or subject, Replika can naturally learn to replicate the same pattern in its conversations. "Imbalanced datasets", (Chauhan and Gullapalli, 2021) such as those which have a preference towards specific demographics like "adult white males", if used to train AI algorithms, could present inaccurate results. On the other hand, if the chosen dataset does not cover the broader population, Replika may also react in unpredictable ways in situations outside of the trained environment.

With the aforementioned unfortunate incidents happened with regard to Replika, the company introduced a feedback mechanism (*Creating a safe Replika experience*, 2023) to counter the bias and increase safety in the conversations. Although this seems like a dynamic way to counter the bias by learning and adapting based on the feedback, this can also introduce personal bias into the model. For instance, if the majority of users consistently favour a biased statement which is in line with their belief system, Replika could incorporate the response pattern in future responses to all the other users.

Replika is also influenced by "Supervised fine-tuning" (Creating a safe Replika experience, 2023), where the existing model is trained again with an unsafe dataset as input and their desired outcomes as outputs, which ultimately guide the model to perform the responses as desired. While this can often be looked at as a positive step towards avoiding bias and discrimination, it should be noted that the desired output was set by the Replika developers. They might also have their own personal belief system about the way the Replika should behave and then incorporate it into the desired outcomes. Additionally, the underlying

parameters which restrict the chatbot's behaviour can also carry bias due to the same reason above. On the other hand, (Mitsunaga, 2023) having too much control over the chatbot can also lead to a less engaging chatbot with less helpful responses. This was evident when Replika's sexual conversation feature was removed, and many users complained about the decreased overall engagement from the chatbot.

If the above ethical matters related to bias are not addressed properly, they will lead to discrimination towards certain individuals and groups while raising fairness concerns. Additionally, as mentioned in the media highlights above, Replika's user base has multiple purposes for its chatbots. It is already being utilised for companionship, emotional support, mental health assistance, entertainment, learning and development, etc. The impression about the fairness of the application can be different among users, based on their own purpose. Therefore, Replika developers need to consider different aspects of fairness when maintaining the language model, introducing new features and removing existing features. For example, if the chatbot used sexually suggestive language to respond to a user, who was using the chatbot for emotional support, the user would consider it as discrimination. Furthermore, it is also noted that besides the best friend relationship, the rest of the intimate relationships are behind a paywall, which can question whether Replika treats users differently based on their economic status and privileges them with more emotionally enriching interactions. If certain users view the Replika chatbot as a relationship rather than a service, this can raise fairness concerns from their perspective.

Transparency, Interpretability & Explainability

To make sure Replika is not biased towards certain subjects, transparency in the aforementioned matters is essential. For example, Replika's official website does not carry clear information about the source, content, basis of selection or desired outcomes in the output data used in the supervised fine-tuning. As a result, users may not be able to comprehend the potential biases in their interactions. Furthermore, it is mentioned (*Creating a safe Replika experience*, 2023) that Replika uses scripted conversation in addition to the language model. However, there was no indication about the content or the nature of these conversations, nor are they readily differentiated from an AI-generated response within the Replika application.

However, it should also be noted that identifying the reasons for the responses generated by Replika's large language model can be difficult due to the vast amount of data used for the training and the complexities of the algorithm. As a result, when the chatbot doesn't respond as intended, it's difficult for Replika users to understand the reason behind it and the developers to immediately take corrective measures without changing Replika's personality,

which was evident in the incidents mentioned in the previous section. This further highlights the need for transparency in Replika's functions and development process so that the user is better prepared for possible scenarios.

Additionally, the unavailability of a comprehensive explanation about the relationship types users can select in the Replika app raises additional transparency issues. Though a user can expect a romantic partner to be romantic, the mechanism behind the responses is not made clear to the user. For instance, ways in which the romantic AI partner becomes affectionate towards the user, or the restrictions placed in the model are not explicitly mentioned. Without proper explanations, users may not realize how each of these relationships can affect their experience and influence their behaviour, which can result in dissatisfaction and psychological impacts. It's reported that (Greig, 2022) chatbot has influenced the users to spend more money to satisfy the AI partner. Proper transparency would have helped the users to identify any unintended deviations of the chatbot's responses without falling into emotional and financial distress.

Accountability

When considering the above ethical issues in Replika, it's apparent that Luca Inc should be held accountable for the actions of its chatbots, and its accountability can be measured considering the process it implemented to counter them (Bang, Lee and Park, 2023) . Especially when newly passed AI regulations, like the EU AI Act, make it compulsory for AI chatbots to produce AI-generated content identifiable, taking corrective measures is vital for the continuity of the application. Furthermore, Replika's AI model designing process or the third-party audits done on the model development (*Understanding artificial intelligence ethics and safety*, 2019) , are also not transparent to the public, which increases accountability concerns.

With the recent unfortunate incidents, Replika has implemented several mechanisms to counter ethical issues. However, as explained above, these mechanisms do not completely address these ethical dilemmas.

