# Using the Design of Adversarial Chatbots as a Means to Expose Computer Science Students to the Importance of Ethics and Responsible Design of AI Technologies

Astrid Weiss, Rafael Vrecar, Joanna Zamiechowska, Peter Purgathofer

# Using the Design of Adversarial Chatbots as a Means to Expose Computer Science Students to the Importance of Ethics and Responsible Design of AI Technologies

Astrid Weiss[0000−0001−7803−9413], Rafael Vrecar[0000−0001−6572−5254], Joanna Zamiechowska[0000−0002−5101−8097], and Peter Purgathofer[0000−0001−5453−5631]

Human Computer Interaction Group
Institute of Visual Computing and Human-Centered Technology
Technische Universität Wien, Karlsplatz 13, 1040 Vienna, Austria
http://igw.tuwien.ac.at/hci/
{astrid.weiss,rafael.vrecar,peter.purgathofer}@tuwien.ac.at
e11936038@student.tuwien.ac.at

**Abstract.** This paper presents a reflection on a master class on "Responsible Design of AI" aimed at raising critical thinking among students about the pros and cons of AI technology in everyday life usage on the example of chatbots. In contrast to typical approaches teaching existing policies and design guidelines, we aimed to challenge students by setting up a project on the "most unethical chatbot imaginable". Our teaching concept therefore builds on students' self-identified issues and concerns and develops guidelines for ethical chatbot design according to students' interpretations of the capabilities and potential applications of these technologies. In our teaching we particularly focused on supporting mutual learning between teachers, students, and experts as foundational aspects. We conclude with reflections from the students regarding how this teaching approach can contribute to establish a critical and reflective mindset for future HCI researchers and developers.

**Keywords:** Education · Teaching · Responsible Design · Responsible Innovation · Artificial Intelligence · AI · Chatbots

## 1 Introduction

The topic of "Responsible Design of AI" is addressed from various disciplinary backgrounds. One can reasonably expect to find courses on this topic in a range of departments, including Computer Science (CS), Engineering, Psychology, Sociology, Philosophy, and many others. This interdisciplinary nature challenges teaching it within a Computer Science curriculum, as a given student body is unlikely to have expertise beyond their single, core area. Because of this, we considered "adversarial" chatbots as an opportunity to expose students to a broad range of new views and new ways of thinking about their work. In particular, chatbots can serve as a useful tool for training students with primary

education in computer science to be aware of how related fields, such as sociology, psychology, philosophy and ethics deal with AI-based technology, and how these perspectives can be useful for practitioners in designing, implementing, and evaluating technologies. This broad perspective should help students to become critical and reflective future HCI scholars and technology designers/developers. In this paper, we explain why we consider chatbots as uniquely positioned and well-suited for introducing computer science students to more socially aware perspectives on AI and ethics and why we intentionally decided to challenge students' thoughts about chatbots through asking them to develop a most "unethical chatbot" during the course of the semester. We detail the different assignments we gave students to provoke their thinking and demonstrate the outcome of the course: a set of guidelines on "ethical chatbot development" students derived themselves from the course work. This description should serve fellow researchers as blueprint to develop HCI courses on related topics, such as accessibility and usability. We conclude with the students' personal reflections on the course.

## 2 "Responsible Design of AI" Exemplified by Chatbots

The scientific discipline of computer science is undergoing a transformation. Due to the increasing entanglement of technology with people and our society, the problems and research questions to be investigated are also changing. Subsequently, in computer science education we have to think of means how to equip students with relevant critical and reflective thinking to shape our technological future [2]. The HCI community is aware of the need to include the societal impacts and ethics into CS education. It was even placed into the ACM/IEEE CS Curriculum more than 20 years ago [9]. However, the implementation of such themes is challenging and often detached from other topics. Project-based teaching of Human-Robot Interaction has already proven to be a useful vehicle for exposing technologist students (e.g., in computer science or engineering) to social aspects of technology [10]. Similarly, a board game-based approach on responsible robotics aims to stimulate thinking beyond technology-centered concepts[1]. We aimed to create a similar learning experience under the constraints of virtual distance learning during the COVID-19 pandemic and therefore decided to work with chatbots as a boundary object to teach "responsible design of AI".

