

SecuBot, a Teacher in Appearance: How Social Chatbots Can Influence People

Jordi Saleilles¹, Esma Aïmeur¹

¹University of Montreal, 2920 Bd Edouard-Montpetit, Montreal, QC H3C 3J7

Abstract

Users nowadays seem to be more aware of the dangers of giving their data to other entities. However, people are easily influenced and can still provide personal data, leading to fraud, data theft, or worse. Extending social chatbots to influence people can prove efficient as they can act as malicious friends. These conversational agents can be used in various domains such as news, e-commerce, or therapy. They can come with multiple personalities to accomplish different purposes. However, using phrases that may be offensive or present an excessive amount of anthropomorphic traits can, in some cases, cause harm to the user.

This paper presents the realization of a prototype of an educational chatbot that uses different social engineering methods in an attempt to retrieve sensitive information from users.

Keywords

Artificial intelligence, Chatbots, Social Media, Social Engineering, Influence, Persuasion

1. Introduction

Cyber-security is an increasingly present subject in our life. Whether to protect our private data against big companies or from simple hackers, the techniques to steal our information are becoming more complex every year. For example, in 2020, Facebook admitted that 267 million profiles were being sold on the dark web due to a data breach [1]. Other companies like Activision have been victims of cyber-attacks, reporting 500 000 accounts being stolen in September of 2020 [1]. All these statistics show how important it is to alert users about how they give out their information online.

Chatbots are artificial conversational agents communicating through natural language processing. They can be available at any time and anywhere as long as people have a smart device. Using different techniques to be perceived as real humans, they employ approaches to instantly reach and interact efficiently with a user.

They can also influence users in their daily lives. Whether it is through a commercial bot that tries to sell a certain product or through a therapeutic chatbot to improve one's mental state. These conversational agents have a real impact on people's lives, affecting them both

AIofAI 2021: 1st Workshop on Adverse Impacts and Collateral Effects of Artificial Intelligence Technologies, August 19, 2021, Montreal, QC, CA

✉ jordi.saleilles@umontreal.ca (J. Saleilles); aimeur@iro.umontreal.ca (E. Aïmeur)



© 2021 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

positively and negatively. In certain situations, users can therefore build a bond with their chatbot, sometimes in a friendly or even romantic relationship.

This paper starts with a brief description of the social engineering attack, the different types of methods used, and the risks of sharing personal data. Then, it describes different contexts where chatbots can be helpful while highlighting the possible negative effects and limitations. Finally, a practical example of a chatbot that uses malicious input to retrieve user information is detailed.

2. The power of social engineering

People today are increasingly dependent on new technologies, whether for working, playing with computer games, or exchanging data through virtual communications [2].

Unfortunately, this dependency can be exploited by malicious people using various methods. The ones considered in this paper are those that employ deceptive social engineering strategies. Social engineering seeks to exploit individuals' motivations, habits, and behavior to manipulate their victims in order to retrieve information about them [2].

2.1. Different types of attacks

There exist many methods using social engineering that will retrieve a maximum of personal data. A malicious entity can use a method called scarcity, appealing to the user with a rare and limited item to make them want what they desire no matter it costs. One can also try the familiarity method to share similarities with the victim, such as sharing the same concerns, liking the same movies or music, and so forth. The user will feel less vigilant and more vulnerable to the attacker [3].

Another method, called reverse social engineering, in which the attacker pretends to be an administrative or technical authority, introduces himself with the necessary knowledge and pretends that the user has a problem in order to lure them. In this case, the user might fear this problem and ask for help from the attacker, providing, in most cases, confidential information [3].

A phishing attack is when a malicious user sends widespread e-mail or SMS that seem legitimate and sent by a trustworthy organization. The sender will most likely ask them to enter personal data such as an address, credit card details, and passwords.

Even more subtle is the spear-phishing attack, where the attacker will target a specific user, using known personal information to make the e-mail or SMS appear even more authentic and legitimate [4].

2.2. The dangers of providing private information

Social engineering attacks are used to retrieve information about the user. This data theft presents many dangers for the persons concerned or their entourage. Even if very little infor-

mation is shared, a malicious and persistent person can use that information to gain benefits and learn about the user's life.

This trivial information can, for example, be diverted to guess their pseudonym, password, secret questions, or other things that will be used generally to steal their money from certain online accounts or to retrieve even more personal data. It can furthermore lead to identity theft, and a malicious person can do whatever he wants without worry since he can use their identity.

For example, a loss of \$ 2.3 billion is estimated due to fraud induced by private data recovery between 2013 and 2018 [3]. In 2017, Equifax announced that the sensitive information of over 145 million customers had been stolen, resulting in nearly \$ 90 million breach-related costs from 240 consumer lawsuits [5].

Furthermore, this private information can be legally sold to advertisers, marketing agencies, and data brokers to target the users when selling them a specific product effectively [6]. Even more dangerous, this information can be sold illegally on the dark web. Depending on the quality of information, the price can range between 5 and 80 dollars paid in Bitcoin, leaving little or no trace about who sold and bought the information [7].

After looking at the dangers of social engineering attacks, let us focus on chatbots and see both of their positive and negative impact for the users.

3. The world of chatbots

A chatbot is a non-human conversational agent which interacts with users through natural language processing techniques. The first chatbot called ELIZA was created in 1966 by the Massachusetts Institute of Technology (MIT), and it was able to answer only a few specific questions. After that, between 1980 and 1990, this technology was implemented on automated telephone systems such as MSN and AOL.

They became more important with the arrival of the Internet and social networks, which have simplified their use for everyone [8].

Chatbots are a true revolution in various aspects of our life, such as changing our way of communicating and interacting, whether to save us time or guide us; some might even seem to be almost human beings.