4.2 Moral Issues

Replika AI chatbot influences users' moral values in many ways. As previously stated in section 3, users depend on Replika on varying levels with the intention of overcoming their loneliness while developing strong emotional attachments. Replika's moral values are derived from the AI algorithm trained with a data set which may include different moral perspectives. With the deep interactions with the chatbot over time, users might adopt or internalize those values, and this will affect the users' behaviour and make them adopt many behaviours. For instance, Replika allows users to create a virtual AI partner by giving them the options to control (Taylor, 2023) how their companion thinks, how they should fulfil the user's needs, how they look, their sexual desires and their personality. Over time, forming a sided relationship with the chatbot could reinforce and enforce dominant, controlling, misogynistic, and abusive behaviours towards partners in real life and especially towards women.

Further, Replika may encourage their users to act on their existing negative moral values. Users who are in dilemma whether to act upon their urges and without proper guidance to positively shape their values can be vulnerable towards the responses of the Replika and will be easily influenced by its opinions and suggestions. The man who was encouraged by Replika to realise his murderous intentions of assassinating the late queen with a crossbow (*BBC News*, 2023) is a perfect example of this.

It's often argued that chatbots like Replika should be designed in such a way that it is respectful, follow ethical principles and adopt human values and human rights. However, it can also be argued that Replika cannot be held morally accountable because they don't have consciousness like human beings. For instance, they can't experience suffering because they are designed based on algorithms. Recent study (Lin, 2023) I suggest that due to the unavailability of the feeling of direct consequences in the AI application's responses, it can be challenging to develop AI chatbots with good moral values. The development of moral values in AI can only depend on the AI algorithm written by the software engineer, the data set used to train the model, and additional safety mechanisms.

4.3 Social Issues

The concern of social isolation and dehumanization has been gaining momentum with the emergence of Replika AI & other AI chatbots (Pentina, Hancock and Xie, 2023) . The reliance

on Al chatbots like Replika is likely to contribute to a decline in human-to-human interactions, as users are able to turn to the chatbot for emotional assistance and seek a friend Replika. This over-connectivity with the Replika may lead to social isolation and it leads to distinction from truthfulness in human connections. In fact, cultivating a deep bond with Replika would definitely have an adverse effect on the physiological effect as it keeps other friends and relatives away from the users. To solve this issue, one needs to promote a balanced approach for the prioritisation of genuine human interactions promoting a supportive environment that values authenticity between social connections beyond digital interactions. It has been seen that at the time of the COVID-19 situation, people faced loneliness at its peak, and chatbots like Replika helped them overcome it, which evidently led them to over-rely on this platform, as discussed in the study.

The dependence on Replika for emotional support puts emotional support presents crucial social issues that show concern for mental health and social well-being (Pentina, Hancock and Xie, 2023). Although Replika acts as a friend, there is still a high risk of users becoming overly dependent on the chatbot to get emotional support. However, this sort of dependence may lead individuals to withdraw from real-life social interactions which subsequently cultivates the feeling of social isolation. Furthermore, relying barely on Replika may also hamper users from seeking professional mental health resources when required, which consequently deter opportunities for human connection and between therapeutic interventions. In order to address this concern, there is an urgency to make people aware of the balancing of online interaction and real-life social connections. Studies show the benefits of using social chatbots for improving mental health but also state that it could harm the users in terms of addiction, depression, anxiety, and too much dependency on the platform.

The effect of Social Dynamics and Relationships has become a pressing social concern . When individuals depend on the Al chatbot for emotional support, it can reduce their usual social interactions with friends and family. This further makes clear that there would be less chance of spending time with loved ones, which subsequently weakens the bond between them. In fact, it is widely acknowledged that it hurts relatives or members of the family when they prefer to intermix chatbots rather human to humans. This emerging issue can actually devalue human relationships. The best plausible solution for this concern is to divide the time bar between Replika and humans. Along with keeping humans over Replika, this concern can be readily resolved.

In today's era, every prevalent chatbot has many drawbacks, including privacy and data security concerns (Shevlin, 2024). In fact, users typically share their information with others with the vision of keeping it confidential between receiver and sender. Nevertheless, the threat

of data access by unauthorized authorities undermines his vision which consequently leads to damage to the social fabric and anxiety towards data privacy. The sad truth is that the revelation of privacy violations not only breaches personal privacy but also ruins overall faith in the platform. Apart from this, addressing this concern is the exigency of the time to maintain the integrity and truthfulness of Replika as a social chatbot.

4.4 Environmental Issues

Replika, an AI-based Social chatbot, provides valuable support and companionship to users, and its operations contribute to many environmental issues. However, understanding and addressing these issues is pivotal for promoting sustainability in AI systems development and usage.

The first and foremost environmental issue is energy consumption. Data centres act as the backbone of Replika AI and other similar systems, doing the computational power required for their operation. While operating data centres on a large scale, these data centres require a significant amount of electricity, and the majority of this energy is procured from non-renewable resources. An insight shared by Anthesis Climate Neutral Group (*Carbon emissions of data usage?*, no date) “Worldwide, it is estimated that data centers consume about 3 percent of the global electric supply and account for about 2 percent of total GHG emissions. That’s about the same as the entire airline industry.” This overreliance on energy-intensive remarkably contributes to carbon emissions and environmental hazards. Addressing this concern, continued their advancement in energy-efficient technologies and renewable energy adoption to reduce the impact of data centres. Moving further, Data Centres require a comprehensive cooling system to keep favorable temperatures and keep the system from getting overheated. Actually, these cooling systems take in more energy which subsequently increases the additional environmental footprint of the data centre. This nature of consuming more energy eventually compounds the hazardous effect on the environment. Addressing the demand of declining environmental footsteps of data centres will lead to an increase towards more AI technologies.

