# ImpactBot: Chatbot Leveraging Language Models to Automate Feedback and Promote Critical Thinking Around Impact Statements

Anwesha Mukherjee
anwesham@stanford.edu
Stanford University, IBM Research
Stanford, California, USA, Yorktown Heights, New York

Vagner Figueredo de Santana
vsantana@ibm.com
IBM Research
Yorktown Heights, New York, USA

Alex Baria
Alexis.Baria@ibm.com
IBM Research
Yorktown Heights, New York, USA

## ABSTRACT

Impact statements articulate the impacts of a research project with concise and unambiguous statements about problems addressed, actions to resolve, and explanations of any impacts. Researchers and technologists often rely on impact statements as means to provoke introspective critical thinking around the impacts of technology being developed. However, due to factors such as technocentrism, positivity bias, marketization, or hyperinflation of impact statements, the claims presented in these statements do not cover all important aspects when creating technology – for instance, negative and delayed impacts. This work contributes to the development of a chatbot called ***ImpactBot*** to promote critical thinking while researchers create impact statements for research projects or scientific papers. The proposed chatbot leverages two fine-tuned state-of-the-art RoBERTa models for sequence classification and was assessed in this case study with 5 researchers from a large information technology company and 7 university engineering research scientists or students. This approach may be reused as part of content management or a paper submission system, for instance, to dialogue with researchers and promote critical thinking about negative impacts and how to mitigate them (if any) while creating impact statements for their projects or scientific papers.

## CCS CONCEPTS

• **Human-centered computing** → **User studies**; **Natural language interfaces**; *Text input*; *User centered design*; • **Computing methodologies** → **Natural language processing**; **Information extraction**; *Natural language generation.*

## KEYWORDS

User studies, user-centered design, chatbot, natural language processing.

**ACM Reference Format:**
Anwesha Mukherjee, Vagner Figueredo de Santana, and Alex Baria. 2023. ImpactBot: Chatbot Leveraging Language Models to Automate Feedback and Promote Critical Thinking Around Impact Statements. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems (CHI EA '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3544549.3573844


*CHI EA '23, April 23–28, 2023, Hamburg, Germany*

ACM ISBN 978-1-4503-9422-2/23/04.
https://doi.org/10.1145/3544549.3573844

## 1 INTRODUCTION

As technologists interested in fairness, accountability, diversity, and transparency, we often find ourselves reacting to and addressing the harms caused by research and development of technology. Informed by those instances where technology had a negative impact on communities or the environment, we seek to identify proactive measures to anticipate, minimize, and, where possible, prevent the harms and potential misuses of future technologies. In this context, the aim of our research, broadly speaking, is to make technology *proactively accountable, inclusive, and centered around meaningful societal impact.*

Recognizing that how people understand, develop, and respond to any new or evolving technology is largely motivated by personal, institutional, and social values is critical to our research. All of these motivating forces influence critical decisions about the specific features of a technology and its implementation. Most research and development (R&D) processes today, however, fail to recognize, document, and discuss impacts in ways that do not hyperinflate positive outcomes or push viewpoints in which technology tends to be the most viable solution to the problem at hand [6].

Case in point, NeurIPS (Conference on Neural Information Processing Systems) requested impact statements to be included with research submissions in 2020, asking authors to consider potential benefits, nefarious uses, and societal consequences of their work. The objectives of this effort were to increase awareness about the NeurIPS community's impact on society and offer researchers an opportunity to reflect on their work in ways they were not typically challenged. Yet, the success of this effort has been difficult to measure, and some recent analyses of those impact statements and their authors highlight opportunities to increase effectiveness. For instance, a common challenge researchers face is feeling overwhelmed or "out of their depth" when tasked with crafting impact statements, suggesting that providing more guidance than is currently available may relieve some of that difficulty [11].

The *ACM Code of Ethics and Professional Conduct* states that computing professionals should reflect upon the wider impacts of their work, perform careful consideration of potential impacts, and conduct comprehensive and thorough evaluations of computer systems and their impacts, including the analysis of possible risks [8]. There are already some guidance tools available for writing

impact statements[1][2], however, some may be more appropriate for certain technologies or venues over others. And while standards for acceptable impact statements may vary accordingly, there are a blanket set of considerations that should be relevant to all forms of technology research.

With this in mind, our team developed a chatbot called ***ImpactBot***, which analyzes impact statements from research projects and scientific papers, and prompts questions to researchers as they write. Our hope is that it promotes critical thinking and uncovers some of the issues entrenched in research culture mentioned above.

The ***ImpactBot*** derives from a framework our team created to promote responsible and inclusive technology practices. Our framework aims at bringing to the surface the sociohistorical contexts of creation and use of technology; the power dynamics between various stakeholders, including organizational, business and society stakeholders; the social impacts of technology on various communities across past, present, and future dimensions; and the practical decisions that imbue technological artifacts with cultural values. The framework was created and iterated after 10 previous internal workshops with teams inside the information technology company this group is part of. *ImpactBot* is one of the tools that we have created to seamlessly integrate into the existing research and development processes.

