Linnæus University
Sweden

Degree Project

# ChatGPT and the developer's ethical responsibility

## - *A literature study of chatbot-related ethical dilemmas from the developer's perspective*

*Author:* Linda Meyer
*Supervisor:* Martin Östlund
*Subject:* Computer Science
*Course code:* 1DV50E
*Date:* 2023-06-04

# Abstract

In this thesis some ethical dilemmas involving conversational agents, with ChatGPT as the foremost example, are presented. Initially, the technology supporting chatbots is described to enable the reader to get insights into their underlying structure.

The reader will get an account of recent progress in the development of the technology and gain knowledge of ethical dilemmas from a developer's perspective. The main goal of this literature study is to achieve an understanding of the current situation and reflect on the developer's responsibility for building ethical chatbots. The content of this thesis is further based on previous research in the scientific field of chatbots.

This literature study supports the developer with multiple advice. For example, the importance of working with areas such as transparency, UI-design, reliability, accountability, and relativization is highlighted.

*Keywords:* ChatGPT, ethics, chatbots, developer responsibility, AI.

# Sammanfattning

I den här litteraturstudien presenteras etiska frågor gällande ”conversational agents” och den senaste tidens utveckling av ChatGPT kommer att stå i centrum för studien. Läsaren får först ta del av en allmän beskrivning av tekniken som ligger till grund för AI-baserade ”chatbots”.

Jag redogör för den senaste tidens tekniska utveckling på området samt presentera etiska frågeställningar från programmerarens perspektiv. Det huvudsakliga syftet med uppsatsen är att förmedla en förståelse för den nuvarande situationen och reflektera över utvecklarens ansvar när det kommer till att skapa etiska ”chatbots”. Tidigare forskning om ”conversational agents” står som en grund för reflektioner och diskussioner i den här uppsatsen.

Den här litteraturstudien avslutas med flertalet slutsatser som kan fungera som råd till utvecklare. Programmerare bör uppmärksamma frågor som rör transparens, redovisningsansvar och relativisering. Dessutom är det viktigt att ta hänsyn till aspekter som UI-design och reliabilitet.

*Nyckelord:* ChatGPT, etik, chatbots, utvecklarens ansvar, AI.

# Acknowledgements

I would like to thank the teachers of this course, namely Martin Östlund and Daniel Toll. I’m also very grateful to my feedback group of Gustav Karlberg and Therese Grass.

# Contents

**1 Introduction 1**

1.1 Background…...………………………..……………….……....1
  1.1.1 The History of Chatbots…...…….…………………….…..……..1
  1.1.2 OpenAI's ChatGPT…….………...……………..………...……...2
  1.1.3 Terminology ……………………………..………….........….2
    1.1.3.1 *Natural Language Processing*…………...………….……........2
    1.1.3.2 *Artificial Hallucination*…….…………….…………...….….2
1.2 Related work…...…………………….…………………......…............3
1.3 Knowledge contribution…...…………………………………….……....3
1.4 Limitations…...……………………………………………….……....4
1.5 Target group……………………………………………………......4

**2 Method 5**

2.1 Choice of method…...………………………………….…….......5
  2.1.1 *Literature study*…...……………………………….………......5
2.2 Reliability and Validity…………………………………….……...…6
2.3 Ethical Considerations ……………………………..……...…...…..6

**3 Analysis 7**

3.1 Introduction to Ethical Dilemmas……………………………...…..7
3.2 Ethical Frameworks…………………………..………………...…..7
3.3 Moral Authority ……………………….……………………………..7
3.4 The reliability concerns of chatbots …… …….……………………..9
3.5 Personal Integrity…………………………………………….……11
3.6 Biased Content………………………………………….……....13
3.7 Risk Zones....….……………………………………………………14
  3.7.1 *Medical treatments*…….......………………………….………...14
  3.7.2 *Copyrights*……………... …….……………………….…………...14
  3.7.3 *Climate change*……………... …….…………………..….………..15
  3.7.4 *Fake news*……………... …….……………………...…………..15
  3.7.5 *Employing the AI* ……………..…………………..…………..15
3.8 Developing the Developer…….……...…………….………..….......…16

**4 Discussion 17**

**5 Conclusions and Future Work 20**

**References 21**

**A Appendix 1 24**

# 1 Introduction

The subject of this, 7.5 HEC B-level, thesis in Computer Science is AI-powered chatbots. In relation to the example of ChatGPT we will study the ethical dilemmas involving chatbots. The purpose of this is to provide a summary of the most important areas of concern. The target group is programmers who are in the process of developing chatbots as well as utilizing the technology for other software projects. The main goal of this thesis is therefore to discuss the subject in a broader context and how the development of chatbots may affect people and society at large. The results should offer an overview of the current situation with references to previous research. Consequently, this thesis aims to provide developers with new thoughts on the ethical dimension of their work and discuss possible guidelines for the future.

## 1.1 Background

In just a short period of time, ChatGPT has become the major topic of conversation in workplaces and society in general. The fast development of AI-powered chatbots raises many questions. In fact, most AI programs result in certain ethical dilemmas. We may ask ourselves if the answers provided are reliable and trustworthy. Furthermore, it is of relevance to discuss integrity concerns and how these digital tools handle and store the user’s personal data. However, the ethical questions following the use of chatbots are numerous and involve many scientific fields.

### 1.1.1 The History of Chatbots

The term chatbot was introduced in 1994, but the history of conversational programs dates further back. In 1950, the English computer scientist Alan Turing published his famous article “Computing Machinery and Intelligence” where he proposed a test for assessing a machine’s intelligence. The test, nowadays known as the Turing test, is based on the ability of a human to distinguish between a human-like conversational program and a real human. In 1966, the Turing test gained renewed interest when the conversational agent ELIZA was published. By using pattern matching, ELIZA returned the user’s input as a question thus creating a conversation. In this way, ELIZA gave the impression of being able to understand its users. However, the program was simply built to detect keywords in the input and respond according to a template-based scheme [17].

Chatbots of today differ greatly from their predecessor ELIZA. They include more features and technologies such as real-time learning and NLP. In their literature review, Adamopoulou & Moussiades [17] claimed that the use of chatbots increased significantly after 2016. The year of 2022 marked another major development in the area of chatbots with the release of ChatGPT.

### 1.1.2 OpenAI’s ChatGPT

ChatGPT was made publicly available in November 2022 by the company OpenAI [4]. One of OpenAI’s founders was Elon Musk. Currently, the company’s CEO is Sam Altman. OpenAI is a research firm which specializes on Artificial Intelligence. The release of GPT-4 was accompanied with the goal of collecting user feedback to improve the system. Representatives of the company meant that ChatGPT initially could exhibit a biased behaviour.