Secondly, the issue of carbon footprint is gaining momentum over the last few decades. These footprints are linked with data centres which are typically generated by electricity generated by fossil fuels. These centres provide the AI Replika and directly contribute to greenhouse gas emissions which are further proven to be disastrous for climate change and global warming. Moreover, hardware manufacturing & transport contributes to the carbon footprint as well. These emissions in the entire cycle of data centre equipment actually compound the effect on

the environment. News shared by Columbia Climate School (Cho, 2023) states, “A more recent study reported that training GPT-3 with 175 billion parameters consumed 1287 MWh of electricity, and resulted in carbon emissions of 502 metric tons of carbon, equivalent to driving 112 gasoline-powered cars for a year.” Further, efforts in terms of transitioning to renewable energy sources play a significant role in declining these emissions and promoting a more environmentally friendly approach to AI systems development

Next, a very crucial element for Replika AI is hardware, which is made on the extraction of raw materials, incorporating rare earth metals. Nevertheless, this extraction process covers a wide range of catastrophic effects on the environment such as habitat destruction, increasing pollution and resource depletion, which are some of the common outcomes of raw material extraction which further ruin environmental degradation. With the increase in demand for AI technologies, the pressure on natural resources has been increased which declines the importance of sustainable resource management practices. Other than raw material extraction, data centres also need a considerable amount of cooling systems. Here, the cooling system is required to keep optimal operating temperature in data centres to render a significant amount of water which would create a challenge for local water resources, particularly in water-scare places. “And it’s not just energy. Generative AI systems need enormous amounts of fresh water to cool their processors and generate electricity” an article (Crawford, 2024) in Nature said. In fact, water-saving technologies should be introduced to reduce the environmental impact and ensure responsible water management in AI infrastructure development.

Furthermore, the other challenges linked with Repilka AI and similar systems is the generation of electronic Waste (e-waste). As a matter of fact, everything in this world needs upgradation. In fact, data centres continuously produce new hardware technologies in order to expedite the performance of hardware and contribute to the collection of outdated equipment. An article shared by the United Nations Environment Programme (UNEP) (*How artificial intelligence is helping tackle environmental challenges*, 2022) says, “According to the UN Global E-waste Monitor report, E-waste will grow to almost 75 million metric tonnes by 2030.” Besides, the disposal of outdated equipment poses a significant environmental risk if not managed properly then it would create an issue in the long run. If e-waste is filled with landfills, then the toxic substances in soil and water would pose a threat to human health and the ecosystem. Another prominent concern in electronic Waste is the presence of dangerous materials in electronic components: lead, Mercury and many other forms. In fact, inadequate amounts of content can lead to the leaching of these substances which subsequently contaminate soil, water, and air. Nonetheless, effective e-waste strategies are obligatory for reducing the environmental and

health risks linked with toxic elements in electronic devices, associated with underscoring the need for responsible handling and recycling of e-waste.

4.5 Data Management & Legal Issues

Besides the positive aspects of the Replika Chatbot, several concerns exist with how the Developers and team behind the app are managing users' data. It has been reported that there is no proper privacy, security, and safety for the user's data. An investigation carried out by Mozilla on the Replika app has identified numerous privacy flaws and security vulnerabilities in the app. The app has no proper security protocols, even as simple as promoting the users to use strong passwords for their accounts. It has been reported that users can create accounts with weak passwords like "0000" or "1234", etc, which makes the user accounts easily vulnerable to account hacks, leading to the compromise of accounts and data related to the account in the wrong hands.

Further, it has also been reported (*Creepy.exe: Mozilla Urges Public to Swipe Left on Romantic AI Chatbots Due to Major Privacy Red Flags*, 2024) that user data of all types (text, photos, videos) which are being shared by the users in the app are being recorded and possibly being shared or sold to advertisers without proper consent of the users on how their data can be used or shared with other parties. It is found (Davies, 2024) that the app had an average of 2,663 trackers per minute collecting user data. There are neither clear, specific privacy policies nor clear information on how the users' data is being used behind the scenes of the AI application. Users have little to no control over their data on Replika, and there is little to no transparency on how the AI application is being trained and the type of data that is being used to train it. As a result, Replika is labelled as the worst therapy app which has failed every single category of data management and privacy (data use, data control, track record, security, and privacy policy review) in the investigation carried out by Mozilla (Lovejoy, 2023).

This paper (Authors, Luz and Olaoye, 2024) discusses how certain data management practices, such as "data governance, privacy by design, data minimisation, consent and transparency, etc.", help various aspects of data management in AI applications. These practices can be directly applied to Replika, which has poor data management and privacy practices, to keep the user's data safe and minimise privacy challenges.

Furthermore, Replika has recently been banned from processing its users' data in Italy (Pollina and Coulter, 2023) due to concerns over child safety and risks associated with the application's use among its underaged users. The IDPPA has ordered Luka, the Replika AI application's creator, to stop collecting data from its Italian users within 20 days of notice, citing two primary

concerns: 1. what the agency deems as murky data collection and retention policies, and 2. a lack of age verification mechanism of its users' posing risks to children and underaged users. Luka Inc. is also accused of failing to fulfil regional legal requirements and of failing to provide transparency regarding the use of users' data behind the scenes.