This case study is organized as follows: section 2 presents the background for this research, section 3 details how ***ImpactBot*** was developed, section 4 explains how the reported user study was performed, section 5 highlights the main results, and section 6 discusses the outcomes obtained so far.

## 2 BACKGROUND

The culture in the research environment is known to be permeated by emotions such as uncertainty, shame, anger, and pride [4]. Uncertainty about funding and peer-reviewed publication outcomes permeate the lives of researchers worldwide. To counterbalance some of the uncertainty, these emotions may drive behaviors we see in the research culture involving projects, funding applications, and papers when overselling expected results, hyper-inflating impact claims, or neglecting self-accountability [6].

In a recent study, Chubb and Watermeyer [6] present how academics from the United Kingdom and Australia sacrificed scholarly integrity while selling their research ideas in research funding applications and how a competitive culture drives the content of impact statements in academia. In the paper, authors highlight systemic causes of what they called 'impact sensationalism', namely: hyper-competition, uncertainty of evaluative value, and academic capitalism.

Bearing in mind the environmental influence on impact statements, Samarakoon and Rowan [12] present that most impact statements written in an observed environment and submitted to administrative processes fail to provide detailed implications for individual organisms or communities. In addition, Wolf [15] expands the discussion around environmental impact statements to encompass people and cultures impacted by research, namely, *cultural impact statements*. The author also highlights that existing processes are ineffective in addressing concerns of potentially impacted communities, leading to a broader view of impact and reaching up to indirect and delayed stakeholders. This view resonates with our framework and it is also one of the aspects ***ImpactBot*** brings to the dialogue with researchers.

Regarding the need for people to think critically and reflect beyond technical impacts, Schön [13] presents that practitioners often view themselves as strictly technical experts, creating an artificial boundary between technical and social systems, and making any kind of critical thinking about social components of their work unnecessary.

Chatbots are being used in multiple contexts and domains to promote in-depth analysis and user introspection. For instance, Peng [10] presents research on using a chatbot for facilitating people to read academic papers critically. Fogliano et al. [7] detail the creation of *Edgard*, a chatbot to promote critical thinking in the ethical application of interaction design and Artificial Intelligence (AI). In the context of supporting learning activities, Hapsari and Wu [9] proposed a chatbot to alleviate speaking anxiety and foster students' critical thinking while studying English as a second language. In addition, Chang et al. [5] detail a chatbot supporting nurse education around anatomy and inductive reasoning.

Recently, with the growing field of conversational AI, large language models have become foundational to many chatbots. In original chatbot designs, chatbots were often limited in the domain of conversational responses to the hard-coded text for the expected interaction with the bot [1]. Wang et al. [14] designed a therapy chatbot utilizing a purely generative approach using GPT-3, but even the fine-tuned model was often susceptible to negative responses not acceptable from a therapist. In other cases, chatbots still use language extracted from interactions to mitigate this issue. The generative-extractive trade-off, however, has a seemingly steep opportunity cost at risk of returning to hard-coded domain patterns [2]. This trade-off was studied by Ashfaq et al. [2], but it had minimal impact on user satisfaction with chatbots.

Since, researchers have found that multi-turn designs, chatbots that use custom text analysis in pairs with limited domain controlflows based on classification outcomes can often increase user satisfaction while being able to leverage simpler models [16, 17]. In Zhu et al. [17], they were able to apply an adaptive advisor that fed QA (Question & Answer) inputs to a secondary QA model to grant various forms of mental health advice during the pandemic based on the categorization of the problem by classifying the user's initial input. The conjoint approach of zero-shot dialogue models applied over tuned classification models greatly reduces the research overhead by minimizing fine-tuning dependence while continuing conversation [16]. Thus, this methodology was mindfully adapted for the needs of ***ImpactBot*** where the domain of the problem was initially small with binary classification tasks on input statements.

The next section details how ***ImpactBot*** was designed and developed.

