Fostering human rights in responsible AI: a systematic review for best practices in industry

Fostering Human Rights in Responsible AI: A Systematic Review for Best Practices in Industry

Baldassarre, Maria Teresa; Caivano, Danilo; Fernández Nieto, Berenice*; Gigante, Domenico, and Ragone, Azzurra*.

***Abstract*—The recent rapid development of Generative AI, and the resulting market growth, has introduced new challenges for social responsibility, an area where companies may need more guidance. In this regard, the literature covers a broad spectrum, from the impact of bias to the potential use of this technology to implement undemocratic surveillance. Another focus area discusses the AI industry's commitment to human rights and social responsibility, examining the diverse actors involved in this commitment and the context-dependent nature of their impact on human rights.**

This work performs a systematic review and a comparative analysis of the strategies and actions taken by four leading companies — OpenAI, Meta AI Research, Google AI, and Microsoft AI — with respect to five critical dimensions: bias, privacy, cybersecurity, hate speech, and misinformation. Our study analyzes 192 publicly available documents and reveals that depending on the diversity of products and their nature, some companies excel in the research and development of technologies and methodologies for privacy preservation and bias reduction, offering user-friendly tools for managing personal data, establishing expert groups to research the social impact of their technologies, and possessing significant expertise in tackling hate speech and misinformation. Nonetheless, there is an urgent need for greater linguistic, cultural, and geographic diversity in research lines, tools, and collaborative efforts.

From this analysis, we draw a set of actionable best practices aimed at supporting the responsible development of AI models, and Foundation Models in particular, that are aligned with human rights principles.

***Impact Statement*—Responsible AI practices have been widely studied. While recommendations fostering Responsible AI exist, these are often high-level statements that are sometimes difficult to translate into concrete implementation strategies. Currently, there is a significant gap between high-level AI ethics principles and low-level concrete practices for AI companies. In this study, we provide actionable advice for AI companies to address the impacts that their implementation choices, especially in the field of Generative AI, may have in the social sphere, with particular reference to the human rights safeguard. Through an in-depth investigation of what current actions, strengths and limitations, have been taken by leading AI companies, we identify a set of best practices, either already in place or that would be good to implement. These are concrete actions that the AI industry can deploy to promote a next generation of AI firmly rooted in a human rights-centered perspective.**

***Index Terms*—Responsible AI, Trustworthy AI, Generative AI, Human-centered AI, Human Rights.**

I. Introduction

Generative Artificial Intelligence (AI) has made significant progress in recent years, providing advantages across diverse sectors, including the creative industry and medicine [106] [14]. Nevertheless, there are domains and contexts in which Generative AI can impact human rights, including but not limited to privacy, non-discrimination, freedom of opinion and expression, freedom of peaceful assembly and various fundamental freedoms. Within this framework, Foundation Models are a rapidly expanding sector. For instance, after its launch, ChatGPT — developed by OpenAI — gathered a user base of over 100 million within less than two months [79]. These developments underscore the pressing need for the progression of *responsible* and *rights-committed* Generative AI.

In this scenario, the private sector — particularly companies specializing in AI development — assumes a pivotal role. In this respect, the Office of the High Commissioner for Human Rights (OHCHR) recognizes two distinct approaches to regulating Artificial Intelligence [118]. The first approach is a risk-based strategy that assigns substantial accountability to the private sector by identifying and addressing risks to attain desired results. The second incorporates human rights principles into every AI development and implementation stage. This approach includes human rights principles throughout the data collection and selection process and the design, development, deployment, and use of the resulting models, tools, and services. As per OHCHR, immediate measures are required today to prevent, tackle, and mitigate the potential adverse effects of Generative AI [118].

The human rights issue in the AI industry is a topic of ongoing and continuous discussion. Various organizations, including the United Nations [119], International Amnesty [7], European Center for Not-for-Profit Law Stichting [35], Access Now [3], AI Now Institute [4], and others, actively seek to collaborate with the industry to advance Responsible AI.

