Toward a Human-Centered Framework for Trustworthy, Safe and Ethical Generative Artificial Intelligence: A Multi-Level Analysis of Large Language Models Social Impact

Berenice Fernández Nieto
berenice.fernandeznieto@uniba.it
University of Bari "A. Moro"
Bari, Italy

ABSTRACT

This research proposal aims to comprehensively explore the trustworthy, safe, and ethical use of Generative Artificial Intelligence (GAI), particularly Large Language Models (LLMs). To this end, we examine the risks and potential social hazards of LLMs, adopting a multidimensional approach— focused on society, human rights, and ethics— involving various stakeholders, including the AI industry, governmental institutions, and regulatory organizations, among others. This strategy allows for offering a research proposal grounded on social and technological dimensions and providing a comprehensive diagnosis, including perceived challenges in the AI industry, the regulatory debate, ethical dilemmas, etc. By delving into these areas, we aim to design a post-audit tool to ensure models are trustworthy, socially responsible, and in alignment with human rights. Additionally, we aim to encourage responsible AI Innovation through Ethics-Driven Incentives.

Supervisor: Prof. Danilo Caivano, danilo.caivano@uniba.it, University of Bari "A. Moro"

Co-supervisor: Dr. Azzurra Ragone, azzurra.ragone@uniba.it University of Bari "A. Moro"

ACM Reference Format:
Berenice Fernández Nieto. 2024. Toward a Human-Centered Framework for Trustworthy, Safe and Ethical Generative Artificial Intelligence: A Multi-Level Analysis of Large Language Models Social Impact. In *28th International Conference on Evaluation and Assessment in Software Engineering (EASE 2024), June 18–21, 2024, Salerno, Italy.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3661167.3661177

1 INTRODUCTION

Throughout history, science has prioritized social well-being and human progress. However, many innovations can inadvertently lead to adverse outcomes, often unrelated to their primary objectives. Thus, thorough diagnostics are crucial to identify potential hazards. While it's impossible to predict all negative consequences, pinpointing sensitive areas enables the implementation of safeguards and monitoring of social issues that scientific innovation should not exacerbate.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
EASE 2024, June 18–21, 2024, Salerno, Italy
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-1701-7/24/06
https://doi.org/10.1145/3661167.3661177

Currently, the domain of Artificial Intelligence (AI) bears a significant responsibility. This responsibility entails a continuous and reflexive commitment to society, as AI generates various social impacts spanning from the labor market to human rights.

On the other hand, the perception of innovation in AI should not be stigmatized and seen as an impending risk to humanity. Instead, it should be considered an opportunity to cultivate novel dynamics of accountable collaboration, construct efficient oversight mechanisms, and pursue technical progress that prioritizes individuals and acknowledges complex social contexts.

The latter point is particularly relevant since each AI artifact operates in distinct environments where they can contribute to the expansion and deepening of unequal social dynamics. An illustration of this is the social impact of outsourcing employees to label data for Large Language Models (LLMs) training [17]. Although AI companies did not create the outsourcing scheme where workers receive little economic compensation and fewer labor rights [14], they do play a role in these unequal dynamics and contribute to various social issues when opting for this kind of employment scheme. Consequently, a human-centered vision is required at all stages of AI development.

2 RESEARCH MOTIVATION, RQS AND OBJECTIVES

The accelerated development of Large Language Models in recent years has generated heightened concern regarding Artificial Intelligence and its resulting societal implications. One of the primary concerns revolves around the potential negative consequences, with workforce displacement being a common worry due to increased task automation. Simultaneously, an important debate is occurring regarding ethics and the pressing need for timely and efficient regulation. This discourse encompasses various complex aspects, including copyright matters, privacy considerations, efforts to address bias, cybersecurity measures, and the risk of spreading disinformation. Given this context, it is crucial to analyze LLMs' social effects comprehensively. This proposal intends to do it through a three-dimension analysis, which includes society, human rights, and ethics. This strategy seeks not only to understand the social effects of generative AI but also to contribute to the development of responsible models from a human rights perspective.

Regarding research questions, our central question is:

RQ1. How do LLMs impact society in terms of trustworthiness, ethical considerations, and their impact on human rights?