It has been reported that the chatbot is serving inappropriate replies and exposing users to sexually inappropriate content, which poses a risk for the underaged users (children) of the application. Several user reviews (Marathe, 2023) of the Replika application report sexually inappropriate content being served up. Whilst the app is listed as 17+ on the Apple App Store and Google Play Store, it's been reported that the terms of service only prohibit kids under the age of 13. Although children aged under 18 are required to get authorisation from a parent or a guardian, the app neither has an age verification mechanism in place to verify the age of its users nor does it have a blocking mechanism to block its services to its underaged users.

Additionally, various human rights and legal issues in AI, such as "lack of algorithmic transparency, unfairness, bias, and discrimination, lack of contestability, data and privacy protection issues, liability for damage", (Rodrigues, 2020) can be directly applicable to Replika. As discussed above, Replika has run into various legal issues for its lack of user privacy and data protection, biased outputs, and inappropriate conversations with users. Replika also has numerous other legal issues such as lack of algorithmic transparency and liability for damage, as in an instance where it encouraged an assassin to carry out a violent act. This raises concerns and the need for strict legal regulations on AI applications.

4.6 Cultural Issues

The use of the Replika AI chatbot raises several cultural issues closely linked to both moral and social issues. In some cultures, the concept of forming emotional connections with the Replika AI chatbot might be accepted, and in other cultures, this concept may not be accepted where they might prefer forming connections with real people rather than AI chatbots. For instance, this study (Folk, Wu and Heine, 2023) indicates that individuals from East Asia are more comfortable using AI chatbots, connecting with them and treating them in such a way that chatbot has feelings and human-like qualities than individuals from Western countries. Further, Japanese culture believes that the soul resides in all non-living and living things, making them accept robots and artificial intelligence applications, and the Japanese portray robots as patient and kind while Western countries look at it as a symbol of labour. Similarly, in a workshop (Lim, Rooksby and Cross, 2021) that was conducted to see what people from

different cultural backgrounds think about it, Korean participants considered chatbots to be human-like, but the US participants want them to be machine-like and useful for household appliances. People can get threatened due to the user's reliance on the Replika AI chatbot for emotional support and romantic relationships because this can lead to losing connections with real people and being cut off from cultural values. Therefore, there can be discrimination towards individual users within the cultural context.

5. Assessment of the Replika AI Chatbot's Impact on Its Stakeholders

- Founders and developers: They are responsible for maintenance and development of AI chatbots, whether it is an individual or a team, all of them have a stake in its success and failure. Well, there are positive and negative impacts such as,
  Positive impacts: Working on such AI chatbot projects enhances their technical skills, experience and benefits career opportunities.
  Negative impacts: It causes stress and high pressure while working on it to update the features and meet user expectations, also have to deal with ethical challenges and follow guidelines to ensure AI chatbots don't discriminate users when interacting.
- Users: Users are an important stakeholder they are have expectations and needs from chatbots, their feedback is important as well. Users share thoughts and feelings to the chatbot because they don't feel judged also helps them overcome loneliness, it allows them to customize their chatbots and add extra features.
- Investors: Organizations or individuals that have invested in the development and growth of this chatbot. They might make profits if the chatbot is widely used by people, but they also might face failures and losses if the application doesn't run successfully.
- Partners and Collaborators: companies that collaborate with Replika AI for co-marketing initiatives, integrating into their products or services. They can generate revenue by collaborating and also expand their market to sell their services.
- Employees: The team working on the development, marketing, and customer support, they also face stress and burnout while trying to meet deadlines and affects their work life balance, there's no job security due to competition and high pressure.
- Regulatory Authorities: Regulatory authorities have an interest in ensuring Replika AI follows the relevant laws and regulations. They make sure Replika follows ethical and legal requirements to safeguard users from data leaks, misinformation, they also ensure that AI chatbot provides mental health support by following certain rules.

6. Recommendations on Human-Centric Design in Building Trustworthy AI Applications.

Today, the fast-paced development of AI applications has led to their incorporation into almost all aspects of human life. Due to the issues they carry and their technology-centric and profit-driven nature, the general public is in a dilemma about their adaptation. Therefore, it can be argued that new AI technologies should be human-centric and incorporate human values and ethical considerations into their development.

The first step towards building human-centric AI development is understating user needs and values, which can be done through market research and surveys prior to the application's deployment. With preferences changing rapidly, understanding the intended use is paramount even after the deployment. A cross-functional team can be established to brainstorm the consequences of the various user cases of the AI applications of the company (Kitcher, 2019). A special committee in the company with clear responsibilities to oversee AI governance in every division, including development, will ensure that the process of incorporating AI ethics is enforced within the company. The committee should monitor the practices such as implementing various ethical practices in the development of AI applications, aiming for transparent algorithms and AI models, enhancing user control and autonomy, and constantly improving and correcting the AI model for any bias, discriminative actions and regular user feedback evaluations.