### 2.1 Pedagogical Goals

The primary superficial goal of the course is to critically examine potentially disruptive technologies, in this specific case, adversarial chatbots, in theory and practice. A main goal of the course is that students learn the skills necessary to extrapolate possible futures from this exploration and analysis, and to learn to see the difference between the hyperbole usually surrounding new technologies.

[1] https://responsiblerobotics.eu/wp-content/uploads/2019/12/Annex5.pdf

A focus on the social implications of a technology, which most likely all of the students have experienced in some way in their own lives, requires a shift in how they approach, discuss, and work through the challenges and ideas in the project work. Instead of providing concrete truths about what a chatbot is and building small systems with extensive trial and error testing (which would be a more "traditional" CS approach), we wanted to show them techniques for engaging with the technology and its implications. Consequently, we put emphasis on oral discourse and peer-feedback.

The class format was a weekly 90-minute slot in Zoom. It was a small seminar-like course with six participating students. The class followed a theme model where each class had a topic with associated readings or other inputs students should reflect upon. Additionally, a Slack channel was created for asynchronous discussions and inputs and Cryptpad was used to document every session. Each actual class can be broken down into three components: (1) input from the instructors, (2) discussion of the status of the student project, (3) reflection on prepared materials from the previous week. The following inputs and assignments were part of the course.

The first individual assignment was a *movie review*. Students were asked to watch a movie of their choice in which AI and robots play a major role. They were asked to pay close attention to human-robot interaction and how it is depicted in the film: What role does the technology have on society? What are their effects? How do people react (positively/negatively)? Further, they should comment on what they thought were the hard/easy social and technical problems involved with developing human-robot interaction of the sort shown in the movie, including potential ethical issues.

Next, a *reading reflection of two chapters of "Robot Futures"* [5] should serve as the scientific basis and discourse starting point on "Responsible Design of AI" for instructors and students. Students were asked to prepare short written reflections that not simply summarize what the chapters were about, but reflect their own understanding and thoughtful discussion of the material. Some of them focused on specific quotations made in the readings that they considered particularly relevant. Others described the most critical insights or posed the most important questions the readings raised.

Additionally, students were asked to *research on Turing Test [7], Chinese Room [8], and the Mirror Test [3].* For this assignment, we asked them to pair up and research one of these topics. It turned out that the Turing Test was already well-known by the students as it is a prominent example when talking about Artificial Intelligence. However, the other two provided several new insights regarding aspects such as agency, human-likeness, and deception. They deemed to be a valuable starting point to discuss the similarities of human and machine intelligence, as well as crucial aspects in human-agent interaction, such as embodiment, personality simulation, and the Uncanny Valley phenomenon [4].

For one session we *invited a chatbot developer* to share her experience from a developer's perspective. Dr. Barbara Ondrisek, who calls herself "Bot Mother"

and an enthusiastic software developer, presented several of the chatbot projects she was involved in and explained what makes a chatbot "successful" from a business perspective. Her talk also served as a starting point for the students how to set-up their own project work. For example, Dr. Ondrisek explained which ways of "exiting" the conversation have to be considered when letting a chatbot handle customer requests. Moreover, she also stated to always think about a suitable "personality"/"character" for the chatbot to create a positive user experience.

The final individual assignment of the course was to perform an *auto-ethnography of Woebot* (https://woebothealth.com). Woebot is a chatbot developed to serve as a digital therapist. Students were introduced to the concept of auto-ethnography [1] and provided with additional literature on the method. They were asked to install the Woebot chatbot on their phone, try it for several days, and share their experiences in class. For students who did not feel comfortable with the task, we offered a literature research on Woebot instead. Some students shared that they were experienced chatting with the bot as "uncomfortable and weird". This task led to critical discussions about the interaction of "AI and health care" in general and "persuasive technology" specifically. Additionally, some students used the opportunity to find ways to break the conversation flow with the bot and achieved surprising results, which also gave valuable insights for their own student projects.