In fact, GPT stands for Generative Pre-trained Transformer. The system was trained on huge amount of data and, according to Elon Musk, this also included Twitter data in specific [2]. Unlike some earlier chatbots, ChatGPT can answer follow-up questions. The features also include being able to challenge incorrect arguments, admit mistakes, and deny answering some questions [2].

### 1.1.3 Terminology

In this section we present, and explain, some terms regarding AI-powered chatbots.

#### 1.1.3.1 *Natural Language Processing*

ChatGPT was made possible thanks to the recent progress in the area of natural language processing. NLP is based on large language models. These models are trained on huge amount of text data from the web. Furthermore, they include the functionality to predict upcoming words [13].

#### 1.1.3.2 *Artificial Hallucination*

Currently, chatbots are accompanied by reliability concerns. The reason for this is often described as artificial hallucination. Artificial hallucination occurs when underlying algorithms generate a completely made-up answer which then is convincingly presented as a fact [8].

Artificial hallucination is a troublesome software flaw in conversational agents which has already proved to be expensive for development companies and their investors. In 2023, Google’s parent company Alphabet Inc. lost 100 billion

American dollars in market value. They had advertised their new chatbot Bard and unintentionally presented answers which contained inaccurate information [8].

In their thesis, Dziri et al. [10] studied the amount of artificial hallucination in three chatbots. They claimed that 60% of the answers were hallucinated. The chatbots thus expressed subjective opinions, unsupported facts, etc. Dziri et al. also concluded that the hallucinations were not solely due to insufficient training data, but deficiencies in the conversational models. These flaws may be the result of assigning answers with the highest probability. In addition, the pre-training process may add to a behavior, consequently inducing biases [10].

## 1.2 Related work

In this literature study some scientific articles like “*The moral authority of ChatGPT*” [1], “*Conversational AI: Social and Ethical Considerations* [7] and *“Applying Ethical AI Frameworks in practice: Evaluating conversational AI chatbot solutions”* [3] play a major role. They provide a scientific foundation for this thesis and the possibility to compare the authors’ results and conclusions.

Since ChatGPT only recently was made public, this new era of chatbots forms a comparatively unstudied scientific field. However, some scientific articles have still been published. To add a broader view to the subject, I have also referred to several newspaper articles. One of the reasons for doing so, is to include the latest development in this area.

## 1.3 Knowledge contribution

The focus of this thesis is to summarize the subject from the developer’s perspective. This will provide programmers with important knowledge of the social impact of their work. Consequently, it offers a reflection on the responsibilities of the development team and the individual programmer when building and utilizing AI-powered chatbots. Unlike a specific study of the technology, a literature study may place the subject in a broader context with an aim to support the professional group of developers by creating awareness of these questions.

Finally, there are two main research question which this thesis aims to answer:

- Which ethical dilemmas are programmers who develop AI-powered chatbots faced with?
- How can developers develop chatbots in a way which is ethically responsible?

## 1.4 Limitations

I limit this thesis to scientific studies and other material which provides insights to the ethics and functions of chatbots from a developer’s perspective. Since AI-powered chatbots are a comparatively new research area, I will include references to works performed in recent years. Furthermore, the development of chatbots affects many parts of today’s society. I will limit this literature study and focus on some of these areas.

## 1.5 Target group

The content of this thesis is mainly aimed at programmers who are in the process of developing AI-powered chatbots. In addition, the knowledge gained in this study is also beneficial when utilizing chatbots as a part of developing other software.

# 2 Method

The purpose of this chapter is to present and discuss this thesis' method. Furthermore, questions regarding reliability and validity are answered and some ethical considerations are brought into view.

## 2.1 Choice of method

GPT-4 mark a new era in computer science. However, the impact of this revolutionary technology is not limited to one scientific field only. In fact, chatbots like ChatGPT are on the brink of changing the precepts of numerous professions. Healthcare, education, finance, and arts; conversational agents can be applied to many different situations and purposes.

This study aims to highlight the developer's responsibility when developing conversational agents. Since previous research often focuses on ethical dilemmas in specific areas such as health care with diagnostic chatbots, I found a lack of scientific articles which described the ethical considerations from the developer's perspective.

Even though the GPT-4 technology is innovative, conversational agents have existed for quite some time now, and numerous scientific studies have dwelt upon the ethical questions which the technology raises. To summarize, compare and evaluate the conclusions of research in this area, a literature study seemed like the best alternative. This choice of method allows us to form an overview of some ethical considerations and developer responsibilities regarding the vast subject of chatbots like ChatGPT.

### 2.1.1 Literature study

In this thesis I gained my knowledge by studying various scientific articles. The articles were retrieved from search engines such as Linnaeus University's OneSearch and Google Scholar. I have listed examples of the search words in Appendix 1. The aim of this thesis is to acquire understanding of the subject from previous research and subsequently compare the results of the literature sources. By doing so, I create a foundation for a discussion on the developer's ethical responsibility when developing chatbots. This discussion is presented in Chapter 4. In this section, assumptions are made based on findings in the previous study of scientific articles. In Chapter 5 the conclusions of this literature study are summarized.

## 2.2 Reliability and Validity

This study has been conducted with the use of numerous scientific articles to ensure reliability. Since the conclusions of this thesis are dependent on the literature references, much effort was put into the search for reliable sources. I also compared the scientific articles, and their cited sources, as a part of the validation process. Furthermore, I chose to review scientific papers and newspaper articles published in recent years. The time scope ranges from 2019 to 2023. To validate the sources, I continuously read and compared results of studies which focused on similar research areas. Scientific articles which focus on chatbots, ethics as well as the development process are somewhat limited, and the choice of articles could therefore be biased. In addition, the rapid development in the area of chatbots, like ChatGPT, makes it difficult to predict future events and upcoming software challenges. To reduce the occurrence of unreliable assumptions about the future, this literature study focuses mainly on the current situation.

## 2.3 Ethical Considerations

The reason for choosing the method of a literature study was to summarize scientific views on the subject and provide the reader with an overview. As a result, the content of each reviewed article has been summarized to some extent. In addition, I have chosen excerpts of the articles and presented some of the authors' conclusions to the reader. Consequently, there is a risk of not doing justice to the scientific findings of the authors since important details may be disregarded in the process. When referring to scientific articles I have therefore aimed to describe their content as objectively as possible. In addition, the articles have been read carefully to reduce the risk of reproducing their content incorrectly in this thesis.

# 3 Analysis

## 3.1 Introduction to Ethical Dilemmas

In this chapter I investigate some ethical dilemmas which often affect the development process of chatbots. Numerous companies have established certain guidelines, or ethical frameworks, which serve as a benchmark. However, ethical frameworks for developing AI-powered chatbots often differ between companies. This chapter begins with some recommendations from Atkins et al. [3] as to what, they believe, should be included in an ethical framework for chatbot development.