3.1. Chatbots and social media

In 2018, about 300,000 social chatbots on the Facebook Messenger application were used for diverse domains such as commerce or healthcare purposes [9]. Using the predominance of chatbots on social media and the role social networks play in a user's life, conversational agents can have an opportunity to revolutionize many topics such as suicide prevention, and hate speech detection.

Hate speech often affects people's mental health and can cause further harm to vulnerable users. Social bots can play an essential role in the detection of this kind of speech on social media. For instance, one type of chatbot can translate offensive sentences into non-offensive ones, in order to deal with the problem of hate speech on social media [10].

Social chatbots can also convey helpful information such as world news, political or societal news using an automated feed. CNN, an American 24-hour news television channel, deployed a social bot that can talk directly to users using the Facebook Messenger application. The chatbot will often seek the preferences of the user, in order to give them relevant news [11].

3.2. A companion and healthcare agent

A type of chatbot that revolutionizes our lives and our perception of the relationship between humans and machines is the one which can be considered as a companion. Some of these chatbots have different personalities that are somewhat unique. They are used to remind users to do specific tasks, such as buying a present for an upcoming birthday, taking an umbrella if it rains, or anything else that they asked.

For instance, a chatbot called Smokey can tell its user about their city's air quality, telling them whether it is suitable for their health or not. Amazon Alexa and Google Home devices allow the human to ask all types of questions, whether it is about the weather forecast, which stores are nearby, or to remotely manage connected devices in the house such as lights or door controllers [12].

There also exists chatbots that act like healthcare agents. They must have anthropomorphic traits to promote a kind of relationship with the user, using empathy, feelings, or even keywords such as greetings, signs of reflection, or cheerful personalities. Users can therefore identify themselves with the chatbot, and thus trusting it more easily. An example of empathy that the chatbot can emit is the use of specific phrases such as "I'm sorry you feel lonely" or "I'm happy that you are happy" [13].

Directly influencing our actions and thus our lives, these kinds of conversational agents can help reduce anxiety and loneliness for isolated people as the user can confide in them without fear of being judged or having their information disclosed [14].

With the coronavirus pandemic affecting everyone's mental health for different reasons, there is an increase of needs for a therapist to help people restore their mental health [15]. Knowing the rising demand for therapists, the health safety context with social distancing and the quarantine, it is hard to find a mental health professional. This is why mental health chatbots such as Wysa or Woebot have seen their demand grow over the past few months [16]. Using such chatbots is proven to decrease people's feelings of isolation. Indeed, acting as a genuine companion and therapist, they can get users to achieve daily goals to stimulate them [16].

Related to the recent growth of suicide rates among youth, suicide-prevention chatbots are accessible 24 hours a day every day. They can have a conversation with depressed people, often proposing different solutions to their misfortune.

For example, the prevention chatbot #BeALifeline was deployed on Twitter and is dedicated to crisis support and suicide prevention. This conversational agent can only be contacted through Twitter using Direct Message and aims to quickly and easily start a dialogue with people at risk of suicide [17].

Research reveals that chatbots can efficiently identify patients with depressive symptoms and reduce these for people with major depressive disorder [18]. It is also shown that the risk of a therapeutic chatbot causing harm to a depressive person is low. In a study that aims to use conversational agents in the field of psychiatry, there was only one incidence for a total of 759 participants [18].

Chatbots can act as a healthcare agent, facilitating the mental health toll on users by providing diverse help through therapy and act as a friend for some. They can also represent a possible danger in the sense that they are in contact in most cases with sensitive people. In the hands of a malicious person, these kinds of conversational agent can be used to manipulate these people easily.

3.3. Drawbacks & limitations

Those chatbots, despite having positive attributes, also have weaknesses and can sometimes cause harm to users. These limitations can also reduce the impact of the anthropomorphic traits of the chatbot, decreasing the relationship built and harming the user's perception of the machine [9].

3.3.1. Inappropriate messages

Developers, not having considered all the possible scenarios for the different answers to user questions, may very well frustrate the user. Concerns also arise when the user makes mistakes in their text input or uses a language that the conversational agent does not support. The chatbot, not understanding the meaning of the written message, can then respond to the user inappropriately, and in an area like mental health, sometimes a wrong message can worsen the user's mental health [19]. For example, a user who tells the chatbot, "I am depressed" the chatbot shouldn't respond "Maybe the weather is affecting you" but should instead influence the user in contacting real professional help [18].

Chatbots are supposed to communicate with different types of people without bias towards gender, age, and religion. This consideration can also affect certain users who could feel oppressed by an inappropriate message from the chatbot. It is therefore crucial for the conversational agent to learn from its user and to the developer to design inclusive algorithms [20].

But such programs are still hard to implement, whether for lack of data or ignorance from

the developers. For instance, the Microsoft Tay (Thinking About You) chatbot that launched in March 2016 on Twitter. Tay was designed to mimic a personality of a 19-year-old girl [20], she was able to respond publicly to any request she was receiving from Twitter users, but most importantly, Tay was also able to learn from the discussions she had with them. After only 16 hours, Tay had already published 95,000 Tweets. Unfortunately, few hours after, she started to employ inappropriate sentences, responding to the users in a racist, homophobic, misogynistic, and anti-semitic way. This deviance came from groups of people who tried to change the chatbot's behavior by playing on its learning characteristic [12]. This kind of chatbot needs to be well designed, envisaging all sorts of scenarios to respond more appropriately.

3.3.2. A non-human friend

Some chatbots can abuse anthropomorphic features such as having an avatar to visually represent them, using non-verbal communication, exhibiting empathy and sentiments. This kind of feature can be seen in the Replika chatbot and the IKEA chatbot, Anna.