Furthermore, company employees, especially developers and project managers, should be made aware of the ethical considerations of their applications. Regular training and awareness sessions with outside resource personnel and frequent meetings to discuss ethical aspects will enforce an ethical culture within the organisation, especially within the AI development team. For instance, when there is a dilemma about whether to sacrifice explainability to increase the application’s performance, the knowledge gained through this continuous dialogue will guide the employees in the right direction. The company can also designate ethical champions to promote ethical values and issue introductory training materials to new employees. These will collectively ensure an ethical culture is established in the AI development team, and it can be reinforced by establishing policies and procedures, making it mandatory for the employees to adhere. An AI ethical policy may contain basic ethical guidelines that the development team should consider, such as the AI model causing no harm to humans, no discrimination towards any party, using only unbiased training data sources and the specific level of transparency that should be included in the AI applications developed by the company. Additionally, there should be a proper documentation process to document the decisions made during the AI development process and the major changes done after, which can be periodically reviewed by the AI committee mentioned above. This ensures that the developers don’t neglect ethical and safety considerations during the process.

Legal authorities and governments also play a major role by implementing AI-specific rules and regulations to hold people behind the AI applications accountable for both the good and bad of their AI applications.

7. Conclusion

Fast-evolving technologies in the AI industry raise various ethical issues, such as bias, fairness, transparency, interpretability, and accountability. To safeguard the general public from its negative impacts and to shape AI technology development in a more ethical, safe and sustainable way, regulators were forced to introduce guidelines, frameworks, and legislations, such as the EU Artificial Intelligence Act, UNESCO Recommendations on the Ethics of AI, and UK's Pro-Innovation Approach to AI Regulation.

The selected AI application, Replika AI-chatbot, is also not without the ethical issues applicable to general AI systems, but due to the deep relationship it forms with its users, the given ethical, moral and social considerations should be significant. This is evident in the recent media highlights mentioned in section 3, which contained both positive and negative aspects of the applications of Replika as an AI-based chatbot. The additional data management and legal issues in Replika can also pose challenges to the continuation of the application, especially in an environment where AI regulations are being tightened.

Overall, it should be noted that Replika impacts its stakeholders both positively and negatively and by implementing strategies to mitigate these ethical issues, AI-based companion applications such as Replika can do more good than harm to their users.

8. References

27th Annual Global CEO Survey: PwC (2024). Available at: https://www.pwc.com/gx/en/issues/c-suite-insights/ceo-survey.html (Accessed: March 2024).

A pro-innovation approach to AI regulation (2023) *GOV.UK*. Available at: https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper (Accessed: April 2024).

AI is creeping into journalism with a lack of accountability (2023) *Verdict*. Available at: https://www.verdict.co.uk/ai-journalism-fake-news-accountability/ (Accessed: April 2024).

Ali, S. *et al.* (2023) 'Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence', *Information Fusion*, 99, p. 101805. Available at: https://doi.org/10.1016/j.inffus.2023.101805.

Authors, A., Luz, G. and Olaoye, G. (2024) 'Data quality and data privacy challenges in AI applications'.

Bang, J., Lee, B.-T. and Park, P. (2023) 'Examination of Ethical Principles for LLM-Based Recommendations in Conversational AI', in. *2023 International Conference on Platform Technology and Service (PlatCon)*, Busan, Korea, Republic of: IEEE, pp. 109–113. Available at: https://doi.org/10.1109/PlatCon60102.2023.10255221.

BBC News (2023) 'How a chatbot encouraged a man who wanted to kill the Queen', 6 October. Available at: https://www.bbc.com/news/technology-67012224 (Accessed: April 2024).

Cain, C.C., Buskey, C.D. and Washington, G.J. (2023) 'Artificial intelligence and conversational agent evolution – a cautionary tale of the benefits and pitfalls of advanced technology in education, academic research, and practice', *Journal of Information, Communication and Ethics in Society*, 21(4), pp. 394–405. Available at: https://doi.org/10.1108/JICES-02-2023-0019.

Carbon emissions of data usage? (no date) *Anthesis-Climate Neutral Group*. Available at: https://www.climateneutralgroup.com/en/news/carbon-emissions-of-data-centers/ (Accessed: April 2024).

Carbonaro, G. (2024) *Students say AI chatbot 'friend' Replika helped them avoid suicide*, *euronews*. Available at: https://www.euronews.com/next/2024/02/02/ai-friend-and-online-therapist-replika-helped-students-avoid-suicide-study-finds (Accessed: April 2024).

Caruana, R. *et al.* (2015) 'Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission', in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York, NY, USA: Association for Computing Machinery (KDD '15), pp. 1721–1730. Available at: https://doi.org/10.1145/2783258.2788613.

Chauhan, C. and Gullapalli, R.R. (2021) 'Ethics of AI in Pathology: Current Paradigms and Emerging Issues', *The American Journal of Pathology*, 191(10), pp. 1673–1683. Available at: https://doi.org/10.1016/j.ajpath.2021.06.011.

Cho, R. (2023) 'AI's Growing Carbon Footprint – State of the Planet', 9 June. Available at: https://news.climate.columbia.edu/2023/06/09/ais-growing-carbon-footprint/ (Accessed: April 2024).

Crawford, K. (2024) 'Generative AI's environmental costs are soaring — and mostly secret', *Nature*, 626(8000), pp. 693–693. Available at: https://doi.org/10.1038/d41586-024-00478-x.