### 2.2 Student Project

Students were assigned a project with the requirement that they should develop an adversarial chatbot, and over the course of the semester reflect what makes their chatbot unethical. Students were encouraged to develop their own ideas, but received input from the guest lecture on example chatbots and how they were implemented. Students were free to decide where to put the energy in their projects, for example, on using actual machine learning for their chatbot, visual design, or a specific story board. One project was completed by team of two students and four projects were done individually, all achieving valuable results and learning curves.

The primary goal of the project was for students to learn first-hand how many aspects of chatbots, besides "known machine learning aspects", are impacting ethics and responsible design. We believe that this was important for developing respect towards the efforts of responsible technology development and a trained, critical eye for appreciating its benefits and not just the "extra effort" it creates. Students were also encouraged to conduct a minimal user evaluation at the end of the course as part of their project, which helped in creating awareness of research ethics. The primary deliverable for the project was a working demonstration.

## 3 Outcome

The student activities resulted in two significant outcomes: (1) every student participated in conceptualizing and implementing a small chatbot, (2) students

derived a set of guidelines which they deemed to be relevant when designing a chatbot with ethics in mind.

### 3.1 Student Projects

**Inspirational Quotes Bot** Two students paired up to design a chatbot which would support conversations with occasional inspirational quotes found online. This chatbot incorporated machine learning algorithms for the generation of responses based on user input and its own database of sample conversations. The students added a list of inspirational quotes found online to the learning set. As users interacted with this chatbot, its responses were occasionally inappropriate. For example, some inspirational responses could be misinterpreted by a person suffering with anorexia as encouragement to continue losing weight. The two students were also challenged to apply ethical considerations to their user testing protocol. Without revealing the "unethical" aspect of the chatbot, users were encouraged to stop at any point, a wellness-check followup was conducted a week later.

**Fitness Guru** The second project aimed towards building a "nasty" chatbot persona, modelled after a popular sports guru, that would give health advice and motivate people to exercise. The unethical twist was that it made toxic comments to users who entered an unhealthy BMI, instead of, for example, encouraging them to start a healthy diet. The bot was designed to have an offensive and arrogant "personality" with the intent on motivating the user through guilt and "tough love". This project demonstrates how the creator inserts their own values into the design, without considering the needs or cultural differences of the users.

**Phishing Attack Bot** The third project developed around the idea of criminals intentionally using chatbots to gain a victim's trust during a phishing attack. The chosen setup was to open a malicious popup window which shows a quite intrusive warning that the anti-virus software has detected a virus on the victim's computer, and that the "smart" anti-virus customer support chatbot would help resolve the issue. However, this is just a way to gain the users' trust, and eventually to sell the user a "full package" anti-virus solution to solve their issue. While the harm here is intentional and a criminal will not follow these ethical guidelines, this project highlights the innate trust people put into technology without realizing how their data and privacy could be compromised.

**Customer Service Agent** In this project the student decided to custom build a chatbot which would be deployed as a customer service agent on a commercial website to help a user resolve a problem, but would continuously encourage the user to leave a positive review about the product. This promotes the commercial interests of the company, over those of the user, in an annoying and unhelpful way. The intent of the chatbot is also deceitful and opaque, where the user

believes its purpose is to solve a problem, the underlying motive is to increase the product rating. To eventually redirect the user to the review link, the chatbot conversation is programmed to follow several pre-defined paths.