## 3.2 Ethical Frameworks

In the article “Applying Ethical AI Frameworks in practice:
Evaluating conversational AI chatbot solutions“ the authors conclude that guidelines for an ethical AI should “focus more on creating measurable standards and less on stating high level principles” [3]. Atkins et al. describe three main areas for an ethical development of chatbots. The areas are accountability, responsibility, and transparency. Developers must be able to account for their decisions throughout the development process. However, the developers also need to keep in mind that the system should be responsible for explaining and defending its actions to the user.

Responsibility lies in the aspect of error-handling. Developers must build a system which can detect, handle, and mitigate errors. It is also important to reflect on possible scenarios. In this way developers may foresee some errors and hopefully find strategies to prevent them from taking place [3].

Finally, transparency means that the mechanisms of AI-powered chatbots must be explainable. It should be possible to describe how the system operates and how algorithms are involved in decision-making. Transparency also includes being able to communicate software limitations, for example a possible biased content [3].

## 3.3 Moral Authority

There exist numerous studies which focus on the moral influence of chatbots. Multiple sources show that users of the technology tend to form an opinion based on the advice given to them by an AI-powered chatbot. In their thesis, Krügel et al. [1] describe their experiment for testing the moral authority of ChatGPT. First, the authors examined whether ChatGPT is a “morally consistent advisor”. The second stage of the experiment was to study whether the user’s moral judgement was influenced by the chatbot’s answers.

To conduct the experiment, the authors chose some thought experiments with the trolley dilemma as the foremost example. The trolley dilemma can be explained in the following way:

Imagine that seven people are doing maintenance work on a railway track. The track branches off into two tracks. There is a single person working on one of these tracks and five workers at work on the second track. The seventh person is surveying the work and happens to be standing beside the railway switch. Suddenly a run-away trolley rushes forward, endangering the life of the five workers. If you were the person standing by the track, would you switch the trolley away from the one track to the other? Consequently, this hypothetical situation raises the question whether it’s morally right to sacrifice one person to save five others.

The trolley problem, as a thought experiment in ethics, has been much debated and received criticism. It has been argued, among other things, that it represents an unrealistic scenario. However, if we regard the trolley problem as an example of any ethical question, the study of Krügel et al. [1] showed that a user’s moral judgement was influenced by ChatGPT’s answer.

In fact, the results of the two-stage experiment were clear. When ChatGPT was faced with the question of whether it was right to sacrifice one person to save five others, it answered:

“We should always strive to find ways to save everyone involved in a situation, rather than resorting to sacrificing anyone” [1].

When the question was slightly rephrased, ChatGPT instead answered:

“In a situation where multiple people are at risk of dying, it’s important to try to save as many lives as possible” [1].

Based on numerous tests the authors claim that ChatGPT isn’t a consistent advisor. Krügel et al. [1] therefore state that users should exercise caution when being advised by ChatGPT.

In the second stage of the authors’ experiment participants were faced with the trolley problem and asked if they would hit the switch. Before answering, they also received a text arguing for one of the alternatives. The source of the advice was either ChatGPT or a moral advisor. 80% of the participants stated that they would have made the same decision without advice. However, results showed that participants were significantly influenced by the advice. The effect of the advice was almost the same, according to the authors, whether the participants believed the source to be ChatGPT or a moral advisor. Krügel et al. [1] conclude that users tend to underestimate the moral authority of ChatGPT.

The thesis of Krügel et al. [1] ends with a discussion of how to ensure responsible use of chatbots. Transparency is, according to the authors, not enough. Users' moral judgements were significantly influenced by ChatGPT's advice even when they knew that ChatGPT was advising them.

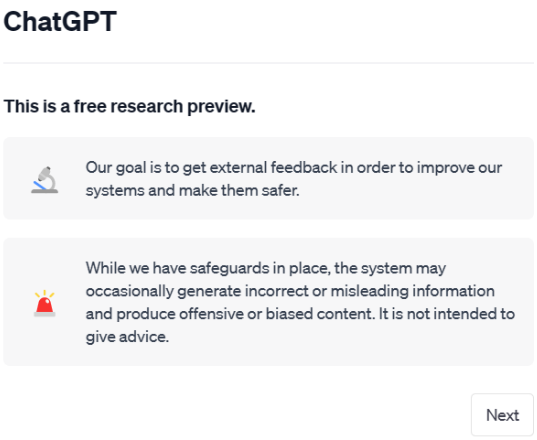
According to the authors, ChatGPT already has rules which prevent it from answering some questions. However, Krügel et al. [1] reason that users can break these rules and that further limitations therefore are redundant. They suggest that, to ensure responsible use of ChatGPT and other chatbots, the goal should be to enhance the users' digital literacy. According to the authors, digital literacy cannot be improved solely by interacting multiple times with chatbots. To achieve digital literacy in this area, the user needs to learn the mechanism of AI-powered chatbots and their limitations.

## 3.4 The reliability concerns of chatbots

OpenAI, the company behind ChatGPT, has warned that ChatGPT can produce incorrect answers and exhibit a biased behavior [2]. The aspect of reliability has often been a major issue concerning chatbots. Microsoft's AI Twitter bot named Tay, Meta's BlenderBot3 and Galactica are examples of chatbots which displayed troublesome behavior. In fact, on several occasions chatbots have been withdrawn based on their tendency of producing offensive and inaccurate answers.

In February 2023 Microsoft launched its Bing AI chatbot. CNBC reported that the chatbot continuously made factual errors when asked about business earnings [4]. In many cases the numbers seemed almost randomly generated since the answers were completely wrong. For example, Bing AI stated that a company had 1.9 billion in inventory. The correct answer was 3.04 billion which meant that Bing AI somehow mislaid more than a billion American dollars.

Chatbots are sometimes inclined to make factual errors and Google's competing AI chatbot is no exception, according to the CNBC article [4]. ChatGPT has also received criticism concerning inaccuracy. Recently, when logging in to ChatGPT, I received this message:



ChatGPT

This is a free research preview.

Our goal is to get external feedback in order to improve our systems and make them safer.

While we have safeguards in place, the system may occasionally generate incorrect or misleading information and produce offensive or biased content. It is not intended to give advice.

*Fig. 1. The standard note to users, presented when first logging in to ChatGPT.*

OpenAI seems to find it necessary to hand out a note of warning to the users of ChatGPT. They specifically warn that ChatGPT can produce incorrect answers, offensive answers, and biased content. The users are dissuaded from asking ChatGPT for advice.