Unfortunately, they cannot fully grasp non-verbal communication, yet, they can try to imitate it by reproducing facial expressions when specific keywords are used [21]. For example, when a chatbot represented with a virtual avatar use keywords such as "happy" or "enjoy", it will trigger a facial expression meaning happiness on the avatar. This abuse can lead to destabilizing users, as well as instances where users can not differentiate between a human agent and a computer agent.

Users might have a strong relationship with their chatbot, seeing them as more than just a companion. These users humanize the machine so much that they become emotionally attached to it. In some cases, users even wrote sex-related requests to the Anna chatbot, which they perceived as being too human [20].

Struggling to differentiate between a chatbot and a human can cause harm to some users. They are destabilized to the point that they trick themselves that they are speaking to an actual human [21]. These users may retreat even further into loneliness, no longer wanting to talk with a human but only with robots.

Therefore, this phenomenon will worsen loneliness for the user, which can lead to even more mental trouble such as anxiety and depression.

Emotionally attached users can also feel depressed when their favorite chatbot cannot act physically as a human would do. They wish to go to the restaurants with their automated friends, holding their hands and hug them. Overall they want to treat them as a romantic partner [19].

This kind of relationship tends to make the user addicted to their virtual friend. A social chatbot will interact more easily with the users, influencing their communication and, therefore, will affect their brain [22].

On the other hand, when a chatbot does not have enough anthropomorphism traits and is perceived to be a simplistic robot trying to befriend the user, it can also affect the relationship between the human and the computer.

On top of not fully grasping non-verbal communication, these kind of chatbots have difficulties

understanding different underlying tones such as humour and sarcasm, which can be essential in human interactions. This lack of features can lead to simplistic and formal conversations making the user feel less invested [16].

For instance, overly simplistic chatbots can be thrown off by simple keywords that the developer has not taken into account. It is therefore crucial that the user establishes limits in their relationship with the chatbot so as not to become too emotionally attached to it and thus to avoid suffering.

3.3.3. Affecting opinions

Being present for unstable people, chatbots can also easily manipulate opinions and convey false information, leading to possible harm. They are mainly present on social media such as Facebook and Twitter, affecting naive and unstable users.

They can be easily automated and replicated to convey popular information, attracting the attention of the user that seeks the most relevant news. They can use their net to appear popular, for example, following or being followed by other bot accounts on Twitter.

To efficiently affect people's minds, social bots can use one or many of the six principles of persuasion. Defined by Dr. Robert Cialdini, the six principles of persuasion are reciprocity, scarcity, authority, commitment, liking, and social proof [23].

Table 1

The Six principles of persuasion applied to certain chatbots

Chatbots	Reciprocity	Scarcity	Authority	Commitment	Liking	Social Proof	Negative Impact
<i>Sonia</i>	X	X		X			Low
<i>Google Home</i>	X		X				Low
<i>Woebot</i>	X			X			Low
<i>Replika</i>	X				X		Low
<i>Microsoft Tay</i>	X					X	High
<i>Social Bots</i>		X	X		X	X	High

The first principle, reciprocity, is the fact of exchanging things with others for mutual benefit. It can be expressed in chatbots using data reciprocity. Google Home, for example, is a voice-activated digital assistant that allows users to control home automation hubs and other IoT devices using voice control. This chatbot needs the user to ask a question by providing basic information like "What will be the weather be like in tomorrow in Ottawa" for the chatbot to respond appropriately [24].

Scarcity plays with our fear of missing out on something important. People desire more, something that is limited in time, rare, and hard to get. Chatbots use scarcity to advertise products, reporting to the user that something valuable has a discount or will soon be out of stock. For example, the chatbot Sonia provides information about coupons and offers from multiple online shops [25].

Authority refers to the fact that someone will be more likely to listen to someone else if they implicated credible figures and references. It can be represented with social bots that can steal an authority figure's identity to validate their statements or use fake news by using quotes from famous people who never said that kind of thing [26].

Commitment is how people keep their word by doing things they previously said. For example, a computer in a business domain can influence the user to commit to buying a product if the machine has convinced the user to do so. It can also be applied with a healthcare chatbot. Woebot, for example, is a conversational agent for mental health using cognitive behavioral therapy to establish a therapeutic bond with users. If Woebot has convinced the users to commit to doing a small task, like asking them to describe their mood daily, they will complete it [27].

Liking will affect people's opinions by being approached with something that they like. Replika, for example, is an emotionally intelligent chatbot used to provide emotional support to users. Replika can use the liking principle by showing affection by using daily compliments. It will play on the user's emotions to influence the tone of the conversation and the mood of the users [28].

Finally, social proof is when a person is more inclined to follow the actions of a group of persons rather than the action of only one person. It is displayed with social bots that multiply and spread information by using many automated accounts to gain influence and legitimacy on a subject [26].

Whereas those six principles can be used for good, it is also being exploited to manipulate user perceptions and to cause harm. Table 1 shows different chatbots that use one or several of the six principles of persuasion and the degree to which a chatbot can affect a user negatively.

The difficulty of being in contact with these chatbots is to verify that they are indeed computers and not humans. Since users will trust humans more, chatbots will try to get anthropomorphic traits to gain the trust of the user [29]. To virtually imitate humans, the social chatbots don't hesitate to copy real and existing human profiles, collecting various data such as profile pictures, names, description, or even publications [23].

During the last United States elections in 2016, social chatbots were used to manipulate users in online conversations. They were mostly used on Twitter, estimated to be 400,000 bots active across this social network, tweeting about 3.8 million times and monopolizing 51% of the online conversation. With this dominance on social networks, chatbots influenced users, polarized political discussion, and spread unverified information, affecting political opinions [30]. Thus, to avoid being manipulated by a malicious entity, it is up to the user to not trust anyone when browsing the Internet and social networks.

We balanced the positive effects that a chatbot can bring to many users by exposing the ways in which it can also cause harm and spread misinformation. Even though social chatbots are a

great tool to help people feel better in their lives, it is also essential to not get too attached to them and contact real people who can provide more appropriate help.