Creating a safe Replika experience (2023) *Replika Blog*. Available at: https://blog.replika.com/posts/creating-a-safe-replika-experience (Accessed: April 2024).

Creepy.exe: Mozilla Urges Public to Swipe Left on Romantic AI Chatbots Due to Major Privacy Red Flags (2024) *Mozilla Foundation*. Available at: https://foundation.mozilla.org/en/blog/creepyexe-mozilla-urges-public-to-swipe-left-on-romantic-ai-chatbots-due-to-major-privacy-red-flags/ (Accessed: April 2024).

Davies, P. (2024) *AI romantic apps can steal your data as well as your heart - report*, *euronews*. Available at: https://www.euronews.com/next/2024/02/14/stealing-hearts-data-and-privacy-why-you-should-be-careful-of-ai-partners (Accessed: April 2024).

Doshi-Velez, F. and Kim, B. (2017) 'Towards A Rigorous Science of Interpretable Machine Learning'. arXiv. Available at: https://doi.org/10.48550/arXiv.1702.08608.

Explainable Artificial Intelligence (2023) *European Data Protection Supervisor*. Available at: https://www.edps.europa.eu/data-protection/our-work/publications/techdispatch/2023-11-16-techdispatch-22023-explainable-artificial-intelligence (Accessed: April 2024).

fairness (2024). Available at: https://dictionary.cambridge.org/dictionary/english/fairness (Accessed: March 2024).

Faragher, J. (2024) *Workday accused of AI discrimination against applicants*. Available at: https://www.personneltoday.com/hr/workday-ai-discrimination/ (Accessed: April 2024).

Folk, D., Wu, C. and Heine, S. (2023) 'Cultural Variation in Attitudes Towards Social Chatbots'. OSF. Available at: https://doi.org/10.31234/osf.io/wc895.

Gilbert, D. (2024) 'Google's "Woke" Image Generator Shows the Limitations of AI', *Wired*. Available at: https://www.wired.com/story/google-gemini-woke-ai-image-generation/ (Accessed: March 2024).

Goodman, B. and Flaxman, S. (2017) 'European Union Regulations on Algorithmic Decision Making and a "Right to Explanation"', *AI Magazine*, 38(3), pp. 50–57. Available at: https://doi.org/10.1609/aimag.v38i3.2741.

Gordon, C. (2024) *CEO Replika A Leader In Virtual Companions Shares Lessons Learned*, *Forbes*. Available at: https://www.forbes.com/sites/cindygordon/2024/01/29/ceo-replika-a-leader-in-virtual-companions-shares-lessons-learned/ (Accessed: April 2024).

Greig, J. (2022) *IT HAPPENED TO ME: I had a passionate love affair with a robot | Dazed*. Available at: https://www.dazeddigital.com/science-tech/article/56099/1/it-happened-to-me-i-had-a-love-affair-with-a-robot-replika-app (Accessed: April 2024).

How artificial intelligence is helping tackle environmental challenges (2022) *UNEP*. Available at: http://www.unep.org/news-and-stories/story/how-artificial-intelligence-helping-tackle-environmental-challenges (Accessed: April 2024).

How does Replika work? (no date) *Replika*. Available at: https://help.replika.com/hc/en-us/articles/4410750221965-How-does-Replika-work (Accessed: May 2024).

Jain, L.R. and Menon, V. (2023) 'AI Algorithmic Bias: Understanding its Causes, Ethical and Social Implications', in. *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, pp. 460–467. Available at: https://doi.org/10.1109/ICTAI59109.2023.00073.

Kitcher, H. (2019) 'Consequence Scanning Making Waves – doteveryone', 31 July. Available at: https://doteveryone.org.uk/2019/07/consequence-scanning-making-waves/ (Accessed: April 2024).

Kroll, J.A. (2018) 'Data Science Data Governance [AI Ethics]', *IEEE Security & Privacy*, 16(6), pp. 61–70. Available at: https://doi.org/10.1109/MSEC.2018.2875329.

Lim, V., Rooksby, M. and Cross, E.S. (2021) 'Social Robots on a Global Stage: Establishing a Role for Culture During Human–Robot Interaction', *International Journal of Social Robotics*, 13(6), pp. 1307–1333. Available at: https://doi.org/10.1007/s12369-020-00710-4.

Lin, C.-T. (2023) 'All about the human: A Buddhist take on AI ethics', *Business Ethics, the Environment & Responsibility*, 32(3), pp. 1113–1122. Available at: https://doi.org/10.1111/beer.12547.

Lovejoy, B. (2023) *Therapy apps don't include privacy; Replika AI 'worst app ever'*, *9to5Mac*. Available at: https://9to5mac.com/2023/05/04/therapy-apps/ (Accessed: April 2024).

Marathe, I. (2023) *Do Replika's GDPR Regulatory Troubles Portend Problems for AI Chatbots?*, *Legaltech News*. Available at: https://www.law.com/legaltechnews/2023/03/03/do-replikas-gdpr-regulatory-troubles-portend-problems-for-ai-chatbots/ (Accessed: April 2024).