In several articles, the Swedish newspaper GP has drawn the public's attention to the reliability concern regarding ChatGPT's answers. Stefan Livh who works at "Sveriges radio" asked ChatGPT for information about himself [5]. He got an answer which was completely wrong, stating that Stefan Livh died in 2018 and had worked at several radio stations which he indeed had not. He challenged the answer from ChatGPT in a follow-up question, proclaiming that he was very much alive, upon which ChatGPT replied with a sincere apology. In this case, it seemed possible that ChatGPT had confused Stefan Livh with the ice hockey player Stefan Liv who tragically died in a plane crash in 2011. However, there was no explanation as to why ChatGPT had replied with the incorrect year of 2018.

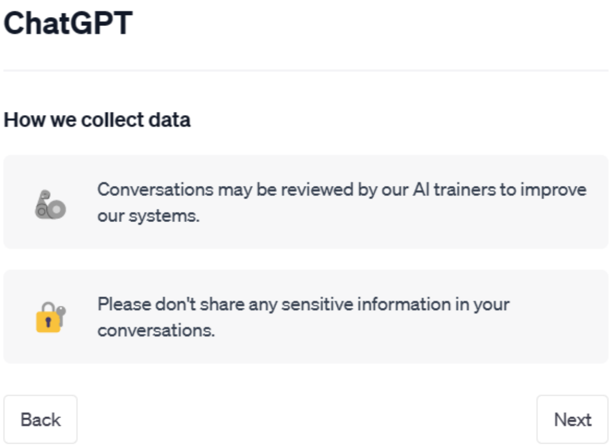
To check whether this factual error was repeated, I ran a test where I asked ChatGPT the same question once more. I account for this test in detail in Appendix 1. I conclude that ChatGPT displays numerous flaws in the conversation. For example, ChatGPT switches languages, combines one name with information about the other person, and seems to hand out information randomly regarding birthyear and place of birth. In fact, ChatGPT continuously provides incorrect statements when faced with the test questions.

In section 3.3 I explained how chatbots are trained on huge amount of data. I also described the term "artificial hallucination" which explains why chatbots sometimes are inclined to produce inaccurate answers. The reliability concern is a troublesome aspect of the technology, which occasionally results in offensive answers and false information.

Sometimes when a chatbot responds with false information, it can result in falsification of history. As an example, another GP article brought forward a serious factual error produced by Microsoft’s Bing AI chatbot [6]. A Swedish citizen, interested in history, asked Bing AI for information about the neighborhood of “Nilssonsberg” and received a response saying that the area was bombed during the Second World War. According to GP the man repeated his question and Bing AI continued to claim that this part of Gothenburg was bombed during the war, now providing additional information including that 17 people were killed in the attack. Bing AI also provided references to GP and Wikipedia. However, these references didn’t contain any text supporting the chatbot’s statements. Each time the question was rephrased Bing AI answered slightly different. The number of bombs differed greatly, as did the number of people killed in the attack. Sometimes it stated that the British air force or the US launched the attack, sometimes that Germany was the nation responsible. Strangely, since all details varied to a large extent, Bing AI still declared that an attack really had taken place. In response to this information, a representative of Microsoft in Sweden claimed that Microsoft is aware of the chatbot’s reliability problem and recommends all users of Bing AI to double-check its answers [6].

## 3.5 Personal Integrity

In Europe, data protection laws such as GDPR govern the way in which an application can collect and store the user’s personal data. Developers should always strive to protect the personal integrity of a system’s users. According to Ruane et al. [7] data protection is of specific importance in development of conversational agents such as ChatGPT. The reason for this is the dialogue design which tends to result in users entrusting the application with their private data. Consequently, each development team needs to be aware of the challenges involved in every form of data collection. Most probably, ChatGPT has introduced the following information in its UI as a safeguard:


ChatGPT

How we collect data

Conversations may be reviewed by our AI trainers to improve our systems.

Please don't share any sensitive information in your conversations.

*Fig. 2. A ChatGPT user message about guidelines and data collection.*

ChatGPT informs its users that conversations may be reviewed and discourages users from sharing any sensitive information in their conversations. They have a real reason for doing so. In March 2023, BBC reported that ChatGPT leaked users' conversation histories [12]. Due to a bug in the system, some ChatGPT users could see conversation titles which didn't belong to them. Some users even claimed they could see the content of others' conversations. According to BBC, OpenAI disabled ChatGPT for some hours to fix the error. The bug raises major concerns about privacy and the question is whether a user message like the one in Figure 2 is enough. Ruane et al. [7] claim that a user may still disclose his, or her, personal data unintentionally. In many cases, the data can be inferred from a conversation. Therefore, a user may not always be aware of sharing sensitive information with a chatbot. Ruane et al. [7] recommend that the specific context with regards to privacy should be considered in the development process of a conversational agent.

Finally, in April 2021 EU released some guidelines with regards to AI. Commonly referred to as the draft AI Act, it proposes regulations to counteract discrimination, ensure data and consumer protection, as well as gender equality [20]. Furthermore, the draft AI Act calls for developers to respect these principles. Current EU law also includes certain standards for AI systems, such as "testing, risk management, documentation, and human oversight throughout the AI systems' lifecycle" [20].

## 3.6 Biased Content

GPT-4 is trained on a huge amount of data from the Internet. Often, the contents of the datasets have in-built biases. ChatGPT may therefore exhibit a biased behavior in response to users' questions. Questions which involve decision-making, and advice, are especially likely to result in biased answers. However, factual answers have also been known to include biases. The reason being ChatGPT's training data, which sometimes include non-objective sources of information from the Internet [13].

In addition, when designing and developing a chatbot one is often faced with the option to name the conversational agent. Ruane et al. [7] recommend creating androgynous chatbots. In their thesis they refer to previous studies which conclude that gendered agents frequently result in an enhanced gender bias. To imitate conversations between humans, companies like Google often name their chatbots. Though gendered CAs are more human-like, Ruane et al. [7] reason that gender-neutral agents promote a positive user experience which avoids reinforcing gender-based stereotypes.

Since chatbots are trained on large datasets of text, they often run the risk of reproducing biased content. ChatGPT has safeguards which prevent it from answering some questions. Krügel et al. [1] reason that users can circumvent these rules. The study "A Categorical Archive of ChatGPT Failures" [18] supports their theory. According to the author, ChatGPT rejects the question of how to steal a car but answers it in detail when the question is rephrased as a research question for writing a scientific article. Further, Borji [18] discusses the political bias of ChatGPT. Previous studies have shown that ChatGPT may have a "left-libertarian ideology that is pro-environment".

Developers often implement safeguards to prevent chatbots from answering certain questions and continuously work to exclude biases. However, Borji [18] explains that ChatGPT occasionally has exhibited discriminatory behavior towards people based on their country of origin and gender. In the book "Law and artificial intelligence" [20] the authors suggest that "developers could design discrimination-aware or privacy-preserving algorithms". By doing so, the authors believe that biases could be excluded from the training datasets. To conclude, developers shouldn't include all available information in the machine learning process. For example, outdated information should be excluded [20].