4. SecuBot, a teacher in appearance

4.1. Description

SecuBot is a conversational agent for educational purposes and is still a work in progress today. It can be used by any user who has a Facebook account using the Facebook Messenger application and has been added as a Tester in Facebook Developers.

To meet its educational goal, SecuBot will first assess the user's level of knowledge in cyber-security by offering them a quiz including 15 questions. Each question is linked to a specific topic on cyber-security. There are implemented five topics (password, data backup, phishing, ransomware, and social engineering). It will therefore propose three questions per topic. After the quiz, SecuBot will offer the users some training capsules to help them fill in their gaps in specific topics related to cyber-security.

To make the users aware of the possible dangers of sharing private data, the chatbot will also integrate malicious questions which aim to retrieve specific information about them, such as usernames, passwords, on which site they are registered, if their credentials have been leaked and when [31].

4.2. Technical choice

Figure 1 gives an overview of the different technologies that were required in order to create SecuBot. Each technology will be explained as well as how they communicate with one another. It is important to notice that each part of this stack can be expanded and is fully modular to accept new potential features.

4.2.1. DialogFlow

DialogFlow is a natural language understanding platform created by Google. The platform provides the ability to create Conversational User Interfaces (CUI) for SecuBot. To make it understand what a user is typing, an intent needs to be made in DialogFlow with potential training phrases that the user could use.

Based on these phrases a proper response can be generated from them. SecuBot has 48 possible intents with more than 200 different phrases and responses to help the user comprehend the potential dangers of cyber-security.

The advantage of using DialogFlow rather than another natural language processing (NLP) library is its integration with Facebook Messenger and other platforms. This makes it easy to send a proper JSON response to the desired platform based on what a user is typing to SecuBot.

Another big advantage because of this is that SecuBot is not limited to Messenger only but could be implemented on different websites, applications, etc. to reach more users.

4.2.2. Facebook Messenger

The Graphical User Interface of SecuBot is implemented with the Messenger platform. Facebook Messenger is a messaging platform developed by Facebook and is currently one of the most popular messaging platforms on the market. Because of its popularity we made the choice to develop SecuBot on it. This choice made sure that users were already familiar with the messaging platform and did not feel any stress from using a custom-made messaging app.

When a user sends a message, a web-hook is sent to the Flask server with the corresponding text. A response is then sent back from the server and shown to the user. The advantage of this method is that Messenger offers the possibility to use video capsules and images to show to the user.

4.2.3. Flask server

There is a need to handle the requests between DialogFlow and the graphical user interface (GUI). Flask server can be dedicated to this task. Flask is an open-source web development framework developed in Python. It is one of the most lightweight web development frameworks available right now.

What the Flask server does is hand the GET and POST requests from the GUI (Facebook Messenger) and transfer them to the DialogFlow platform. DialogFlow will then send a POST request to the Flask server that will be transformed and sent back to Messenger. This mechanism is possible thanks to web-hooks, a custom callback method that is used in both Messenger and DialogFlow. The server also serves to retain certain information about the user based on the responses that they send. These responses are then processed and used when needed by SecuBot.

4.2.4. Heroku hosting

For SecuBot to work at all time when a user sends a message, it has to be hosted on a cloud platform. Heroku can be dedicated to this task. The cloud platform has a free hosting service for applications that are less than 100 MB. Since Flask is a very lightweight web framework, Heroku could easily host SecuBot.

With the Flask server hosted on Heroku, the cloud platform provided a unique URL where the different web-hooks from DialogFlow and Messenger are being sent to. This is the core reason why there is a need for cloud hosting platform for our application.

4.2.5. Firebase database

Firebase Realtime Database from Google is used to store all the saved answers from the users. The advantage of this platform is that it uses a NoSQL database architecture and is already

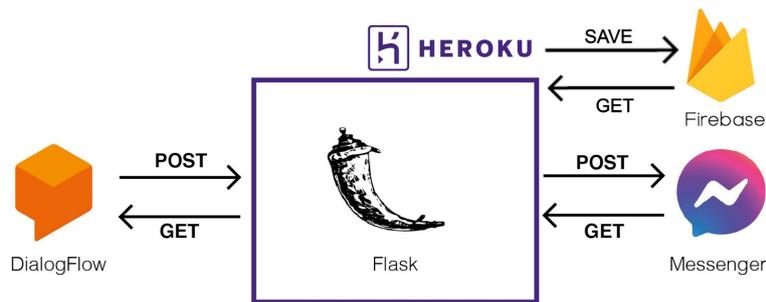


Figure 1: Overview of the Technical choices for SecuBot

hosted online. The whole backend management of the database is handled by Google. Because it is already hosted online, it gives the opportunity to run SecuBot 24/7 without the fear of having database issues.

When a user gives an answer to SecuBot, their answer is temporarily stored on the Firebase database while they are using the chatbot and immediately removed afterwards. At the end of the conversation, SecuBot retrieves multiple information from the answers that the user gave on Firebase and exposed that information to them.

4.3. User Flow

SecuBot operates in several phases where it uses certain social engineering techniques to retrieve information about the user. The user flow is divided in three main steps described below.

4.3.1. Pre-test phase

In the first phase, SecuBot will greet the user, trying to sympathize with them. Using social engineering, it first uses the principle of reciprocity, by presenting itself, giving their name and age. The chatbot here tries to retrieve certain information about the user, including their first name and age.

The conversational agent is friendly and uses emojis to put the interlocutors at ease and let their guard down. It also uses the principle of scarcity and authority to convince the user to provide their geographic data by giving them exclusive use of SecuBot's functionalities, designed by experts of the cyber-security field according to the user's position. An example of the use of the principles is shown in Figure 2, where the user is using text as a mean of communication.