**Building Rapport** The final project implemented a social chatbot focused on gaining the trust of the user by convincing them that they are speaking with a real human. The chatbot attempts to emulate natural human conversation by strategically inserting humor, jokes, slang, and misspelling into the conversation flow. This can then be used towards various unethical ends, such as phishing, or spreading misinformation on social media. This is an ambitious task to achieve convincingly, and is a major focus of research and development in artificial intelligence. The student gained experience in working with Google DialogFlow and Google Assistant technologies to create a small working prototype. The experiences were mixed as it turned out that designing a chatbot despite using existing machine learning libraries is more complex than anticipated beforehand.

### 3.2 Guidelines on Ethical Chatbots

The last classroom activity was a retrospective look at the results from the projects, other assignments, and the discussions. In this session the students together with the instructors derived the following guidelines on ethical chatbot design:

- Be transparent: Openly show that your chatbot is a chatbot.
- Use a label or explain yourself: Be open about your intentions.
- Do not try emulating human behaviour: E.g., avoid human names, emotions.
- Avoid assigning gender to your chatbot.
- Provide a way to chat with a human and offer a way to end the conversation.
- Make an effort to detect urgency and relay the chat to a human.
- Be aware of vulnerable users/groups.
- Be aware of people with cognitive limitations/disabilities.
- Always question the values you encode into the chatbot when offering options, try to be comprehensive; offer a "none of the above" option.
- Aspire to be sensitive of context, e.g., culture; do not assume everyone celebrates on December 24.
- Inform users about how you handle their data and input, as is mandated by GDPR, and give people a chance to opt out.
- Assume the worst about your chatbot, and design for it.

Comparing these guidelines with the "Social and Ethical Considerations" of Conversational AI (CAs) [6] offers initial insights into potential successes of the course and future areas for improvement.

The authors of these Social and Ethical Considerations explain the importance of trust and transparency, and how it enables users to make informed choices in their interaction with a CA. They suggest making the CAs status as a non-human, and its motivations and capabilities explicit, which goes in line

with the students' guidelines. In this paper, it is further explained that users assume a CA is neutral and unbiased, and that their data is secure, when this is often not the case. They describe the topic of user privacy as "paramount", and increasingly important to address on a societal and legislative level, due to the encroachment of these technologies into more aspects of our lives. Users are mostly unaware of the amount and scope of data that is collected on them, and how this data is used. In the case of CAs, the authors are in favor of legislative and legal compliance with respect to data privacy protection.

The authors are also aligned with the students' guideline of designing CAs to be androgynous. Not only does this avoid reinforcing gender stereotypes that are purposefully or unintentionally programmed into the CA, it also avoids influencing the user's interaction with the agent. Dehumanizing and not anthropomorphizing the chatbot also helps avoid a user subconsciously placing undue trust into it. Finally they discuss the dangers of unsupervised learning, where the CA could learn profanity, abusive language, or personal data from users, and then incorporate this into future conversations with different users. The importance of controlling the learning data and the chatbot responses was demonstrated in the "inspirational quote" chatbot.

The comparison of state-of-the-art "Social and Ethical Considerations" of Conversational AI (CAs) [6] with the guidelines derived by the students demonstrate that the course managed to convey the relevance that a developer or designer must always consider the unintended ethical consequences of their work in order to mitigate against possible harm. As we have seen from the student projects, even well-meaning chatbots can become problematic. As the projects progressed, the ethical considerations became increasingly nuanced and complex. Emerging themes of transparency, privacy, users' rights, protecting people, and the larger social impact of technology became central points of discussion. As such, the principles of digital humanism, and of putting the well-being of the person at the center of design considerations, underscore the guidelines constructed over the course.