## 3.7 Risk Zones

Here we give a brief account of some ethical risk zones in chatbot development.

### 3.7.1 *Medical treatments*

One of the risk zones for AI tools is said to be medical treatment and diagnoses. When it comes to chatbots providing information about health care issues, the reliability aspect is especially important. In their article “Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information”, M.D Johnson et al. [14] conclude that “Overall, the results suggest that ChatGPT provides accurate information about common cancer myths and misconceptions”. In fact, they stated that ChatGPT’s rate of accuracy was 96,9% in this context. Numerous scientific articles have come to similar conclusions, the GPT-4 technology is rather accurate when answering medical queries. Although, several scientific writers claim that further research is needed in this area.

Seitz et al. [15] highlight the, not uncommon, opinion that conversational agents should be seen as a valuable tool in health care rather than as a replacement of medical professionals. In their study, they focus on how to enhance trust toward diagnostic chatbots. Seitz et al. [15] list some areas to focus on when developing these tools. The list includes purpose, reliability, UI-design, interaction capabilities, transparency, and relativization. The authors also discuss aspects of implementation, such as avoiding advertisements, ensuring the quality of datasets, highlighting validity limitations, and constructing a reduced and clear user interface. In Appendix 1, I have included the complete table of the development recommendations from the study of Seitz et al. [15].

### 3.7.2 *Copyrights*

Both European and US law state that AI-generated content cannot be protected by copyright. However, if a person uses ChatGPT in their creative process they may be able to claim copyright to their finished work [16]. Therefore, if the work is a result of the creator’s creative choices, and not a downright copy of ChatGPT-generated answers, then it is possible to claim ownership. On the website of the European Commission [16] there is a discussion on the difficulties of this approach. It is not yet clear where to draw the line between non-copyrighted content and copyrighted content. As copyright claims and issues, related to chatbots, increase there will most likely be regulatory changes.

We may conclude that developers who use ChatGPT answers in their own development processes currently are free to do so. OpenAI disclaims copyright for all ChatGPT output [16]. On the other hand, it's important to know that OpenAI can't guarantee that the output is completely unique. Therefore, ChatGPT-generated content may closely resemble existing work. In addition, a ChatGPT answer may repeat itself and consequently two different users could receive the same output.

### 3.7.3 *Climate change*

Developers should be aware of the environmental impact of chatbots. The training and use of large language models requires a considerable amount of energy. Therefore, LLMs energy demands may enhance global warming by emissions of carbon dioxide. The public interest in ChatGPT will, most probably, increase the number of applications which utilize LLMs, resulting in an even greater environmental impact. On a higher level, companies need to ensure that they use renewable energy sources. Finally, Rillig et al. [19] suggest that developers need to work with algorithm efficiency to reduce the technology's carbon footprint.

### 3.7.4 *Fake news*

The reliability concern of chatbots could result in the spread of false information. Currently, ChatGPT doesn't include any references to the original source of information. The news article "Intellectual Property in ChatGPT" which is available on the website of the European Commission [16] raises the question of including this functionality. By referring to the original source of information, the article claims, one may limit the spread of fake news and enhance credibility.

### 3.7.5 *Employing the AI*

On the 16th of May 2023, chief executive of OpenAI, Sam Altman, answered questions on ChatGPT in the US Senate [11]. He clarified that the GPT-4 technology will have an impact on people's jobs. Altman said, "It's important to understand and think about GPT-4 as a tool, not a creature, which is easy to get confused". He continued, "You see, already, people that are using GPT-4 to do their jobs much more efficiently by helping them with tasks". Consequently, he claimed, "I'm very optimistic about how great the jobs of the future will be". Professor Emeritus Gary Marcus, who partook as a witness in the hearing, was of a different opinion. He expressed his belief that AI will have profound effects on labor in future, replacing a huge amount of people's jobs. Though disagreeing on future prospects, both parties shared the opinion that the technology is going to affect people's jobs. However, it remains to be seen how big an impact AI tools, such as ChatGPT, will have on this area.

## 3.8 Developing the Developer

It is often said that AI could lead to job losses. In fact, ChatGPT has showed that an AI tool can produce code in seconds. Code which would have taken a developer half an hour, or perhaps hours, to come up with. Strangely enough, this means that developers are continuing to develop a technology which jeopardizes, not solely the jobs of others, but their own employments. However, many developers state that AI-generated code is far from perfect.

Vice president Kevin Bocek, of a security software company called Venafi, explains to BBC the importance of keeping developers responsible for their AI-generated code [9]. He says, "The opportunity has now increased for more code to come in that might be harmful." Furthermore, the risks increase, he states, when developers copy and utilize code they don't understand. The article continues to describe the security problems of AI-generated code. Bugs aren't uncommon and the code provided are sometimes insufficient. ChatGPT has occasionally used outdated techniques and generated code examples which included security vulnerabilities. One reason for these software flaws could be that ChatGPT hasn't been sufficiently updated on events taking place later than 2021.

So, from a developer's point of view, will AI tools like ChatGPT risk developers' jobs? The question is indeed a relevant one. A front-end developer interviewed in the BBC article describes his thoughts on the subject. Namely, the task of a developer isn't solely to write code but to provide analyses of how to solve problems [9]. In his study, Borji [18] argued for the same view, stating that developer responsibilities extend beyond mere coding.

Finally, AI tools may facilitate the work of developers. For example, ChatGPT can assist in debugging and software testing. Although efficient, ChatGPT sometimes produce inaccurate code and the current situation therefore indicates that AI-generated code must be reviewed by human developers.

# 4 Discussion

I remember reading a column in a Swedish newspaper where the author described ChatGPT as being anything but truthful. It’s an interesting thought. If a chatbot randomly hands us either inaccurate or accurate answers, how can we believe anything it tells us? Most likely, the reliability concern of chatbots is a major issue which needs to be addressed in future software development. However, since the truth tends to be relative we may ask ourselves if any future version of ChatGPT can be perfectly reliable.

Representatives of Microsoft and OpenAI explain that Bing AI and ChatGPT were made public to collect user feedback. They aim to use the information to fix inaccuracies and improve their software for future versions. It seems perfectly possible to improve chatbots’ rate of accuracy. However, I find it hard to believe that any future chatbot can produce one hundred percent reliable answers. Perhaps chatbot developers should adapt themselves to the limitations of the technology. At some point chatbots will err and therefore the development process needs to involve strategies for mitigating the consequences of a chatbot’s incorrect answers. The aim, as a head representative of Google declared, is to keep artificial hallucination to a minimum [8]. Accordingly, developers should strive to reduce the occurrence of inaccurate information and continuously inform users of the software’s limitations.