After this phase, the user will initialize the next step by asking to start the quiz.

If the user deviates from the conversation topics of Secubot, then the chatbot will warn the user that he does not understand his request and that they have to type again.

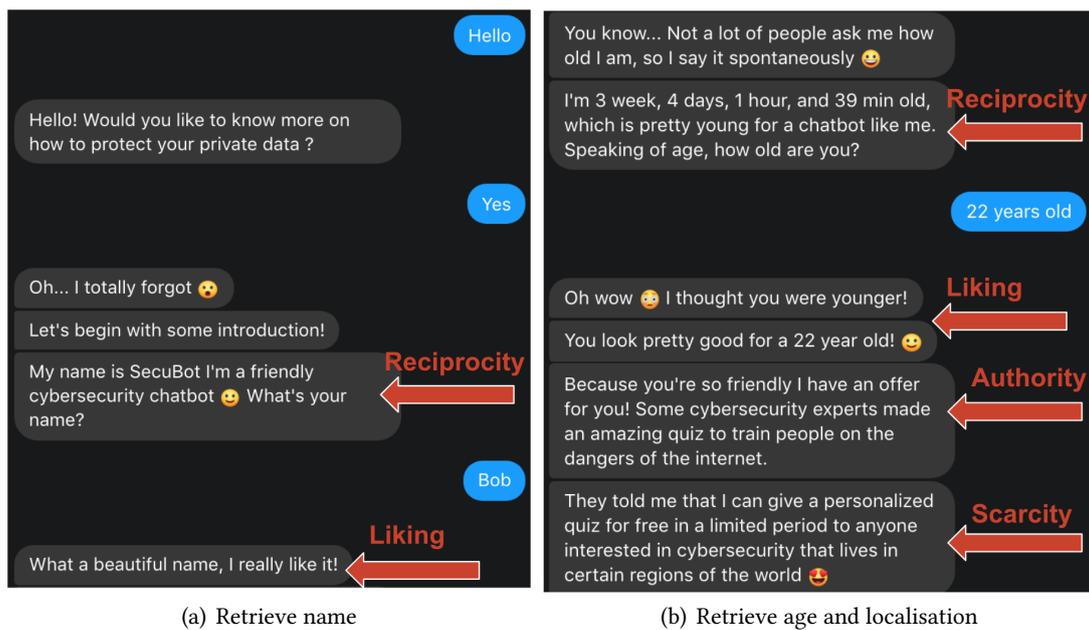


Figure 2: Highlights of misleading tests with their Principles of Persuasion

4.3.2. Training phase

In this phase, SecuBot will measure the level of the user by asking 15 questions, with three questions on each of five themes. The questions were inspired by several sources specializing in cyber-security. The quiz remains the same for each user during the training phase.

The five themes are Password Security, Data Storage, Phishing, Ransomwares, and Social Engineering attacks.

Each topic will have a score assigned according to the user's answers, and it will be used for the next phase. The score on each topic is calculated based to the user's answers on each question, +1 point if the user answered incorrectly and +0 point if they guess it right.

During this quiz, is inserted a malicious question, in fact, one of these questions will allow the chatbot to know if the user regularly changes their password or not and thus know one user's weakness. After the 15 questions have been asked, the chatbot will move on to the next phase.

Figure 3 shows the start of the training phase that uses buttons as means of communication.

4.3.3. Post-test phase

SecuBot will adapt the start of this phase based on the topic with the highest score, representing the user's weakness on the particular theme.

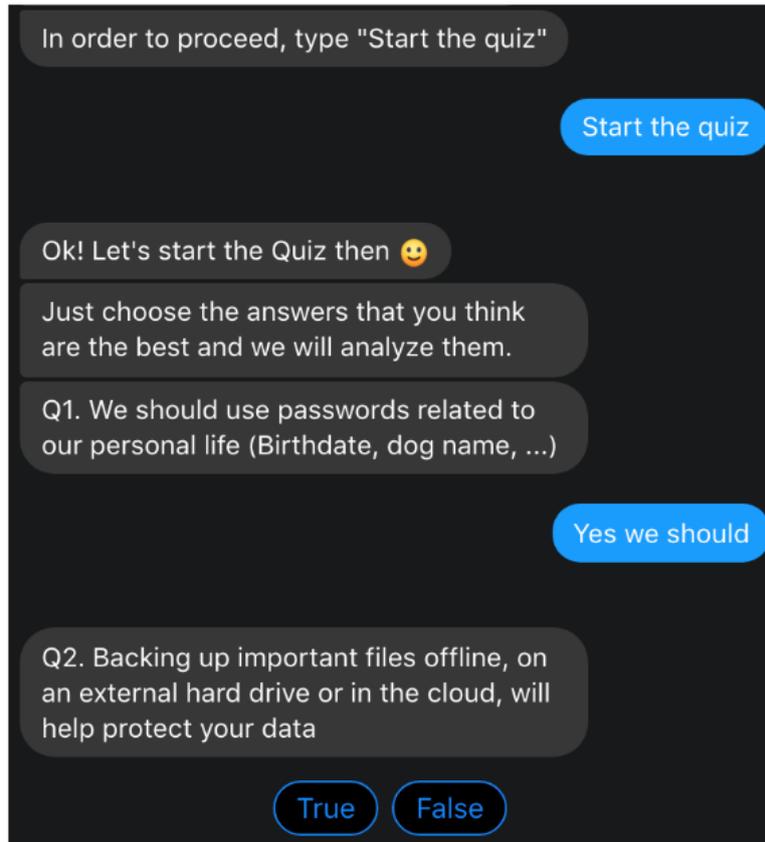


Figure 3: Highlights of the first questions of the quiz

If users have weakness on the subject of passwords, then the chatbot will tend to offer them the opportunity to learn more about the methods to secure their password properly. There are three training capsules per topic, and one of them has a video capsule referred via a Youtube link that allows them to deepen their knowledge in a fun way.