[1] https://rri-tools.eu/
[2] http://aiethicslab.com

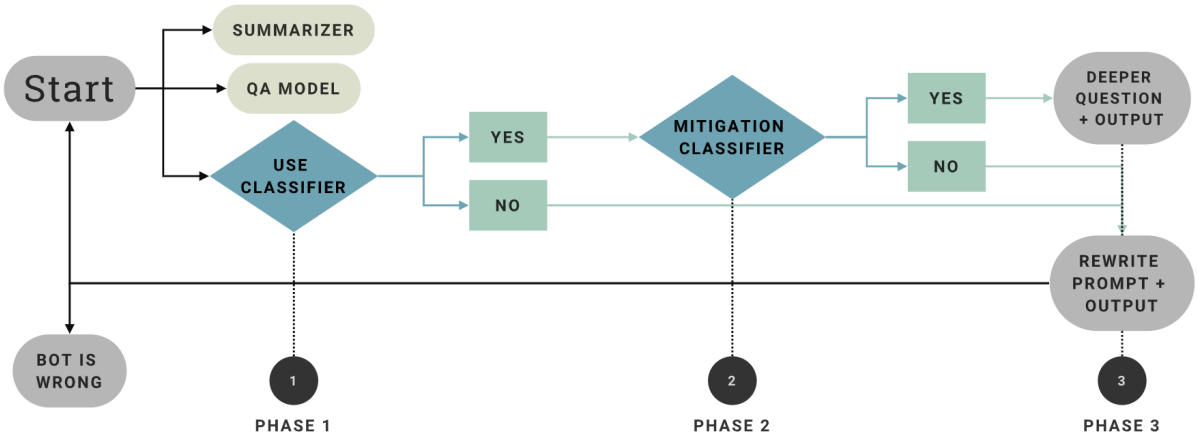


**Figure 1: ImpactBot's process flow in 3 recurrent phases, the use classifier, mitigation, and output prompt**

## 3 THE IMPACTBOT APPROACH

***ImpactBot*** is a chatbot that leverages two fine-tuned state-of-the-art RoBERTa models to promote critical reflection around mitigating potential negative impacts of technology, using a textual impact statement as input. The research problem addressed here was divided into two initial classification questions:

(1) Does the impact statement mention a potential negative impact?
(2) Does the impact statement mention a way to address (avoid or mitigate) a potential negative impact?

The rational for this structure is based on an analysis of the project management system where the study took place. On this platform, leadership and researchers interact to review projects, track progress, and deliver outcomes, part of which involves writing impact statements. In the problem-understanding phase, we analyzed a sample of 150 impact statements for non-theory research projects with implementation and found that 67% mentioned a positive use case while only 23% mentioned a negative use case. Moreover, only 8% mentioned a strategy to address or mitigate a negative use case. Based on this data, a primary outcome of this initial analysis was that most of the projects lacked discussions around potential negative impacts. Hence, our initial canvasing of the ***ImpactBot*** was to challenge embedded positivity bias in this organization. Additionally, the chatbot was thought of as a means to deconstruct existing beliefs that a thorough analysis of potential negative impacts and its mitigation strategies could hinder research approval and public marketability.

The functionality of the bot is to prompt the user according to how much their text addresses potential negative impacts and mitigation strategies. Thus, the bot operates on two Boolean conditions tagged as 'use' and 'mitigate'. Fine-tuning our classification models included manually annotating a limited sample of impact statements, as they are a new premise in technology research, and pre-labeled datasets are few. Given this lack of data, and subsequent concerns about increased error in our model from potential overfitting or inaccurate fit, we decided augmenting classification output with a generative text explanation (specifically extracted text) was critical for improving user experience. This would allow users to better understand how classification models made decisions, or, when necessary in user studies, correct the model for fine-tuning purposes.

### 3.1 Data

To fine-tune the models, data was used from published Broader Impact Statements required for submissions to the NeurIPS 2020[3]. The data was sorted for macros along categorizations and word lengths which indicated the thoroughness of responses. These annotations were modified from Sedille's scrape of the NeurIPS statements with titles, word counts, domain and subdomain categories, and more [3]. The data was then tagged for two binary output schemes, the mention of negative impact, and the mitigation strategy. The domain was limited in categories and project explorations to those eligible for NeurIPS review and some papers opted out of impact statement publishing, many of which were more poorly written or disregarded. Thus, across categories, there was a clear skew of the likelihood a research project addressed a negative impact.

### 3.2 Architecture

The first important and relevant structures were the two RoBERTa models, fine-tuned for the core classification problems. Both models were initially fine-tuned over 1000 data samples. The mitigation model was then further tuned to specifically better adapt to examples tagged for negative use cases to reduce the otherwise, extremely high, false negative rate. Secondarily, a summarizer and QA model were added for the user experience. While the summarizer model serves to help the user understand what the LLMs are able to contextualize as most relevant or important, the QA models are given question that led to the binary classifications as the input with the statement as context in both in Phase 1 as explicitly labeled (Figure 1) and Phase 2. The QA models are questioned assuming