While the secondary questions are:

RQ2. How do governments, policymakers, and companies deal with it?

RQ3. How can the AI industry address the challenges derived from LLMs' impact on Human Rights?

Concretely, our main objectives is as follows:

- Examine the multifaceted implications of LLMs, integrating ethical considerations and their societal effects into a holistic framework for their trustworthy integration. This framework, aimed at AI companies, will provide a post-release audit tool to uphold safety and human rights standards.

Our secondary objectives are:

- To thoroughly examine how governments, regulatory institutions and AI companies navigate and adapt their practices in response to the challenges posed by LLMs.
- Define and comprehend LLMs' ethical and human rights challenges, fostering responsible AI innovation through Ethics-Driven Incentives.

3 WORK PLAN

A thorough analysis of the societal and human rights implications of LLM demands a comprehensive approach that encompasses various dimensions, such as:

(1) **Internal Operations dimension:** This encompasses the internal operations of AI companies, including their policies, personnel practices, and research methodologies.
(2) **Product Development dimension:** This focuses on the models developed by the AI industry, examining their design and features.
(3) **External Interactions dimension:** This pertains to the interactions of AI companies with their external environment, including compliance with laws and regulations, engagement with governments, interactions with users, and collaborations with other companies.

By exploring these dimensions, we can conceptualize the AI industry's companies as dynamic entities comprised of diverse elements. These elements facilitate the development of models that engage with and impact their environment. Moreover, it allows us to examine how companies interact with regulatory institutions, other companies, and users. Adopting this holistic perspective allows us to recognize the significant societal challenges presented by LLMs.

As a visual aid, we can use a tree as an analogy for the model, see Fig. 1. The roots represent the internal operations, including policies, regulations, staff, and research, enabling the company's function. The trunk symbolizes the product development dimension, leading to the innovative models. Finally, the branches depict the various impacts that companies generate in their environment, as well as their interactions with other actors.

The external dimension includes studying how companies, institutions, and legal frameworks respond to LLMs' presence and the subsequent effects on companies' commitment to their social environment. This also includes examining the ethical dimension, encompassing a broad scope of concerns, including bias, privacy, cybersecurity, environmental impact, transparency, explainability, and accessibility in LLMs. To examine these dimensions, we will also use quantitative approaches to evaluate patterns in regulatory

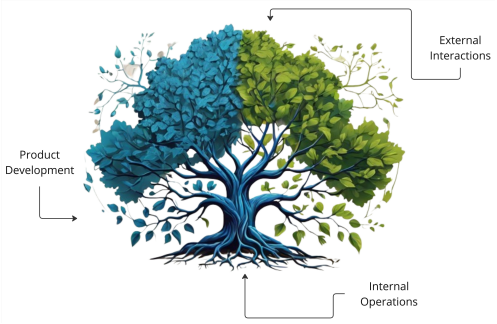

Figure 1: Research Approach

modifications. Additionally, qualitative research methods will be employed to interview ethics experts, industry professionals, and academics, thereby obtaining a comprehensive perspective of the social and ethical environment.

3.1 Instruments for Data Collection

To effectively conduct this study, a mixed methods research methodology will be employed, encompassing a diverse array of tools for data collection [8], such as:

- **Surveys**: We will conduct surveys to capture quantitative data on various aspects of LLM impact, including economic shifts, legal modifications, and ethical considerations.
- **Interviews**: We will conduct in-depth interviews with relevant stakeholders, including AI industry representatives, legal experts, policymakers, and ethics experts. These interviews will provide qualitative insights and a variety of perspectives regarding the impact of LLMs.
- **Document Analysis**: We will comprehensively analyze legal documents, policy reports, media content, and extensive academic literature to triangulate findings and thoroughly understand the research's dimensions [12].

3.2 Mixed Methods Research

The mixed methods approach in this research will ensure that our study accounts for the complexity of LLM's impact on society across the three established dimensions. Our study design encompasses the analysis of qualitative and quantitative data to respond to our research questions using surveys and interviews, in which we will employ both open-ended and closed-ended questions.

We plan to employ two mixed methods: 1) follow-up Quantitative Data Collection and Analysis (QUAN) with QUAL Qualitative Data Collection and Analysis (QUAL) to aid in interpretation, and we will 2) use QUAL data to build QUAN for interpretation [6] see Figures 2 and 3.