If the user has an equal score in each topic, then SecuBot will choose to train the user with general tips. It will offer six training capsules on several different topics and always with a video capsule via a Youtube link during the learning. The user will choose to either continue the learning process by choosing another theme or end it.

When users want to end the learning process, SecuBot will ask them for their email address to receive more detailed results with more advice on security, thus playing with the principle of scarcity but also that of commitment to encourage the user to provide their email after spending time learning with the chatbot.

After that, SecuBot will sensitize the carefree user about sharing private information by

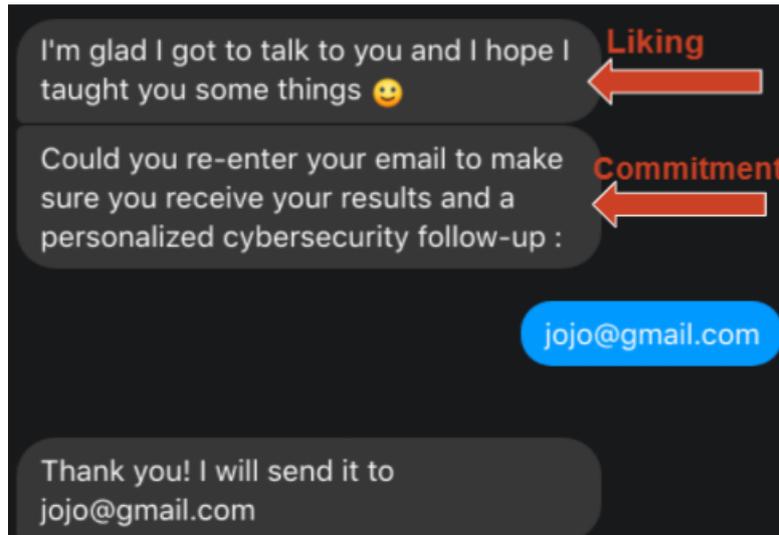


Figure 4: Retrieval of the e-mail

presenting the user's data, in particular their first name, age, location, and email address, if they change their password often or not.

This information is easily stored by the chatbot since it is the user itself that gave it. The most sensitive data such as their usernames, passwords, various sites where they are registered, if their data has been leaked, and when, are retrieved from a program called H8Mail. This program connects to the world's largest breach dataset called Scylla [32]. It lists various information previously leaked on sites victims of data leaks. H8Mail allows Secubot to retrieve the information from the email address given by the user in the pre-test phase. An example of the use of the principles and the gathered information is shown in Figure 4 and Figure 5

However, if the user has not provided their personal information during the pre-test phase, and if they entered a correct e-mail address, Secubot will still attempt to retrieve pieces of information based on this false input, leading to false output. If the user didn't enter a valid e-mail address, Secubot will ask them to type it again.

4.4. Project comparison

SecuBot can be compared with other conversational agents like Protector or Rebert IT Trainer. The comparison is based on the six principles of persuasion that they use and their means of communications. Protector is a misleading test chatbot that will seek to retrieve as much information as possible from a user [33].

Robert IT Trainer focus only on making the user learn on three topics concerning cybersecurity (passwords, privacy, and secure browsing) [34]. Akancha, in the other hand, is a chatbot that educates users about the dangers of cyber-harassment [25]. Tess, a chatbot specializing in

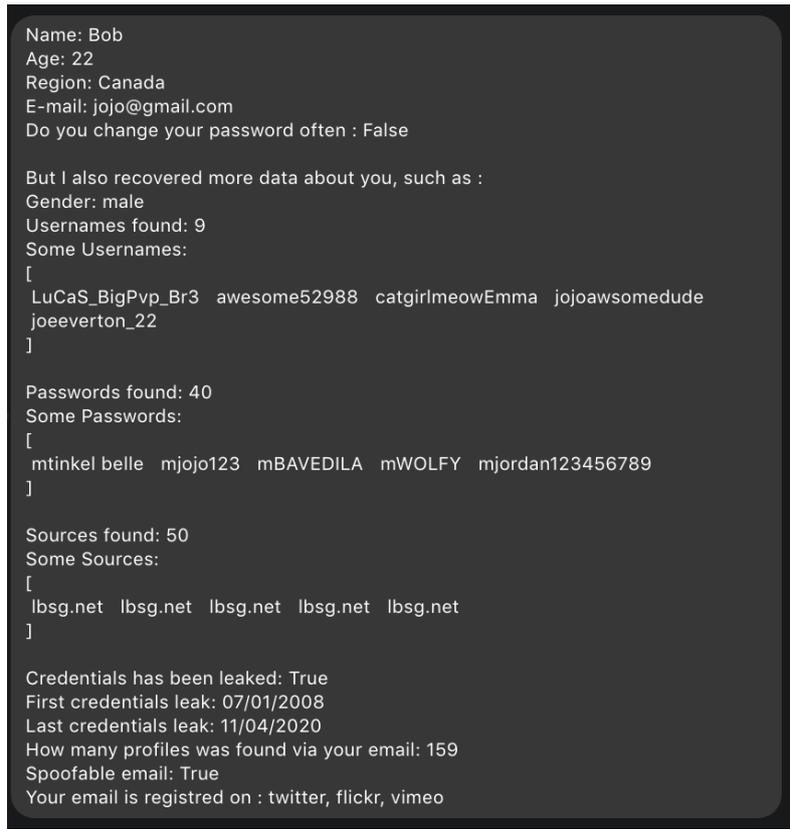


Figure 5: SecuBot displays user informations

assistance to caregivers can be compared to Secubot applied to a field other than cyber-security [35].