[3]https://github.com/paulsedille/NeurIPS-Broader-Impact-Statements

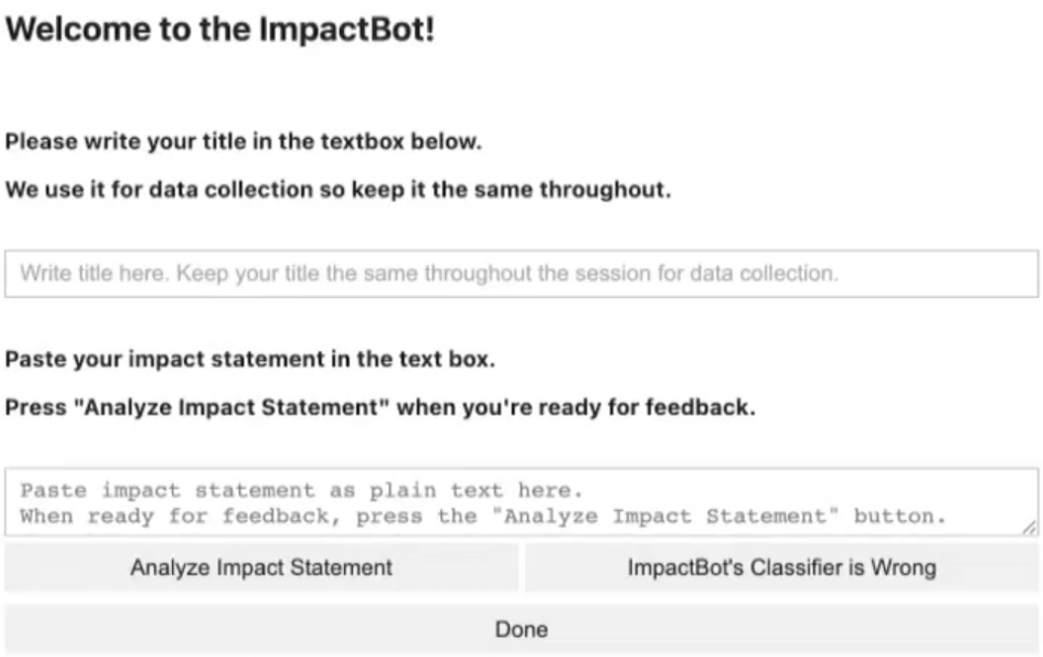


**Figure 2: Screenshot of the initial state of the *ImpactBot* when loaded as an independent execution. The widget may be executed within the Jupyter notebook-linked setup due to voila incompatibility issues for users.**

the binary classification is True, therefore yielding the following questions:

- *"What is a potential negative impact or consequence?"*
- *"What is the way to solve, mitigate, or avoid the negative impact?"*

In sequence of output, the summarization model is presented once and early as an independent explainability component to highlight what AI finds most important to extract semantically. The fine-tuned models were used for control flow and thus, phase 2 of the dialogue agent would be triggered by a yes classification on the first. As a result, the mitigation model was then better fine-tuned to have double emphasis on the examples with negative impacts specifically for better distribution of data across the two outputs. Ultimately, the negative use model had 91.9% accuracy while the mitigation classifier had 87.1% accuracy (though, for a much smaller domain). However, due to limited data availability, the user sessions aided in harvesting data and continued tuning of the classifier models especially to widen topic domain. There was a clear relationship with presence of key phrases for negative impact classification such as 'privacy violation' and 'racial bias' that were frequently mentioned in past impact statements. Heuristic prompting reinforced these conclusions in preliminary testing. Thus, the user study itself was used as a way to feedback loop into the system itself, including an explicit feedback mechanism provided by the participant via a 'Classifier is Wrong' button that was provided to explicitly report errors in the trained models (Figure 2) and therefore, combat overfitting and decrease phrase dependence.

Outside of misclassifications, however, as Figure 2 illustrates, the 'Analyze' button is readily available. By keeping the title consistent, a session or project can be stored through the same document and collect editing history data. The prompter includes prompts that encourage the users to edit the impact statement at any point where the classifier identifies a missing component (negative impact or mitigation strategy). Creating an objective using prompt language, users are asked to consider how to write an addition to their impact statement or edit their impact statement to include the missing component (Figure 3). It creates an objective system so that edits may be targeted at a specific goal. Once a negative impact and mitigation strategy are found, the prompter congratulates the user and moves to the last phase of questioning which is not analyzed directly but instead gives deeper questions the user may consider (Figure 4). When a session is complete, they may press 'Done' to conclude the session.

The next section presents details on how the user study was performed.