These two methods will allow us to examine AI companies developing Generative AI. Further, tools like document analysis and interviews for exploring AI ethical and human rights concerns will be enriched through this methodology [8].

We will also provide surveys to a population sample to ascertain if the qualitative results can be generalized to the entire population [8].

Figure 2: Mixed Methods type 1

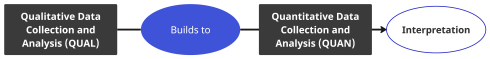

Figure 3: Mixed Methods type 2

To ensure the responsible development of the post-audit tool, experts in ethics and human rights will be involved in the analysis of the ethical and social dimension.

Furthermore, we will employ the respective validation methods for each type of research method. For method 1, we will establish the validity of the scores of the quantitative measures and discuss the validity of the quantitative findings. Meanwhile, for method 2, we will review the validity of both the quantitative and qualitative data, all following Creswell's recommendations [6].

4 DATA ANALYSIS

After completing the data collection phase, we will conduct data analysis. The research methodology will use qualitative and quantitative methods to analyze the collected data. We will adopt qualitative approaches to investigate nuanced thoughts, attitudes, and opinions, while with quantitative methods, we will uncover patterns, trends, and correlations within the information. We will also conduct comparative analysis [18, 21] to gain a deeper understanding of the relative strengths, weaknesses, and characteristics of each item, providing valuable insights for decision-making or further study.

The analysis will be also performed via specialized software and instruments to guarantee precision and dependability, such as Atlas.ti[1], Dedoose [2] and NVivo[3], all of these tools used in mixed-methods research.

5 RESULTS EVALUATION

The qualitative data will be subjected to a rigorous *thematic analysis* [4, 19], *member checking* [3, 5], and *triangulation* [10, 15]. Simultaneously, quantitative data will be subjected to *hypothesis testing* [11, 13], *regression analysis* [2], and *cross-tabulations* [7] to identify essential correlations and predictors. One notable aspect of this research is its capacity to establish connections among several dimensions, facilitating the detection of overarching themes. This cross-dimensional analysis will provide a holistic understanding.

[1]https://atlasti.com/
[2]https://www.dedoose.com/
[3]https://lumivero.com/

6 MAIN OUTCOME DELIVERY

Based on the conclusions drawn from data collection and data analysis phases, we will build our main output an **evaluative framework to evaluate human rights and ethical impact of Large Language Models**.

This tool will offer recommendations to AI companies, especially, to their Ethics and compliance teams, Corporate Social Responsibility (CSR) departments, Diversity, Equity, and Inclusion (DEI) tems, and Public policy and government Affairs teams. It will be designed to facilitate post-release audits, ensuring the models' trustworthiness and safety. We aim to guarantee that LLMs do not cause harm to society and adhere to human rights standards.

The development of this instrument will be a collaborative effort, encompassing consultations with experts from diverse disciplines, such as AI ethics, cybersecurity, legal scholars, and representatives from AI companies. The project will be conducted iteratively, incorporating feedback loops and making adjustments as necessary to ensure that the framework is resilient, flexible, and aligned with the dynamic technology landscape.

7 VALIDITY THREATS AND MITIGATION STRATEGIES

Throughout the course of this project's development, a variety of potential threats to validity may emerge. In the following section, we outline the significant threats to the validity of our study on the impact of LLMs on society and human rights and suggest appropriate mitigation measures.

Using surveys, we recognize both the advantages and limitations associated with their use. Surveys serve as the initial phase of our research effort, as they offer a convenient means of data collection that is comparatively less demanding in terms of time and effort when contrasted with interviews or other forms of qualitative research conducted in person. The aforementioned point serves as the foundation for a more comprehensive investigation into the problem under examination. Surveys prove to be valuable in providing preliminary insights that can be leveraged for subsequent investigations. Nevertheless, surveys are subject to certain limitations, such as the potential for respondents to encounter difficulties comprehending specific questions, resulting in responses that may not accurately reflect their authentic understanding. This is the reason why we also plan to conduct in-person interviews using open-ended questions.