## 4 Student Feedback

One of our pedagogical goals for the course was to show students reflective techniques for engaging with technology and its implications; through the provocative task of thinking about unethical chatbots for their student project we aimed to foster reflective stances. As a voluntarily final submission, we asked them to hand in a written reflection on their learning experience. Quotes such as the following suggest that we succeeded "*I honestly can say that the lecture substantially changed my view on chatbots. First and foremost because I did not see how many negative implications can be caused by malfunctioning chatbots as people probably can be encouraged by the bot to do things to harm themselves or others.*"

Based largely on a constructivist learning theory, we believe that what you learn is to a great extent determined by the diverse and holistic ways you are enabled to think about a subject matter. Therefore, our meta-goal was that

students learn skills that are necessary to extrapolate possible futures from the chatbot example. Again, a student feedback suggested that this was indeed the case: "*Furthermore, I underestimated the ethical implications that come with designing these. For me, chatbots before were just small gadgets which regularly totally mess up and annoy me when changing details on my, e. g., phone contract as they are often used to replace human assistance in my experiences [...] the lecture was insightful from many perspectives and also thinking about malicious use cases broadened my horizon in a way that we always have to think twice [...] when elaborating if our intended design can have negative implications [...]*" Another student stated that the course was "adventurous": "*We didn't have such strict tasks with rigid deadlines like in other courses but together we explored unethical chat bots, which I find really adventurous. [...] Frankly I never had in my technological education any focus on ethical aspect or social consequences of what I as an engineer create. Very big advantage of this course was opportunity to train creative thinking, as we weren't just ordered to perform very concrete strict tasks but had an opportunity to think about possible usage of our chat bot*".

## 5 Conclusion

This paper illustrates why teaching "Responsible Design of AI" through the means of "adversarial" chatbots can be a useful mechanism for exposing CS students to a broader, socially-embedded view on technology. Students are exposed to readings and discussions that illustrate why it is important to be aware of social considerations, as such competence can influence algorithm and technology design and dictate how a technology is used, integrated, and ultimately, decide on whether it is successful or not. These are perspectives generally not covered in a technology-centered curriculum.

The brief overview of topics covered and the format of the course (including a student project) gives an indication how to engage students to reflect on the social implications of technology. Scaffolded by our input and the projects we gave them, students used the class as an opportunity to explore a range of social perspectives on technology. Furthermore, they designed chatbots and conducted evaluations, mostly using qualitative approaches. Overall, we believe that this paper illustrates the potential of adversarial chatbots as an accessible topic for providing CS students with education on how to be more socially reflective regarding their work.

## Acknowledgements

We would like to thank all students who took part in the course described in this paper and who put so much effort into engaging with the material and developing exciting projects despite all COVID-19 limitations.

## References


1. Ellis, C., Adams, T.E., Bochner, A.P.: Autoethnography: an overview. Historical social research/Historische sozialforschung pp. 273–290 (2011)
2. Frauenberger, C., Purgathofer, P.: Ways of thinking in informatics. Commun. ACM **62**(7), 58–64 (Jun 2019). https://doi.org/10.1145/3329674, https://doi.org/10.1145/3329674
3. Haikonen, P.O.: Reflections of consciousness: The mirror test. In: AAAI Fall Symposium: AI and Consciousness. pp. 67–71 (2007)
4. Mori, M., MacDorman, K.F., Kageki, N.: The uncanny valley [from the field]. IEEE Robotics & Automation Magazine **19**(2), 98–100 (2012)
5. Nourbakhsh, I.R.: Robot futures. Mit Press (2013)
6. Ruane, E., Birhane, A., Ventresque, A.: Conversational ai: Social and ethical considerations. In: AICS. pp. 104–115 (2019)
7. Saygin, A.P., Cicekli, I., Akman, V.: Turing test: 50 years later. Minds and machines **10**(4), 463–518 (2000)
8. Searle, J.: Chinese room argument, the. Encyclopedia of cognitive science (2006)
9. Tucker, A.B.: Computing curricula 1991. Commun. ACM **34**(6), 68–84 (Jun 1991). https://doi.org/10.1145/103701.103710, https://doi.org/10.1145/103701.103710
10. Young, J.E.: An hri graduate course for exposing technologists to the importance of considering social aspects of technology. J. Hum.-Robot Interact. **6**(2), 27–47 (Sep 2017). https://doi.org/10.5898/JHRI.6.2.Young, https://doi.org/10.5898/JHRI.6.2.Young