This work aims to contribute to these efforts by attempting to fill the gap between high-level principles and low-level concrete practices that can be deployed by the AI industry to address the challenges and concerns of one particular area of Responsible AI, the one connected to human rights protection.

Submitted on November 1, 2023

M. T. Baldassarre is with the University of Bari "A. Moro", Bari, 70125, Italy (e-mail: mariateresa.baldassarre@uniba.it).

D. Caivano is with the University of Bari "A. Moro", Bari, 70125, Italy (e-mail: danilo.caivano@uniba.it).

B. F. Nieto is with the University of Bari "A. Moro", Bari, 70125, Italy (e-mail: berenice.fernandeznieto@uniba.it). She is also with IMT Lucca, 55100, Italy.

D. Gigante is with the University of Bari "A. Moro", Bari, 70125, Italy (e-mail: domenico.gigante1@uniba.it).

A. Ragone is with the University of Bari "A. Moro", Bari, 70125, Italy (e-mail: azzurra.ragone@uniba.it).

This paragraph will include the Associate Editor who handled your paper.

* Authors are listed in alphabetical order. Corresponding authors: Berenice Fernández Nieto (berenice.fernandeznieto@uniba.it) and Azzurra Ragone (azzurra.ragone@uniba.it).

The human rights doctrine can, on one side, contribute to reducing the gap between high-level AI ethics principles and low-level concrete practice, and on the other side, it can support the definition of AI ethics principles, as human rights can support value alignment for AI systems across a range of different national and social contexts [42]. The human rights vision can help close the gap between research and practice and between scientists and civil society, providing a common vocabulary and a shared understanding [42].

The outcomes of our research yield the following research contributions:

- Through a **systematic review**, we have highlighted the specific measures taken by a set of big AI players (i.e. OpenAI, Meta AI Research, Google AI, and Microsoft AI) in addressing five different dimensions: *Bias*, *Misinformation*, *Hate speech*, *Cybersecurity*, and *Privacy*. These actions directly influence the protection of fundamental human rights, encompassing but not limited to the right to non-discrimination, health, security, access to information, free expression, and privacy, among others. Additionally, our assessment has identified both **strengths** and **weaknesses** in the company's measures and strategies.
- By conducting a **thematic analysis** of the most important publicly available documents related to each parameter in each selected organization, we have identified recurring themes that revolve around ethics, the need for diversity and representation, and the importance placed on collaboration and research.
- In light of these findings, we offer a set of **actionable best practices** that these and other companies can embrace to enhance their performance on these parameters. This, in turn, will strengthen their approach to Responsible and Trustworthy AI, solidifying their commitment to upholding human rights.

Building upon the research conducted in the study addressing the maturity of Responsible AI in [16] and the examination of challenges and gaps within the AI industry [1], our work makes a valuable contribution by proposing a series of best practices to be implemented within the AI industry. It is essential to emphasize that these recommendations draw inspiration from existing measures while also presenting a potential remedy for identified weaknesses. Collectively, these practices intend to uphold human rights within the burgeoning AI industry.

The paper is structured as follows: Section II provides background information about corporate social responsibility and human rights in the AI industry. Section III introduces our methodology, describing all the steps of the systematic review, including comparative and thematic analysis. Section IV provides a comparison of the companies' strategies with respect to the selected analysis parameters, and outlines the recommended best practices. Finally, Sections VI delves into the threats to validity, and Section VII presents our conclusion and future work.

II. Background and Context

Nowadays, a substantial body of scholarly literature focuses on human rights and artificial intelligence [106], [100], [111], [41], [102], [105], [64], [31], [113], [5]. This discussion spans a broad range of topics from fighting bias to the possible use of AI to implement non-democratic surveillance models.

Rowena Rodrigues [106] examines AI's legal and human rights concerns, such as the lack of legal personhood, intellectual property, and liability for damages. Her study evaluates the suggested solutions and the state of their implementation while considering persistent gaps and challenges. Overall, the analysis highlights the significant impact of AI applications on human rights, for instance how liability issues associated with harmful uses can affect people's right to life and access to effective remedies [106].