Aside from the difficulties presented by surveys, various other potential threats to validity need to be considered. **Response bias** is an example of a possible threat, which occurs when participants may give inaccurate or skewed answers due to social desirability bias or other factors [16]. We will implement randomized response methodologies to address this potential risk and ensure anonymity and confidentiality in survey administration [16]. **Ethical considerations** also threaten the validity of our research, especially when focusing on rights such as privacy, non-discrimination, fairness, and freedom of expression. Given the complexity and sensitivity of these issues, ethics and human rights professionals will be included to ensure appropriate consideration of ethical and human rights standards [20].

Our research will follow ethical and human rights standards due to the complexity and sensitivity of these issues. This requires informed consent from participants, data anonymization and confidentiality, and **ethical approval from relevant institutional review boards** [9].

8 EARLY RESULTS

Early findings encompass a systematic literature review that extensively explored the negative and positive impacts, emerging patterns, and potential opportunities associated with Generative Artificial Intelligence, focusing on ChatGPT[1]. The findings indicate that although the assessment of generative AI models' social impact is still in its early stages, significant areas of concern have surfaced, particularly regarding privacy and the risk of bias [1]. Moreover, with the increasing adoption of generative models across diverse social contexts, an urgent need arises to address and mitigate issues related to inequality, prejudice, discrimination, and stereotyping [1].

Another early finding consists of examining the beneficial impacts of Large Language Models (LLMs) on society, focusing on their potential to facilitate informal learning. The investigation evaluates how LLMs, specifically ChatGPT, can encourage citizens' civic engagement by helping them comprehend complex legal documents, such as the AI Act. Through the administration of 73 online surveys featuring closed-ended and multiple-choice questions and the application of a readability test to the texts utilized, our study revealed that participants exposed to a simplified ChatGPT-generated text demonstrated a superior grasp of the legal content compared to those who solely read the original text. Nevertheless, further investigation is necessary to clarify the effectiveness of these models in improving informal learning and citizen civic engagement. Details of this study are outlined in "Large Language Models to Enhance Informal Learning and Foster Citizens Civic Engagement" authored by Maria Teresa Baldassarre, Berenice Fernandez Nieto, Domenico Gigante, and Azzurra Ragone. The article is presently undergoing review.

Finally, another preliminary result entails a systematic review and comparative analysis of strategies and measures implemented by four prominent companies — OpenAI, Meta, Google, and Microsoft — across five critical dimensions: bias, privacy, cybersecurity, hate speech, and disinformation. By scrutinizing 192 publicly available documents, our study reveals that, depending on product diversity and nature, certain companies excel in researching and developing privacy-preserving and bias-reducing technologies and methodologies. They provide user-friendly tools for managing personal data, establish expert groups to explore the social impact of their technologies, and possess significant expertise in combatting hate speech and disinformation. However, there is an urgent need for greater linguistic, cultural, and geographic diversity in research lines, tools, and collaborative efforts. From this analysis, we distill actionable best practices to foster the responsible development of AI models, particularly Large Language Models, aligned with human rights principles. Titled "Fostering Human Rights in Responsible AI: A Systematic Review for Best Practices in Industry," this study authored by Maria Teresa Baldassarre, Danilo Caivano, Berenice Fernandez Nieto, Domenico Gigante, and Azzurra Ragone is currently under review.

9 NEXT STEPS

The next steps involve enriching our evaluative framework through the integration of expert inputs and stakeholder consultations. This collaborative effort aims to enhance the comprehensiveness and effectiveness of our assessment approach. Additionally, we will explore existing ethical frameworks to understand their applicability and limitations in the context of Generative AI. This exploration will provide valuable insights into the ethical considerations inherent in our evaluation process. Subsequently, we will consolidate our evaluative framework, particularly focusing on the development of a post-audit tool. This tool will serve as a key instrument in integrating human rights principles and ethical considerations into LLMs. Furthermore, we are committed to encouraging responsible AI innovation through the design of ethics-driven incentives.