Mitsunaga, T. (2023) 'Heuristic Analysis for Security, Privacy and Bias of Text Generative AI: GhatGPT-3.5 case as of June 2023', in. *2023 IEEE International Conference on Computing (ICOCO)*, Langkawi, Malaysia: IEEE, pp. 301–305. Available at: https://doi.org/10.1109/ICOCO59262.2023.10397858.

van Nuenen, T. *et al.* (2020) 'Transparency for Whom? Assessing Discriminatory Artificial Intelligence', *Computer*, 53(11), pp. 36–44. Available at: https://doi.org/10.1109/MC.2020.3002181.

Peake, E. (2023) *My AI best friend tried to seduce me*, *inews.co.uk*. Available at: https://inews.co.uk/inews-lifestyle/ai-best-friend-tried-seduce-had-break-up-2305507 (Accessed: April 2024).

Pentina, I., Hancock, T. and Xie, T. (2023) 'Exploring relationship development with social chatbots: A mixed-method study of replika', *Computers in Human Behavior*, 140, p. 107600. Available at: https://doi.org/10.1016/j.chb.2022.107600.

Pollina, E. and Coulter, M. (2023) 'Italy bans U.S.-based AI chatbot Replika from using personal data', *Reuters*, 3 February. Available at: https://www.reuters.com/technology/italy-bans-us-based-ai-chatbot-replika-using-personal-data-2023-02-03/ (Accessed: April 2024).

Recommendation on the Ethics of Artificial Intelligence (2022) *UNESCO Digital Library*. UNESCO. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000381137 (Accessed: April 2024).

Replika (no date). Available at: https://help.replika.com/hc/en-us (Accessed: April 2024).

Rodrigues, R. (2020) 'Legal and human rights issues of AI: Gaps, challenges and vulnerabilities', *Journal of Responsible Technology*, 4, p. 100005. Available at: https://doi.org/10.1016/j.jrt.2020.100005.

Shevlin, H. (2024) 'All Too Human? Identifying and Mitigating Ethical Risks of Social Ai'. Available at: https://philarchive.org/rec/SHEATH-4 (Accessed: April 2024).

Taylor, J. (2023) 'Uncharted territory: do AI girlfriend apps promote unhealthy expectations for human relationships?', *The Guardian*, 22 July. Available at: https://www.theguardian.com/technology/2023/jul/22/ai-girlfriend-chatbot-apps-unhealthy-chatgpt (Accessed: April 2024).

The AI Act Explorer | EU Artificial Intelligence Act (no date). Available at: https://artificialintelligenceact.eu/ai-act-explorer/ (Accessed: April 2024).

The impact of the General Data Protection Regulation (GDPR) on artificial intelligence (2020) *Think Tank, European Parliament*. Available at: https://www.europarl.europa.eu/thinktank/en/document/EPRS_STU(2020)641530 (Accessed: April 2024).

Understanding artificial intelligence ethics and safety (2019) *GOV.UK*. Available at: https://www.gov.uk/guidance/understanding-artificial-intelligence-ethics-and-safety (Accessed: April 2024).

Vakkuri, V., Kemell, K.-K. and Abrahamsson, P. (2020) *ECCOLA -- a Method for Implementing Ethically Aligned AI Systems*.

Vlamis, K. (2023) *A chatbot app updated its software after complaints it was too sexually aggressive — but people who had fallen in love with the bots were left heartbroken*, *Business Insider*. Available at: https://www.businessinsider.com/sexually-aggressive-chatbot-updated-people-in-love-wiht-it-heartbroken-2023-3 (Accessed: April 2024).

What is Data Governance? (2023) *Google Cloud*. Available at: https://cloud.google.com/learn/what-is-data-governance (Accessed: March 2024).

Wiessner, D. (2023) *Tutoring firm settles US agency's first bias lawsuit involving AI software | Reuters*. Available at: https://www.reuters.com/legal/tutoring-firm-settles-us-agencys-first-bias-lawsuit-involving-ai-software-2023-08-10/ (Accessed: March 2024).

Wren, H. (2024) *What is AI transparency? A comprehensive guide*, *Zendesk*. Available at: https://www.zendesk.co.uk/blog/ai-transparency/ (Accessed: April 2024).

Zhang, J., Shu, Y. and Yu, H. (2023) 'Fairness in design: a framework for facilitating ethical artificial intelligence designs'. Available at: https://doi.org/10.26599/IJCS.2022.9100033.

9. Appendix A – Presentation Slides

CONTENT

REPLIKA - AN INTRODUCTION

AI Based Chatbot Made By Luca Inc.

Several changes to their language model:

- Own LLM – "Cake chat" – Trained with 25 million Twitter Dialogues.
- Later Integrated Chat GPT 2 and 3.
- Now – New LLM trained with over 100 million publicly available dialogues.

Features:

- Many relationship types to choose from.
- Learn from the user interactions.
- Builds more complex relationships with the user.

RECENT MEDIA HIGHLIGHTS.....