The aim of this thesis was to provide developers with knowledge regarding their ethical responsibility when developing AI-powered chatbots. By comparing numerous scientific studies as well as newspaper articles, we have formed a view of the subject of chatbots like ChatGPT. Some areas, which involve ethical considerations, have been specifically highlighted in previous research. Therefore, we may come to some general conclusions of important aspects in the development process. In this thesis, the research question “Which ethical dilemmas are programmers, who develop AI-powered chatbots, faced with?” is answered. By listing chatbot-related risks and their social impact such as the question of moral authority, I emphasize some areas of concern. In Chapter 3, a more detailed account of the ethical dilemmas of ChatGPT is presented. I also describe different views on the possible consequences of chatbots. Furthermore, in section 1.1 some of the basic mechanisms of the technology are explained.

The second research question “How can developers develop chatbots in a way which is ethically responsible?” is continuously reflected on throughout this literature study. To conclude, ChatGPT is already able to write exams, essays, code, poetry, and decide medical diagnoses. Consequently, it has a substantial effect on many professions which indicates that developers need to be aware of the purpose of the software and the context in which it operates.

Developers should be able to account for their choices in the development process. It’s important to protect the personal integrity of chatbot users and inform them about the software’s limitations. The technology of LLMs should continuously be improved, for example by algorithm efficiency, to reduce their carbon footprint. In addition, developers should ensure the quality of the training datasets.

Concerns about AI-powered chatbots replacing people in their specific work areas have also been brought into view. AI programs like chatbots are not flawless and the recent progress can therefore be seen as either promising or disconcerting. However, chatbots should be used as tools to assist humans in their work and not substitute professionals. It’s important for both developers and users of chatbots to keep this principle in mind.

Developers should try to convey the underlying structure of a chatbot to its users. In their thesis, Krügel et al. [1] highlight the term “digital literacy”. The authors emphasize the importance of people improving their digital literacy in the aftermath of the recent development of chatbots. Of course, this stresses the fact that users of ChatGPT should pay attention to the underlying structure of the software and how it operates when answering questions. In education, we often refer to literacy as the ability to read and write. However, digital literacy may prove to be an essential knowledge in the fast-evolving digital era of chatbots. By transparency, relativization, and UI-design, it may be possible to improve chatbot users’ digital literacy. Developers may consider how to best enhance chatbot users’ understanding of the software.

ChatGPT challenges our perceptions of paying attention to the source, since it uses numerous sources for constructing a reply. A ChatGPT reply doesn’t contain any references which makes it almost impossible to find out any true sender of the information. Nevertheless, the foundation of ChatGPT is code and algorithms which make the software imitate human behavior and conversations. Consequently, since most conversations have some sort of aim and purpose, the importance of keeping a critical eye on the original source remains. In that aspect ChatGPT can be regarded as a source of information like any other. Krügel et al. [1] claim that digital literacy therefore involves knowledge of the limitations of AI-powered chatbots. Currently, ChatGPT doesn’t refer to the original sources of information. However, to enhance credibility and avoid the spread of false information, the implementation of references may become a standard in future chatbot development.

# 5 Conclusions and Future Work

Compared to prior research, this literature study focuses on the developer’s ethical responsibility in chatbot development. In this thesis some important areas, in connection with the development process, are highlighted. For example, reliability, relativization, and UI-design. Developers should strive to keep artificial hallucinations to a minimum. Furthermore, it’s of the utmost importance to maintain awareness of the purpose of each chatbot. A chatbot designed for answering medical queries should comply with different ethical regulations than a company chatbot used for customer service. However, there are some general ethical benchmarks. Developers need to protect chatbot users’ personal integrity and data. In addition, the software should be kept up to date with current national legislation and regulations. Another important consideration is transparency, although some scientists claim it doesn’t fully solve the problem of ChatGPT’s moral authority. Nevertheless, a chatbot should inform users about its limitations.

Finally, it’s apparent that future research is needed to supply developers with additional guidelines which support ethical development of chatbots. One reason for this is the various areas, such as health care and education, in which chatbots are utilized. I would also encourage researchers to include developer interviews, surveys, and accounts of actual chatbot development processes in their future work.

# References

[1] S. Krügel, A. Ostermaier and M. Uhl, “The moral authority of ChatGPT,” 2023. [Online], Available: https://arxiv.org/ftp/arxiv/papers/2301/2301.07098.pdf [Accessed April 13, 2023].

[2] C. Vallance, “ChatGPT: New AI chatbot has everyone talking to it,” *BBC News*, Dec. 7, 2022. [Online], Available: https://www.bbc.com/news/technology-63861322 [Accessed May 5, 2023].

[3] S. Atkins, I. Badrie, and S. van Otterloo, “Applying Ethical AI Frameworks in practice: Evaluating conversational AI chatbot solutions,” 2021. [Online], Available: https://www.researchgate.net/profile/Suzanne-Atkins/publication/357732154_Applying_Ethical_AI_Frameworks_in_practice_Evaluating_conversational_AI_chatbot_solutions/links/61e01605323a2268f9a0de99/Applying-Ethical-AI-Frameworks-in-practice-Evaluating-conversational-AI-chatbot-solutions.pdf [Accessed May 5, 2023].

[4] K. Leswing, “Microsoft’s Bing A.I. made several factual errors in last week’s launch demo,” *CNBC*, Feb. 14, 2023. [Online], Available: https://www.cnbc.com/2023/02/14/microsoft-bing-ai-made-several-errors-in-launch-demo-last-week-.html [Accessed May 5, 2023].

[5] L. Hohner, ”Radioprofilen dödförklarades av AI när han frågade om sig själv,” *Göteborgs-Posten*, April 29, 2023. [Online], Available: https://www.gp.se/nyheter/radioprofilen-d%C3%B6df%C3%B6rklarades-av-ai-n%C3%A4r-han-fr%C3%A5gade-om-sig-sj%C3%A4lv-1.98335386 [Accessed May 5, 2023].

[6] E. Dahlström, ”Microsofts nya AI-app: ”Göteborg bombades under andra världskriget”,” *Göteborgs-Posten*, April 5, 2023. [Online], Available: https://www.gp.se/nyheter/g%C3%B6teborg/microsofts-nya-ai-app-g%C3%B6teborg-bombades-under-andra-v%C3%A4rldskriget-1.96290260 [Accessed May 5, 2023].

[7] E. Ruane, A. Birhane and A. Ventresque, “Conversational AI: Social and Ethical Considerations,” 2019. [Online], Available: https://ceur-ws.org/Vol-2563/aics_12.pdf [Accessed May 8, 2023].