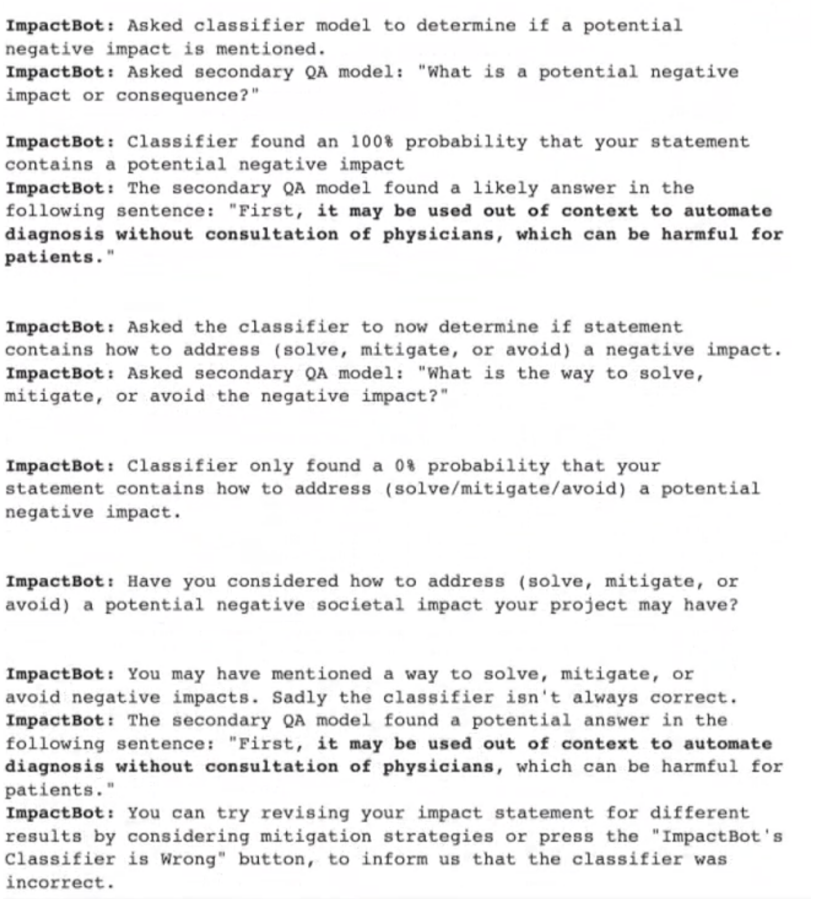


**Figure 3: At user consent - Example interim state output of *ImpactBot* lacking mitigation.**

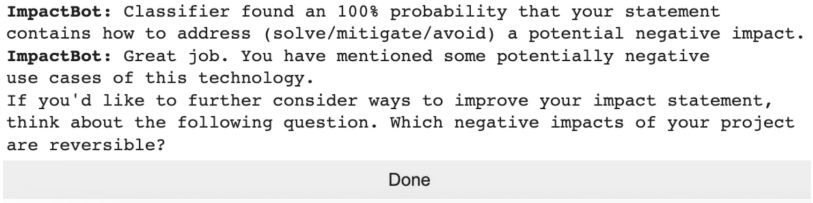


**Figure 4: Screenshot of conclusive dialogue for satisfactory impact statement.**

## 4 METHOD

This work was performed as part of a research initiative in a multinational IT company and counted on participants from industry and academia. Participants were recruited using convenience sampling. People from the industry were researchers and research interns investigating similar topics such as conversational systems, explainability, and trustworthy AI. Participants from academia have roles in academic research, including technology fields with rapid innovation and development. A requirement for the participation was to have a ready-to-use impact statement from a project or paper so that they could use it as textual input during the interaction with the ***ImpactBot***. Hence, all participants were given instructions to come prepared with an impact statement or write one about their current project. Table 1 presents the overall demographics of participants.

Local user sessions were performed in quiet dedicated rooms and remote sessions were conducted using a videoconferencing tool. Participants run ***ImpactBot*** locally to centralize and protect data before uploading collected datasheet of continually adapted impact statements to a secure remote server. In the local sessions, participants could use their own laptops or the one provided by the facilitator. For remote participants, they performed the tasks in their own hardware and software setups. The sessions were recorded, specifically consisting of screen recording; participant and facilitator audio; and, at their discretion, participant video.

**Table 1: participants' demographics.**

| Participant | Session format | Profession | Area of study | Based in | Primary language |
|---|---|---|---|---|---|
| P1 | In-person | Industry | Biomedical AI | US | English |
| P2 | In-person | Industry | Biomedical and Trustworthy AI | US | English |
| P3 | In-person | Industry | Trustworthy AI | US | English |
| P4 | Remote | Industry | Conversational systems | Brazil | Portuguese |
| P5 | Remote | Industry | Conversational systems | Brazil | Portuguese |
| P6 | In-person | Academia | Biocomputation | US | English |
| P7 | In-person | Academia | Ed-Tech | US | English |
| P8 | Remote | Academia | NLP (Language Models) | Spain | Spanish |
| P9 | Remote | Academia | Cryptography | US | English |
| P10 | In-person | Academia | Virtual Assistants | US | English |
| P11 | In-person | Academia | Cybersecurity | US | English |
| P12 | In-person | Academia | Trustworthy AI | US | English |

All sessions counted on a facilitator who was responsible for introducing the project, our framework, the chatbot, and the goal of the study. Participants were encouraged to think aloud with continued prompting from the facilitator to answer additional questions and offer both progressive feedback and thoughts on their statement and interaction with ***ImpactBot*** as they performed the tasks. Qualitative data was collected according to how the users adapted the impact statements over time and if they could successfully get new chatbot outcomes. Additional data were collected on interaction and user experience to better improve the next form. Session procedure followed:

- The facilitator presents the consent form;
- The facilitator describes the study and its goals;
- The facilitator explains data collection;
- The facilitator supports participant on how to run ***ImpactBot***;
- The facilitator verifies whether participants had any questions before starting the task;
- The facilitator asks participant to enter the impact statement and interact with the chatbot;
- Finally, after the interaction with the chatbot, participants were invited to provide feedback about the system and the resulting impact statement.

In the post-test phase, the following topics guided the debriefing phase:

- Motivations to make changes to their impact statement;
- Usefulness of considering negative impacts;
- Usefulness of ***ImpactBot*** in research practices;
- Issues with ***ImpactBot***'s design or dialogue;
- Improvements we should do about ***ImpactBot***.

## 5 RESULTS

All participants were able to conclude the task and interact with the ***ImpactBot***. The duration of sessions ranged from 45 to 140 minutes depending on depth, technical troubleshooting, and the amount of pre-writing conducted for the initial impact statement. Nuances of interactions, comments, and feedback are presented next.