Emilie C. Schwarz [113] presents an additional analysis that aligns with the same perspective, highlighting the accountability of transnational corporations in Artificial Intelligence regarding human rights. The study examines possible uses of AI in different human activities and identifies the specific rights that may be impacted if responsible AI practices are not followed. Among these rights are the right to privacy, freedom of thought, freedom of expression, security, non-discrimination, peaceful assembly, and the right to work, all enshrined in the Universal Declaration of Human Rights [113].

On the other hand, Aizenberg [5] delves into the design of values to incorporate stakeholders and translate fundamental rights into context-dependent designs. The study places human rights in a central position, focusing on values derived from the Universal Declaration of Human Rights and its application and interpretation within the European AI industry.

Our work follows a similar trajectory and aims to help AI companies design Responsible AI systems, that can impact the exercise of human rights, by providing them with practical guidance. Through analysis of five parameters and an evaluative framework, we derive what could be the best practices for responsible AI systems development.

Lastly, it is crucial to recognize that the responsibility for unexpected social consequences resulting from AI development should be shared not only by the companies involved but also by decision-makers, regulatory bodies, service providers, and authoritative figures in sectors that are incorporating or considering the use of Generative AI in their operations. Developing a vision that prioritizes human rights and is socially responsible requires a collective effort.

A. AI Companies' Commitment to Human Rights and Social Responsibility

A social dilemma is a predicament where individuals must determine the best course of action. In this scenario, there are two possible choices: 1) looking out for one's own best interests or 2) looking out for those of the community. In the first scenario, the self-focused option guarantees personal well-being but at the expense of others. In the second scenario, the cooperative option entails personal sacrifice for the benefit of the group [98].

Similarly, the discourse revolves around the AI innovation dilemma, that is to weigh the different benefits of Generative AI against the potential impact on society and human rights. This responsibility falls upon AI companies at various stages, ranging from research and design to development.

Several scholars contend that responsibility is distributed among stakeholders like innovators, providers, and consumers [81].While these analyses aid in delineating the private sector's duties, they overlook the crucial role of ongoing reflexive research into diverse social contexts. In specific instances, a mere caution may not suffice to safeguard rights like privacy.

AI companies have an inherent responsibility to promote social well-being and protect human rights, given that their activities have significant implications for fundamental aspects of society, including democracy. This obligation, although not explicitly established, is indirectly mentioned in documents such as the Universal Declaration of Human Rights, the International Covenant on Civil and Political Rights, and the European Convention on Human Rights [124]. Companies are held responsible to the state in which they are situated, as well as to national laws, such as the penal code, and to the international duties of that state under human rights law [124]. Often obligations are not clearly established, which constitutes an *"accountability gap"* [124]. Another motivating factor for companies to prioritize respect for human rights is their aim to safeguard their reputation. Indeed, even large and prominent companies have been involved in several scandals related to the non-transparency or unfairness in their systems [125], [83], as well as to their propensity to generate misinformation [82]. One mitigation measure could be to develop and encourage voluntary mechanisms for AI companies to assess their products' potential social, ethical and human rights impacts [124], [22]. However, this requires government incentives, guidance from specialized human rights institutions [120], and – in particular – a novel mechanism for multidisciplinary and multisectoral cooperation. Especially as companies today face a legislative environment that imposes progressively greater burdens on social responsibilities and human rights, costing work, financial resources, and time to profit-oriented actors [26], [80], [15].

On the other hand, the impact of the AI industry extends beyond the economic realm and significantly influences social, political, and cultural contexts. Scholars sometimes refer to this tendency as "corporate power" [19]. This term acknowledges their influence in diverse areas, such as the labour market, narratives on social media, security, and other related domains. Therefore, it is crucial to place AI innovation within a multi-layered landscape, as the consequences for human rights depend on the cultural, social, economic, and political milieu [19].