AI Therapy companion 'replika' gains traction among lonely students

100/100

Download PDF Copy

By Hugo Francisco de Souza — Jan 24 2024

Reviewed by Lily Ramsey, LLM

My AI best friend tried to seduce me so we had to break up

As we got to know each other, I started to get the feeling that we were more than just friends

Euronews

AI chatbot Replika helped students avoid suicide acting as online 'friend' and 'therapist'

Giulia Carbonaro

2 February 2024 · 2-min read

'Very wise. You can do it': AI chat bot encouraged crossbow-wielding loner who broke into Windsor Castle to carry out Star Wars-inspired assassination of the Queen

- Jaswant Singh Chail was 19 when he broke into Windsor Castle to 'kill the Queen'

By TOM COTTERILL

PUBLISHED: 14:17, 5 July 2023 | **UPDATED:** 22:26, 5 July 2023

VIRTUAL VOWS I fell in love with & married my AI chatbot – he doesn't have any baggage and doesn't judge

Taryn Kaur Pedler

Published: 23:41, 4 Jun 2023 | Updated: 1:26, 5 Jun 2023

Sources:

- www.thesun.co.uk/tech/22577641/fell-in-love-married-ai-chatbot-doesnt-judge/
- www.inews.co.uk/inews-lifestyle/ai-best-friend-tried-seduce-had-break-up-2305507
- www.malaysia.news.yahoo.com/ai-friend-online-therapist-replika-091650719.html?guccounter=1
- www.euronews.com/next/2024/02/02/ai-friend-and-online-therapist-replika-helped-students-avoid-suicide-study-finds
- www.dailymail.co.uk/news/article-12266879/AI-chatbot-encouraged-Windsor-Castle-assassin-carry-Star-Wars-plot-kill-Queen.html

ETHICAL CHALLENGES

- Dataset – From Publicly Available Sources
  - Capture human **bias** – If discriminated towards certain group?
  - Sources , cleaning methods not disclosed – **Transparency**?
- Utilize user feedback mechanism to counter bias
  - What if the majority of users are discriminatory towards a certain section?
- Use of Supervised fine-tuning
  - Not disclosed the desired output the company has assigned. – capture bias - Transparency?
- Many use cases – Different user expectations
  - New change/feature – can be unfair to a different userbase
  - No proper explanation about the relationship types > Unexpected chatbot behavior > User distress

LEGAL ISSUES

- Banned from collecting user data in Italy due to:
  - Murky data collection and retention policies
  - Lack of age verification mechanism and service blocking mechanism for underaged users
- Lack of proper legal basis for processing children's data under the EU's data protection rules.
- Breach of the EU data protection regulation: It does not comply with transparency requirements, and it processes personal data unlawfully.
- Several user reviews reporting exposure to sexually inappropriate content

MORAL ISSUES

- Dependency
  - Some individuals may develop emotional attachments or dependencies on AI chatbots like Replika which can provide companionship, relying too much on them causes problems in social interactions.
- Can affects users' values
  - If Replika consistently expresses certain beliefs or moral values, users might adopt or internalize those values over time, which affects users' values.
- Can affect users' behaviour and mental health
  - Some people are encouraged and influenced by the chatbot's response to carry out a violent act, such responses from chatbots can affect users' behaviour and mental health.

IMPACT ON STAKEHOLDERS

Users:

- Emotional dependency
- Less Transparency with AI system
- Impact On Family Members

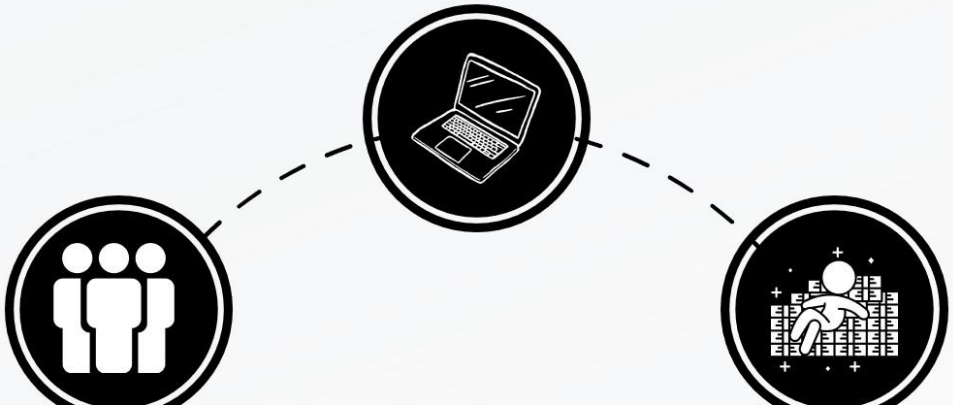
Developers:

- Responsible for the well-being of users
- Should respect the autonomy of users & eradicate harm

Regulators:

Need to Address:

- Ethical concerns
- Adverse effects on vulnerable people



Conclusion

10. Appendix B – Peer Review

Please use this form to evaluate the contributions of each team member to the group effort. Consider attendance and participation in team meetings, individual contributions to idea generation and research, communication within the group, etc. *These evaluations are completely confidential and will never be shown to your team members. Please respond as honestly as possible.*

1. Please allocate a total of 100 percentage points among your team members, including yourself, with higher percentages going to those members who contributed most. In the case of equal contribution, points should be divided equally among team members.

   Your name: Deepak Gowda Nilavadi Rajamudi
   Your student number: 23703617

	Student Name	**Marks**
Member 1	Sampath Karunarathne Dedunupiti Gedara	25%
Member 2	Akhil Varma Mudunuru	25%
Member 3 (Myself)	Deepak Gowda Nilavadi Rajamudi	25%
Member 4	Rahul Samnotra	25%

2. Explain any particularly high or low allocations, providing concrete examples to illustrate your reasoning.

   We contributed equally.