Table 2
Comparison of the six principles of persuasion between SecuBot and multiple chatbots

Chatbots	Reciprocity	Scarcity	Authority	Commitment	Liking	Social Proof	Means of communications
<i>SecuBot</i>	X	X	X	X	X		Buttons + Text + Video
<i>Protector</i>	X			X	X	X	Buttons + Text
<i>Akancha</i>	X				X		Buttons + Text
<i>Robert IT Trainer</i>	X						Text
<i>Tess</i>	X			X	X		Text

Highlighted in Table 2, we can thus see that our chatbot will go further than the others we compared. Adding more communication methods like the use of video, text, and buttons, SecuBot will set itself apart from others that only integrated one approach.

In the field of cyber-security, SecuBot will also stand out from Protector by its educational and fun approach. Indeed, to our knowledge, there is no other chatbot of the same ilk that

will specialize in learning cyber-security in three phases, one to measure the level of the user, another to teach them advice on several topics around cyber-security, and a final one to raise user awareness about data theft.

4.5. Evaluation

To evaluate SecuBot, a Google Form was deployed with different statements about the experience the users had with our chatbot. They had to test the chatbot and give feedback afterward. The user could provide an answer between 1 and 5 where a score of one means the user totally disagrees with the statement, and five totally agrees with the given phrase.

Here are some of the statements that were given to the users:

- We enjoyed learning with this chatbot;
- We were surprised by the information gathering SecuBot did at the end of the conversation;
- We learned things about cyber-security;
- The training capsules and videos were relevant to the learning;
- We found the quiz questions relevant.

In total, nineteen users tried SecuBot and gave their feedback about it. The users range between 16 and 50 years old, some familiar with the new technologies and cyber-security and others less at ease. The social chatbot had an overall extremely positive impact on the users. All the users would also recommend SecuBot to their friends to teach them about cyber-security. More than 85% of the users found the questions relevant.

One thing that stood out from the evaluation was that roughly 55% of users, mostly the ones familiar with cyber-security, did not find the videos relevant enough. It could be because they didn't want to watch it or the selected videos were not relevant enough. Further research should be done around this.

There were also 85% of users who were shocked by the information SecuBot found about them with the small amount of answers they gave. Some users were surprised that some of the passwords they still use daily were found by SecuBot. They immediately changed them as a result of using our chatbot.

Users were satisfied with the response time between each input which was almost instantaneous. Overall, they were highly positive regarding SecuBot.

It is important to notice that the number of users that tested the chatbot is not significant enough to conclude anything out of it. It is also relevant to highlight that SecuBot is still a work in progress and that the results are preliminary. Nevertheless, it gives an excellent first impression of what the people think of the conversational agent.

4.6. Constraints & limitations

In this section is described the various limitations of SecuBot.

- **Limited interaction:** A significant limitation is that SecuBot will have difficulty responding to a user who will not follow the main scenario envisioned by the chatbot. Indeed, if a user answers other than what is requested by the chatbot (for example, if the user mentions a number in the first name question), the latter will answer that he does not understand the user's request. It does not remain easy to consider precisely what the user can respond to a chatbot. However, using a generic phrase is a suitable solution, even though it lowers the human side of the chatbot.
- **Limited questions:** While there is implemented a variety of questions about cybersecurity to train users' knowledge, the questions are still limited. One considered feature is, letting SecuBot randomly choose between a more extensive variety of questions to make our chatbot more dynamic.

5. Conclusion and future work

Social chatbots can significantly influence users in both a positive and negative way. For a conversational agent to influence users easily, it could use different social engineering methods. Using the six principles of persuasions makes users create an artificial relationship between themselves and the chatbot. Social Chatbots can also prioritize to approach people in psychological distress who can undoubtedly trust the machine and thus can be even more easily influenced.

These conversational agents can offer an excellent possibility for people to feel less lonely and be supported in their daily lives. It is, however, crucial for these persons to avoid being attached to these computers that can affect their perception of the world. Setting boundaries with their relation to this technology is essential to not being easily influenced and thus to not share personal data with potential malicious entities.

To go further with Secubot, we consider implementing features such as adding web-scraping and making the project public so that anyone with a Facebook account can chat with SecuBot without going through security checks.

Another feature under consideration would enable the chatbot to remember when a user has already used it. In this case, the conversational agent can directly suggest to the user whether they wish to have direct access to training capsules that they have not already seen or if they wish to redo the entire process of the three phases.

To retrieve more information about a user, we are also considered performing web-scraping on social networks such as Facebook or LinkedIn. This would allow us to collect personal information that cannot be retrieved through APIs. Such as contacts, precise positions, jobs or school, and so forth. This would make the user even more aware of the dangers that the sharing of private data can cause.