REFERENCES

[1] Maria Teresa Baldassarre, Danilo Caivano, Berenice Fernandez Nieto, Domenico Gigante, and Azzurra Ragone. 2023. The Social Impact of Generative AI: An Analysis on ChatGPT. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good* (Lisbon, Portugal) *(GoodIT '23)*. Association for Computing Machinery, New York, NY, USA, 363–373. https://doi.org/10.1145/3582515.3609555

[2] Henning Best and Christof Wolf (Eds.). 2015. *The SAGE Handbook of Regression Analysis and Causal Inference*. SAGE Reference.

[3] Linda Birt, Suzanne Scott, Debbie Cavers, Christine Campbell, and Fiona Walter. 2016. Member Checking: A Tool to Enhance Trustworthiness or Merely a Nod to Validation? *Qualitative Health Research* 26, 13 (2016), 1802–1811. https://doi.org/10.1177/1049732316654870

[4] Virginia Braun and Victoria Clarke. 2022. *Thematic Analysis: A Practical Guide*. SAGE.

[5] Benjamin F. Crabtree and William L. Miller. 2023. *Doing Qualitative Research* (3 ed.). SAGE Publications, Inc.

[6] John W. Creswell. 2014. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (4th ed.). SAGE Publications.

[7] John W Creswell and J David Creswell. 2018. *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (fifth ed.). SAGE.

[8] John W. Creswell and Vicki L. Plano Clark. 2017. *Designing and Conducting Mixed Methods Research* (edición de kindle ed.). SAGE Publications.

[9] Karen De Wet. 2010. The Importance of Ethical Appraisal in Social Science Research: Reviewing a Faculty of Humanities' Research Ethics Committee. *J Acad Ethics* 8 (2010), 301–314. https://doi.org/10.1007/s10805-010-9118-8

[10] Jennifer Esposito and Venus E Evans-Winters. 2022. *Introduction to Intersectional Qualitative Research*. SAGE Publications, Inc.

[11] Patrick M Fay and Evan H Brittain. 2022. *Statistical Hypothesis Testing in Context: Reproducibility, Inference, and Science*. Cambridge University Press.

[12] Uwe Flick. 2018. *The SAGE Handbook of Qualitative Data Collection*. SAGE Publications Ltd. https://doi.org/10.4135/9781526416070

[13] Güzin Gül. 2017. *Robust and Distributed Hypothesis Testing*. Vol. 414. Springer International Publishing. https://doi.org/10.1007/978-3-319-49286-5

[14] R Hudiana and H Susetyo. 2020. Law and Human Right Protection of Outsourcing Labour Law Number 13 of 2003. In *Proceedings of the 3rd International Conference on Law and Governance (ICLAVE 2019)*. Solo, Central Java, Indonesia. https://doi.org/10.2991/aebmr.k.200321.026

[15] Thomas M Mulvihill and Raji Swaminathan. 2023. *Collaborative Qualitative Research*. The Guilford Press.

[16] Delroy L Paulhus. 1991. Measurement and control of response bias. In *Measures of personality and social psychological attitudes*, John P Robinson, Phillip R Shaver, and Lawrence S Wrightsman (Eds.). Academic Press, 17–59. https://doi.org/10.1016/B978-0-12-590241-0.50006-X

[17] Billy Perrigo. 2023. Exclusive: The $2 Per Hour Workers Who Made ChatGPT Safer. https://time.com/6247678/openai-chatgpt-kenya-workers/

[18] Benoît Rihoux and Heike Grimm (Eds.). 2006. *Innovative Comparative Methods for Policy Analysis*. Springer US. https://doi.org/10.1007/0-387-28829-5

[19] Gareth Terry and Nikki Hayfield. 2021. *Essentials of Thematic Analysis*. American Psychological Association.

[20] The Norwegian National Research Ethics Committees. 2022. *Guidelines for Research Ethics in the Social Sciences and the Humanities*. https://www.forskningsetikk.no/en/guidelines/social-sciences-and-

humanities/guidelines-for-research-ethics-in-the-social-sciences-and-the-humanities/

[21] Stanisław Wrycza (Ed.). 2011. *Research in Systems Analysis and Design: Models and Methods: 4th SIGSAND/PLAIS EuroSymposium 2011, Gdańsk, Poland, September 29, 2011, Revised Selected Papers*. Vol. 93. Springer. https://doi.org/10.1007/978-3-642-25676-9

Received 15 March 2024; revised 12 April 2024; accepted 5 May 2024