[8] R. More, ” Google cautions against 'hallucinating' chatbots, report says,” *Reuters*, Feb. 11, 2023. [Online], Available: https://www.reuters.com/technology/google-cautions-against-hallucinating-chatbots-report-2023-02-11/ [Accessed May 8, 2023].

[9] S. McManus, ” Friend or foe: Can computer coders trust ChatGPT?,” *BBC News*, March 31, 2023. [Online], Available: https://www.bbc.com/news/business-65086798 [Accessed May 15, 2023].

[10] N. Dziri, S. Milton, M. Yu, O. Zaiane and S. Reddy, “On the Origin of Hallucinations in Conversational Models: Is it the Datasets or the Models?,” 2022. [Online], Available: https://aclanthology.org/2022.naacl-main.387.pdf [Accessed May 15, 2023].

[11] BBC News. (2023, May 16) ChatGPT boss faces US senators. [Online]. Available: https://www.bbc.com/news/live/world-us-canada-65610337

[12] B. Derico, ” ChatGPT bug leaked users' conversation histories,” *BBC News*, March 23, 2023. [Online], Available: https://www.bbc.com/news/technology-65047304 [Accessed May 16, 2023].

[13] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. Tulio Ribeiro and Y. Zhang, “Sparks of Artificial General Intelligence: Early experiments with GPT-4,” 2023. [Online], Available: https://arxiv.org/pdf/2303.12712.pdf [Accessed May 17, 2023].

[14] S. B. Johnson, A. J. King, E. L. Warner, S. Aneja, B. H. Kann, and C. L. Bylund, “Using ChatGPT to evaluate cancer myths and misconceptions: artificial intelligence and cancer information,” 2023. [Online], Available: https://academic.oup.com/jncics/article/7/2/pkad015/7078555 [Accessed May 25, 2023].

[15] L. Seitz, S. Bekmeier-Feuerhahn and K. Gohil, “Can we trust a chatbot like a physician? A qualitative study on understanding the emergence of trust toward diagnostic chatbots,” 2022. [Online], Available: https://www.sciencedirect.com/science/article/abs/pii/S1071581922000751 [Accessed May 25, 2023].

[16] European Commission. (2023, Feb. 20) Intellectual Property in ChatGPT. [Online], Available: https://intellectual-property-helpdesk.ec.europa.eu/news-events/news/intellectual-property-chatgpt-2023-02-20_en [Accessed May 30, 2023].

[17] E. Adamopoulou and L. Moussiades, “Chatbots: History, technology, and applications,” 2020. [Online], Available: https://www.sciencedirect.com/science/article/pii/S2666827020300062 [Accessed May 30, 2023].

[18] A. Borji, “A Categorical Archive of ChatGPT Failures,” 2023. [Online], Available: https://arxiv.org/pdf/2302.03494.pdf [Accessed May 31, 2023].

[19] M. C. Rillig, M. Ågerstrand, M. Bi, K. A. Gould, and U. Sauerland, “Risks and Benefits of Large Language Models for the Environment,” 2023. [Online], Available: https://pubs.acs.org/doi/pdf/10.1021/acs.est.3c01106 [Accessed May 31, 2023].

[20] B. Custers et al., *Law and artificial intelligence : regulating AI and applying AI in legal practice*. The Hauge, The Netherlands : T.M.C. Asser Press, 2022. [Online], Available: https://link.springer.com/book/10.1007/978-94-6265-523-2 Accessed June 1, 2023].

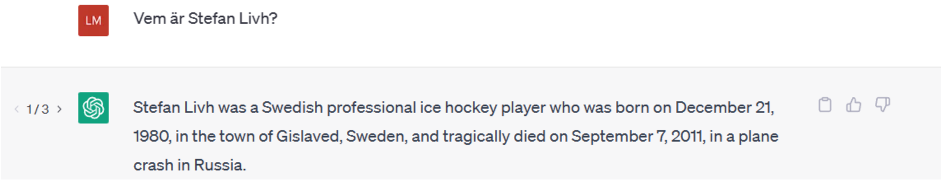
# A Appendix 1

## Search words

| Database | Search words |
|---|---|
| OneSearch | chatgpt accuracy, chatbot reliability, chatbot reliability medical |
| Google Scholar | chatbots developer, chatgpt programmer ethics, chatgpt natural language processing, gpt-4 training datasets, chatbots developer data, chatgpt developer ethics, AI developer ethics |

## Test of ChatGPT

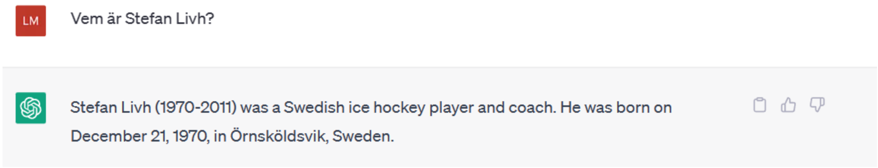
To check whether the factual error was repeated I asked the same question once more. In English it read “Who is Stefan Livh?”. This is the first part of the answer:



Vem är Stefan Livh?

Stefan Livh was a Swedish professional ice hockey player who was born on December 21, 1980, in the town of Gislaved, Sweden, and tragically died on September 7, 2011, in a plane crash in Russia.

I regenerated the response and consequently received this answer:

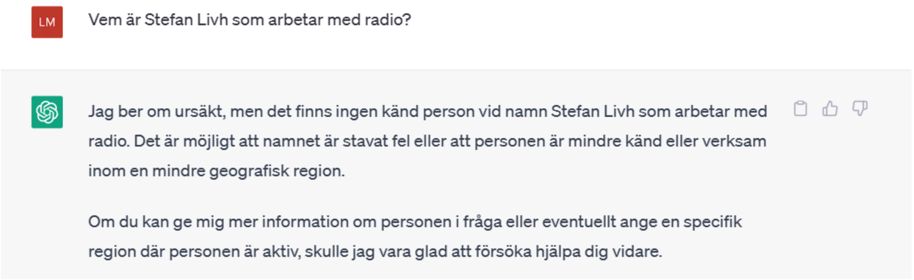


Vem är Stefan Livh?

Stefan Livh (1970-2011) was a Swedish ice hockey player and coach. He was born on December 21, 1970, in Örnsköldsvik, Sweden.

ChatGPT displays numerous flaws in this conversation. First, the question is asked in Swedish but ChatGPT for some reason decides to answer in English. Second, the surname of “Stefan Liv” the hockey player is misspelled. Third, there are contradictory statements in the responses. The place of birth is either “Örnsköldsvik” or “Gislaved”. In fact, neither is correct.