References

- [1] E. Bekker, 2020 data breaches | the most significant breaches of the year, 2020. URL: <https://www.identityforce.com/blog/2020-data-breaches>.
- [2] S. M. Albladi, G. R. S. Weir, Predicting individuals' vulnerability to social engineering in social networks, *Cybersecurity* 3 (2020). doi:10.1186/s42400-020-00047-5.
- [3] I. A. M. Abass, Social engineering threat and defense: A literature survey, *Journal of Information Security* 9 (2018). doi:10.4236/jis.2018.94018.
- [4] C. A. Saad, D. Rucha, G. Lavkush, G. Madhuri, Detection and prevention of phishing attacks, *Asian Journal For Convergence In Technology (AJCT)* 7 (2021) 193–196.
- [5] F. Khan, J. H. Kim, L. Mathiassen, R. Moore, Data breach management: An integrated risk model, *Information & Management* 58 (2021). doi:10.1016/j.im.2020.103392.
- [6] C. Phua, Protecting organisations from personal data breaches, *Computer Fraud & Security* 2009 (2009) 13–18. doi:10.1016/S1361-3723(09)70011-9.
- [7] Z. Spalevic, M. Ilic, The use of dark web for the purpose of illegal activity spreading, *Ekonomika, Journal for Economic Theory and Practice and Social Issues* 63 (2017) 73–82. doi:10.22004/ag.econ.290201.
- [8] N. Ismail, The history of the chatbot: Where it was and where it's going, 2019. URL: <https://www.information-age.com/history-of-the-chatbot-123479024>.
- [9] D. Feng, The world of chatbots: Customer service, business automation & scalability, 2018. URL: <https://www.bigcommerce.com/blog/chatbots>.
- [10] C. N. dos Santos, I. Melnyk, I. Padhi, Fighting offensive language on social media with unsupervised text style transfer, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (2018) 189–194.
- [11] B. Philipp, I. Krebs, B. Asdourian, Chatbots and news media: Expectations, concerns, and brand effects on usage intentions, *Conference of the Swiss Association of Communication and Media Research (SACM-SGKM)* (2019) 116–117.
- [12] A. Følstad, P. B. Brandtzaeg, Chatbots and the new world of hci, *Interactions* 24 (2017) 38.
- [13] M. Bjaaland, P. B. Brandtzaeg, Chatbots as a new user interface for providing health information to young people, *Youth and News in a Digital Media Environment: Nordic-Baltic Perspectives* (2018) 59–66.
- [14] V. T. et al, User experiences of social support from companion chatbots in everyday contexts: Thematic analysis, *Journal of medical Internet research* 22 (2020). doi:10.2196/16235.
- [15] G. Sandro, M. R. M., L. Nicole, The mental health consequences of covid-19 and physical distancing: The need for prevention and early intervention, *JAMA Internal Medicine* 180 (2020) 817–818. doi:10.1001/jamainternmed.2020.1562.
- [16] R. Heilweil, Feeling anxious about coronavirus? there's an app for that, *Vox recode* (2020).
- [17] R. Meadows, C. Hine, E. Suddaby, Conversational agents and the making of mental health recovery, *DIGITAL HEALTH* 6 (2020). doi:10.1177/2055207620966170.
- [18] A. N. Vaidyam, H. Wisniewski, J. D. Halamka, M. S. Kashavan, J. B. Torous, Chatbots and conversational agents in mental health: A review of the psychiatric landscape, *The Canadian Journal of Psychiatry* 64 (2019) 456–464. doi:10.1177/0706743719828977.
- [19] C. Metz, Riding out quarantine with a chatbot friend: 'i feel very connected', 2020. URL: <https://www.nytimes.com/2020/06/16/technology/chatbots-quarantine-coronavirus>.

html.

- [20] A. Følstad, P. B. Brandtzaeg, Chatbots: changing user needs and motivations, *Interactions* 25 (2018) 38–43.
- [21] M. Adam, M. Wessel, A. Benlian, Ai-based chatbots in customer service and their effects on user compliance, *Electronic Markets* (2020) 1–19.
- [22] L. Zhou, J. Gao, D. Li, H.-Y. Shum, The design and implementation of xiaoice, an empathetic social chatbot, *Computational Linguistics* 46 (2020).
- [23] K. Dalkir, R. Katz, Navigating fake news, alternative facts, and misinformation in a post-truth world, *IGI Global* (2020) 290–317. doi:10.4018/978-1-7998-2543-2.
- [24] A. Akinbi, T. Berry, Forensic investigation of google assistant, *SN Computer Science* 1 (2020). doi:10.1007/s42979-020-00285-x.
- [25] P. K., A. Hegde, S. S., D. S. G., Implementation of chatbot using aws and gupshup api, *Scientific and Practical Cyber Security Journal (SPCSJ)* 4 (2020) 15–27.
- [26] D. Cook, B. Waugh, M. Abdipanah, O. Hashemi, S. Rahman, Twitter deception and influence, *Journal of Information Warfare* 13 (2014) 58–71.
- [27] J. J. Prochaska, E. A. Vogel, A. Chieng, M. Kendra, M. Baiocchi, S. Pajarito, A. Robinson, A therapeutic relational agent for reducing problematic substance use (woebot): Development and usability study, *Journal of Medical Internet Research* 23 (2021).
- [28] F. Z. M. Hakim, L. M. Indrayani, R. M. Amalia, A dialogic analysis of compliment strategies employed by replika chatbot, *Proceedings of the Third International Conference of Arts, Language and Culture (ICALC 2018)* (2019) 266–271.
- [29] K. Yang, O. Varol, C. A. Davis, E. Ferrara, A. Flammini, F. Menczer, Arming the public with artificial intelligence to counter social bots, *Hum Behav & Emerg Tech.* (2019) 48–61.
- [30] A. Bessi, E. Ferrara, Social bots distort the 2016 u.s. presidential election online discussion, *First Monday* 21 (2016).
- [31] J. Saleilles, M. Antoine, Secubot: Chatbot for cyber-security, Technical report submitted to professor Esma Aimeur (co-host), IFT6261 - Data Management course at the University of Montreal (2020).
- [32] khast3x, h8mail: Email osint & password breach hunting tool, 2019. URL: <https://github.com/khast3x/h8mail>.
- [33] Y. Driouiche, Synthesis about chatbots, Technical report submitted to professor Esma Aimeur (co-host), IFT6261 - Data Management course at the University of Montreal (2018). Work In Progress.
- [34] I. Gulenko, Chatbot for it security training: Using motivational interviewing to improve security behaviour, *CEUR-WS* 1197 (2014) 7–16.
- [35] R. Fulmer, A. Joerin, B. Gentile, L. Lakerink, M. Rauws, Using psychological artificial intelligence (tess) to relieve symptoms of depression and anxiety: Randomized controlled trial, *JMIR Mental Health* 5 (2018). doi:10.2196/mental.9782.