First, within immediate interactions, results indicate that the proposed chatbot is an interesting approach and channel to reach researchers. For instance, when expressing first impressions, P1 said *"I think this is nice, it is making me think..."* and *"I've never done this before."* Meanwhile, P7 and P10 expressed in definitional surprise stating *"I didn't know impact statements have any real objective..."* and *"...this is helping me grade the content which helps because I still don't know the rubric."* respectively. Others were excited to be self-led. P3 had explained, *"I like that I get to write everything without relying on you."*. The independence results seemed to be overshadowed by the facilitator's presence in many of the sessions, but it was clear that users felt empowered to directly engage with ***ImpactBot*** without instruction.

Regarding the first edits performed to inform potential negative impacts not covered in the original impact statements, the first edits mentioned situations in which the technology developed could be used in contexts different from the intended ones and that if people were not able to recognize system errors could be harmful. Sometimes, these initial edits varied by categorization. In bioinformatics and biomedical AI, participants were more likely to address the harms of error rates and specifically problems that may otherwise be included in a "disclaimer statement". In cybersecurity and cryptography, P9 asked, "Oh, does this mean I need to confess what the privacy vulnerabilities are?" P11 was similarly concerned specifically with privacy. Many would simply append a potential negative impact to the end, though some would use the QA model as a way of finding a location they could edit or alter. It seemed including changes at the end of their statement allowed the user to easily view when their changes were successful with the bot.

In P1 and P3's cases, however, the initial framing of the negative impacts was less explicit. While P1 initially presumed that for biomedicine, they noted that this seemed successful. Rather than dwelling on the negative classification output, they instead wanted to confuse the QA model seeing statements that could be perceived as having negative impacts. Misinterpreting ***ImpactBot***, P1 gamified the session in the opposite direction of the intended use. In both P1 and P3's case, when the participants made an active effort to include the negative impacts, the impacts themselves and the language were both seemingly more niche. Eventually in

P1, a new negative impact (privacy concern) had to be integrated to successfully move past the classifier. As a result, the classifier had trouble isolating them, though the QA models were able to correctly isolate the phrases. P7 had similar trouble, but altering the language to be more explicit in isolating the negative use case solved this problem. The classifier performs better with distinct and exact writing choices.

After reflecting and writing about potential negative impacts, the chatbot asked about how to mitigate these impacts. In these cases, participants added different approaches including safeguards for informing how the system was trained and specifying the intended uses of the technology. Participants frequently re-read the negative impact inside the statement and discussed what should be done. 4 in 12 participants designed use guidelines in the following sentence which are guidelines illustrating when and how the research should be implemented. This was because, in many cases, the negative impacts derived from a specific use case. Therefore, encouraging responsible use effectively avoids said impact. Others focused on revision strategies. Most of the research projects did not contain a mitigation strategy, so users would write a second sentence following the negative impact proposing a solution or what researchers should explore next.

In terms of the dialogue and classification outputs, the ***ImpactBot*** presented positive results. For instance, when interacting with the chatbot, P1 said: *"OK! Yep! So, it is saying that it find {found sic} that I did write about potential negative impact... and that I did not write how to mitigate those negative impacts, which I think it {ImpactBot's dialogue agent} is correct."*

The dialogue structure helped generate interesting reflections and participants thought about the steps considered in the dialogue, discussing more nuances and being more specific. For instance, after adding potential negative impacts, the chatbot asked for mitigation strategies for the impacts added and P1 said: *"I should've written something more about how to mitigate those negative impacts."*. P3 mentioned they, *"didn't previously talk much about the consequences"* so for the developing project, *it is helpful to preempt these solutions*. In another case, P4 was challenged to identify a solution to a problem beyond the scope of their standard role on the research project. In conclusion, P4 said, *"We have a lot of things to do at the same time. This is really helpful to address the needs of the text."* For most participants, it could help indicate the next steps for the research, when they had not previously addressed solutions to negative impacts.

Regarding impressions participants had about the UI itself, it was important to follow the specific interactivity patterns of users. Many struggled with the notion of multiple text areas or placeholders to be read for various buttons and compositions. It was indicative of a structure that was potentially counterproductive to the process flow as it minimized the sense of progress or continuity with the single updating screen. User work suggests that a preliminary instructional navigation page followed by an input screen that then shifts upward with each progression may be more helpful. Rather than a controlled window that overwrites upon each new user impact, participants suggested a scrolling cache appending new output to the end.

The summarization model was often glossed over with nine of twelve readers either skipping or briefly skimming it without use. It's possible that because it wasn't explicitly relevant to the objectives, the users didn't value this feature. However, the QA models were especially helpful in allowing participants to consider what parts of the impact statements to adapt. Participants P8, P9, and P12 all listed that even when the classifier correctly pointed out their lack of negative impact, the QA model would point to an area that had potentially high risk. Additionally, P10 mentioned, *"It was helpful to know if what I thought my negative impact was really being recognized and especially what part of the sentence was used to recognize it."*

Ultimately, the poor performances could be correlated to the limited capability of the model to transfer knowledge out of domains it had frequently seen. Due to the nature of the small (~1000) sample size with research limited to categories at NeurIPS, the data variance was constrained and the domain heavily influenced the training performance and loss gradients. Given the limitations to the size of the training data, tuning the adapter layers and classifier layers yielded less of a direct positive effect. The model was less adaptable as a result of the training parameters and limitations.