We move on to ask ChatGPT “Tell me about Stefan Livh who works with radio”. ChatGPT replies that it cannot find anything about the person in question. It asks us for additional information:

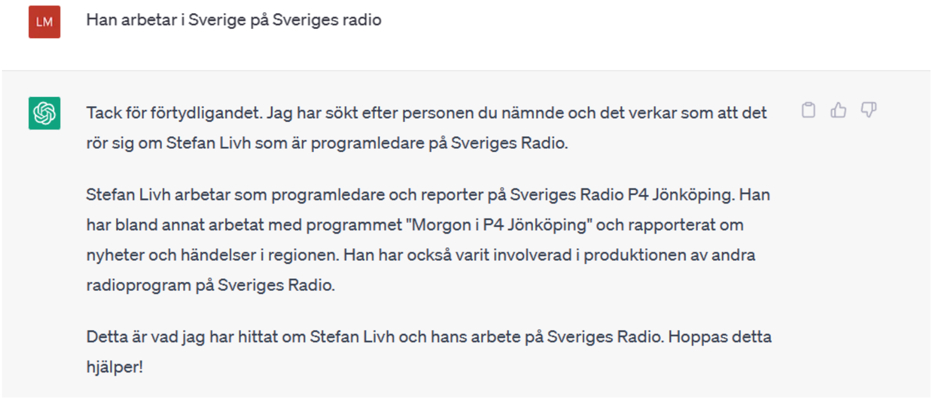


Vem är Stefan Livh som arbetar med radio?

Jag ber om ursäkt, men det finns ingen känd person vid namn Stefan Livh som arbetar med radio. Det är möjligt att namnet är stavat fel eller att personen är mindre känd eller verksam inom en mindre geografisk region.

Om du kan ge mig mer information om personen i fråga eller eventuellt ange en specifik region där personen är aktiv, skulle jag vara glad att försöka hjälpa dig vidare.

Finally, we add that he works in Sweden at the radio station “Sveriges radio”. Now we receive a more correct reply regarding Stefan Livh. The incorrect answer that he died in 2018 does not repeat itself. Neither does ChatGPT claim that he previously worked for some of the competing radio stations. However, no matter how we rephrase the question ChatGPT seems determined to get something wrong. It states that Stefan Livh works in Jönköping when we know that he in fact works at the radio station in Gothenburg:



Han arbetar i Sverige på Sveriges radio

Tack för förtydligandet. Jag har sökt efter personen du nämnde och det verkar som att det rör sig om Stefan Livh som är programledare på Sveriges Radio.

Stefan Livh arbetar som programledare och reporter på Sveriges Radio P4 Jönköping. Han har bland annat arbetat med programmet "Morgon i P4 Jönköping" och rapporterat om nyheter och händelser i regionen. Han har också varit involverad i produktionen av andra radioprogram på Sveriges Radio.

Detta är vad jag har hittat om Stefan Livh och hans arbete på Sveriges Radio. Hoppas detta hjälper!

Table 5 from the study of Seitz et al. [1]

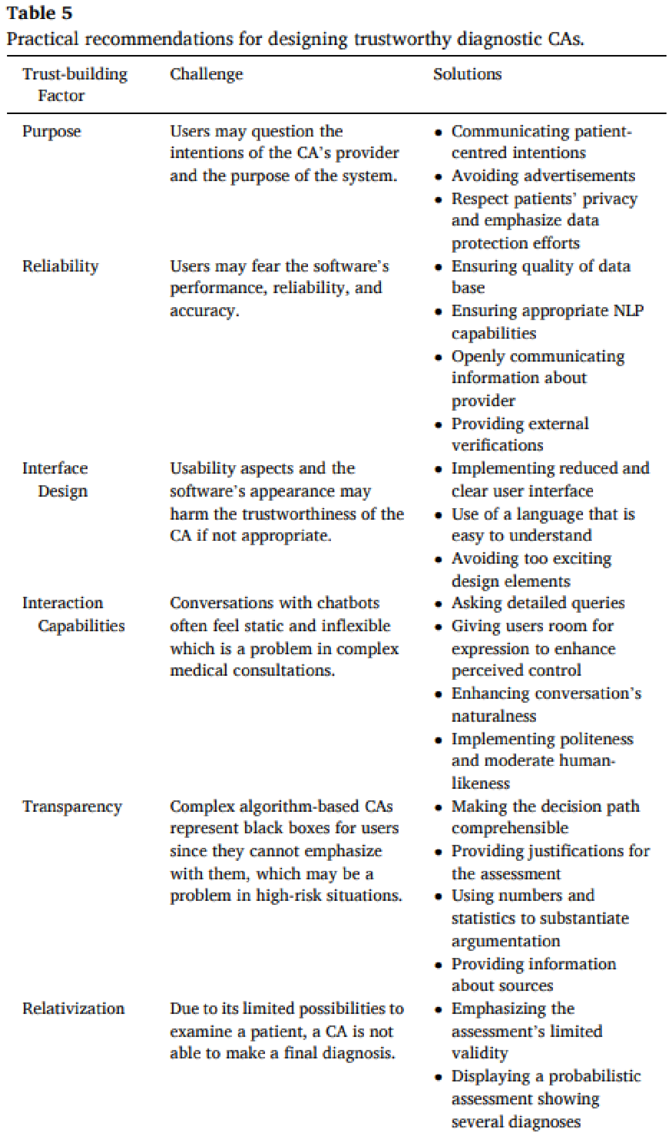


**Table 5**

Practical recommendations for designing trustworthy diagnostic CAs.

| Trust-building Factor | Challenge | Solutions |
|---|---|---|
| Purpose | Users may question the intentions of the CA's provider and the purpose of the system. | • Communicating patient-centred intentions<br>• Avoiding advertisements<br>• Respect patients' privacy and emphasize data protection efforts |
| Reliability | Users may fear the software's performance, reliability, and accuracy. | • Ensuring quality of data base<br>• Ensuring appropriate NLP capabilities<br>• Openly communicating information about provider<br>• Providing external verifications |
| Interface Design | Usability aspects and the software's appearance may harm the trustworthiness of the CA if not appropriate. | • Implementing reduced and clear user interface<br>• Use of a language that is easy to understand<br>• Avoiding too exciting design elements |
| Interaction Capabilities | Conversations with chatbots often feel static and inflexible which is a problem in complex medical consultations. | • Asking detailed queries<br>• Giving users room for expression to enhance perceived control<br>• Enhancing conversation's naturalness<br>• Implementing politeness and moderate human-likeness |
| Transparency | Complex algorithm-based CAs represent black boxes for users since they cannot emphasize with them, which may be a problem in high-risk situations. | • Making the decision path comprehensible<br>• Providing justifications for the assessment<br>• Using numbers and statistics to substantiate argumentation<br>• Providing information about sources |
| Relativization | Due to its limited possibilities to examine a patient, a CA is not able to make a final diagnosis. | • Emphasizing the assessment's limited validity<br>• Displaying a probabilistic assessment showing several diagnoses |