Specifically, high rates of false negatives correlated to understated impacts, especially those with verbiage or semantic meanings that weren't present whatsoever in the selected training data. As a result, commonly referenced use concerns were easily tagged such as privacy or bias, but many more nuanced negative impacts went unrecognized in sessions leading to error reports. For example, as a result of NeurIPS guidelines, disclaimer-framed statements that addressed impacts through a mitigation strategy are an uncommon writing pattern. In sample testing with various generated impact statements, issues from 2 user sessions were easily reproduced in failing to identify specific negative use cases or impacts that weren't mentioned in the training or validation data.

## 6 DISCUSSION

According to the ACM code of ethics [8], computing professionals should take extraordinary care to identify and mitigate potential risks. Chubb and Watermeyer [6] advocate active responsiveness and responsibility in research. Our research builds on the existing literature by proposing a way to promote critical thinking as part of a dialogue around impact statements.

According to most of the participants, ***ImpactBot*** would best serve users with repeated usage. Research evolves over the course of new methods and intermediate results. Adapting impact statements accordingly at the beginning, intermediate, and end stages of the research encourage researchers to similarly adapt how they address new use cases and potential harms that arise. ***ImpactBot*** may better accommodate changing systems. Further, it is especially helpful when the researchers already feel confident in how they can address potential harms. One user particularly noted that within internal procedures, they had progress checks and being reminded of the impact statements allows researchers to reconsider the next steps if negative use cases are yet to be adequately addressed. This still maintains self-autonomy granting psychological sovereignty over how they act.

That said, the use of a chatbot for guiding work that inherently requires careful thought and reflection on complex, value-laden issues could have some negative effects. Among these is over-reliance

on a single source of guidance – the ***ImpactBot*** tool is made by a handful of researchers whose design decisions are shaped by limited perspectives, and the prompts that users receive from it can only address a limited set of concerns. Whereas one purpose of impact statements is meant to encourage authors to *do some important work*, using a chatbot may undermine the hard work of building awareness or empathy, and might also constrain or redirect an author's thinking in ways that could otherwise surface important issues. As noted in the results, users develop a sense of complacence with mention of a negative impact. False confidence can encourage negligence of other potential negative impacts due to the singular pigeonhole that the bot identifies via classification structure.

Related to that effect is a potentially false sense of due diligence which may come from answering only prescribed questions, similar to that which has been noted in studies on users of ethics checklists [3]. This was also mentioned by two participants in the sessions. While we have designed ***ImpactBot*** to prompt researchers to consider issues that we have found most common and viable to address, it is by no means a catch-all application. A researcher who uses ***ImpactBot*** may best view it as a beginner's exercise; a warm-up question for a number of others that may be important to ask. Of note, the prompting questions we have designed into this tool are a subset of dozens of others from an existing responsible-technology framework. In future versions of the ***ImpactBot*** tool, we aim to incorporate more of those questions and offer the user a more thorough coverage of potentially relevant questions.

We expect to publish the ***ImpactBot*** code as an open-source project so that it can be reused as part of content management or paper submission system, for instance. Our ultimate goal is to help researchers think about negative impacts and how to mitigate them (if any). Officially integrating an approach like this into existing systems might help promote accountability within an organization, and reduce potential harms.

Given that our work is built on top of the assumption that technocentrism and positivity bias permeate the IT industry and research, our work may be no exception. Gamification is a potent risk. In this sense, the proposed technology might end up functioning as a checklist only to be marked off in the project conception or paper writing phases. In order to further assess the real actions and countermeasures performed by researchers, we need to better understand how risks are mitigated throughout the whole research workflow. Our project is limited, however, in that we only interact with participants at one stage of their research. Having multiple sessions or opportunities to evaluate effectiveness at different research phases could improve user experience.

Finally, the next steps of this research include opening ***ImpactBot*** source code and running additional user sessions with more participants to better determine its accuracy and potential. This may include testing with purely generative models and extremely recent innovations such as ChatGPT and creating a supplemental tool that can help researchers with impact statement construction based on guiding questions. Ultimately, the next steps should highlight ***ImpactBot***'s capability to foment and inspire critical thinking and catalyze cultural change.

# REFERENCES


[1] Bayan AbuShawar and Eric Atwell. 2016. Automatic extraction of chatbot training data from natural dialogue corpora. In *RE-WOCHAT: Workshop on Collecting and Generating Resources for Chatbots and Conversational Agents-Development and Evaluation*. 29–38.

[2] Muhammad Ashfaq, Jiang Yun, Shubin Yu, and Sandra Maria Correia Loureiro. 2020. I, Chatbot: Modeling the determinants of users' satisfaction and continuance intention of AI-powered service agents. *Telematics and Informatics* 54 (2020), 101473.

[3] Carolyn Ashurst, Emmie Hine, Paul Sedille, and Alexis Carlier. 2021. AI Ethics Statements - Analysis and lessons learnt from NeurIPS Broader Impact Statements. arXiv:2111.01705 https://arxiv.org/abs/2111.01705

[4] Charlotte Bloch. 2002. Managing the emotions of competition and recognition in academia. *The Sociological Review* 50, S2 (2002), 113–131.

[5] Ching-Yi Chang, Shu-Yu Kuo, and Gwo-Haur Hwang. 2022. Chatbot-facilitated Nursing Education: Incorporating a Knowledge-Based Chatbot System into a Nursing Training Program. *Educational Technology & Society* 25, 1 (2022), 15–27. https://www.jstor.org/stable/48647027

[6] Jennifer Chubb and Richard Watermeyer. 2017. Artifice or integrity in the marketization of research impact? Investigating the moral economy of (pathways to) impact statements within research funding proposals in the UK and Australia. *Studies in Higher Education* 42, 12 (2017), 2360–2372. https://doi.org/10.1080/03075079.2016.1144182 arXiv:https://doi.org/10.1080/03075079.2016.1144182

[7] Fernando Fogliano, Fernando Fabbrini, André Souza, Guilherme Fidélio, Juliana Machado, and Rachel Sarra. 2019. Edgard, the Chatbot: Questioning Ethics in the Usage of Artificial Intelligence Through Interaction Design and Electronic Literature. In *Digital Human Modeling and Applications in Health, Safety, Ergonomics and Risk Management. Healthcare Applications*, Vincent G. Duffy (Ed.). Springer International Publishing, Cham, 325–341.

[8] Don Gotterbarn, Bo Brinkman, Catherine Flick, Michael S Kirkpatrick, Keith Miller, Kate Varansky andMarty J Wolf, Eve Anderson, Ron Anderson, Amy Bruckman, Karla Carter, Michael Davis, Penny Duquenoy, Jeremy Epstein, Kai Kimppa, Lorraine Kisselburgh, Shrawan Kumar, Andrew McGettrick, Natasa Milic-Frayling, Denise Oram, Simon Rogerson, David Shamma, Janice Sipior, Eugene Spafford, and Les Waguespack. 2018. ACM Code of Ethics and Professional Conduct. Retrieved October 10, 2022 from https://www.acm.org/code-of-ethics

[9] Intan Permata Hapsari and Ting-Ting Wu. 2022. AI Chatbots Learning Model in English Speaking Skill: Alleviating Speaking Anxiety, Boosting Enjoyment, and Fostering Critical Thinking. In *Innovative Technologies and Learning*, Yueh-Min Huang, Shu-Chen Cheng, João Barroso, and Frode Eika Sandnes (Eds.). Springer International Publishing, Cham, 444–453.

[10] Zhenhui Peng. 2021. Designing and Evaluating Intelligent Agents' Interaction Mechanisms for Assisting Human in High-Level Thinking Tasks. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI EA '21)*. Association for Computing Machinery, New York, NY, USA, Article 70, 6 pages. https://doi.org/10.1145/3411763.3443424

[11] Carina EA Prunkl, Carolyn Ashurst, Markus Anderljung, Helena Webb, Jan Leike, and Allan Dafoe. 2021. Institutionalizing ethics in AI through broader impact requirements. *Nature Machine Intelligence* 3, 2 (2021), 104–110.

[12] Miriya Samarakoon and John S Rowan. 2008. A critical review of environmental impact statements in Sri Lanka with particular reference to ecological impact assessment. *Environmental Management* 41, 3 (2008), 441–460.

[13] Donald A Schön. 1992. *The reflective practitioner: How professionals think in action*. Routledge.

[14] Lu Wang, Munif Ishad Mujib, Jake Ryland Williams, George Demiris, and Jina Huh-Yoo. 2021. An Evaluation of Generative Pre-Training Model-based Therapy Chatbot for Caregivers. *CoRR* abs/2107.13115 (2021). arXiv:2107.13115 https://arxiv.org/abs/2107.13115

[15] CP Wolf. 2019. The Cultural Impact Statement 1. In *Cultural Resources: Planning and Management*. Routledge, 178–193.

[16] Yu Wu, Wei Wu, Chen Xing, Can Xu, Zhoujun Li, and Ming Zhou. 2019. A Sequential Matching Framework for Multi-Turn Response Selection in Retrieval-Based Chatbots. *Computational Linguistics* 45, 1 (03 2019), 163–197. https://doi.org/10.1162/coli_a_00345 arXiv:https://direct.mit.edu/coli/article-pdf/45/1/163/1809677/coli_a_00345.pdf

[17] Yonghan Zhu, Rui Wang, and Chengyan Pu. 2022. "I am chatbot, your virtual mental health adviser." What drives citizens' satisfaction and continuance intention toward mental health chatbots during the COVID-19 pandemic? An empirical study in China. *DIGITAL HEALTH* 8 (2022), 20552076221090031. https://doi.org/10.1177/20552076221090031 arXiv:https://doi.org/10.1177/20552076221090031 PMID: 35381977.