Nevertheless, allocating social responsibilities to the private sector yields both adverse and favourable consequences. One of the negative consequences is that imposing excessive liability on innovators and AI companies may discourage competition, leading to the exit of smaller companies from the market and allowing economically robust entities to dominate the industry [81]. Conversely, one positive result is that the substantial weight of responsibility catalyzes collaboration among companies, service providers, and consumers, intending to mitigate any negative repercussions and cultivate a continuous feedback cycle. Striking a delicate equilibrium between effective regulation and fostering innovation is a highly complex matter, [81], a *Gordian knot* to be resolved in the coming years.

Despite those above, the field of study regarding the social dimensions of AI has grown rapidly in the last decade [21]. This study aims to enhance the understanding of strategies and activities related to human rights within the private sector and contribute to joint and multidisciplinary efforts for a trustworthy, accountable and people-centered AI.

III. Methodology

This qualitative study conducts a descriptive and comparative analysis of the measures implemented by four leading AI companies regarding *bias* reduction, *privacy* protection, *cybersecurity*, fighting *hate speech*, and addressing *misinformation*. We conducted a systematic (grey) literature [59] review of all the publicly available documents of these companies, followed by a comparative analysis.

A. Research Questions

Our research questions are as follows:

- **RQ1.** What Responsible AI practices, in terms of **risk mitigation and human rights protection strategies/actions**, do leading AI companies like OpenAI, Google, Microsoft, and Meta employ when developing AI models, and in particular Generative AI models?
- **RQ2.** What are companies' **strengths**, and **weaknesses** in risk mitigation and human rights protection actions?
- **RQ3.** Which **best practices** can be adopted to enhance risk mitigation and human rights safeguard within the AI industry, with particular reference to Generative AI development?

All the steps followed in our study are highlighted in the following sections (Figure 1).

B. Study design

To conduct our study, we implemented the following steps:

1) **Companies and parameters selection**: This phase involves identifying and selecting the specific companies to be analyzed and determining the evaluation's parameters or criteria.
2) **Establishing a search strategy**: A systematic strategy is defined to find relevant information in the published document sources.
3) **Adopting eligibility criteria for data extraction**: This is done in order to ensure that only meaningful and reliable sources are included.
4) **Analysing and synthesizing selected data**: The obtained data is evaluated, organized, and synthesized to make a meaningful analysis.
5) **Performing thematic analysis**: Using Atlas.ti we identified recurring themes within the synthesized data, providing a deeper understanding of the overarching trends and characteristics.

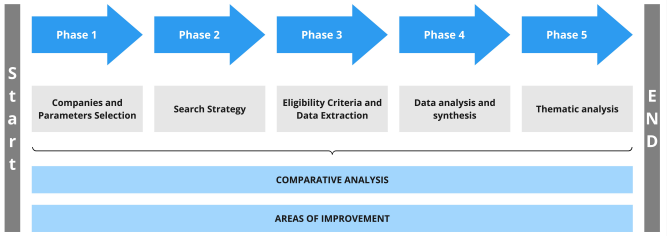

Fig. 1. Study design steps

The following sections provide a detailed description of each step.

C. Companies and Parameters Selection

In order to analyze how some of the Responsible AI issues are addressed by the industry, we conducted a comprehensive analysis of **four leading companies** operating within the AI industry: OpenAI, Meta AI Research, Google AI, and Microsoft AI. These companies were selected based on their substantial impact on the AI market and research, and their relevance across various domains [104].

Each of the four aforementioned companies exhibits a wide range of approaches and regulations for Responsible AI development, each with its own particular mission and vision.

OpenAI is a notable corporation renowned for its innovative work in Artificial Intelligence and great emphasis on research. OpenAI's mission is to create safe Artificial General Intelligence for humanity's benefit. ChatGPT and Dall-E are two of its products, and its research spans across the text, audio, and image domains [90].

Meta AI Research is a Meta Platforms, Inc. division. Meta's Fundamental AI Research Team (FAIR) is devoted to various AI-related topics. Meta's vision is to disseminate its research within the AI community and encourage collaborative efforts to develop Responsible AI [70].

Google AI has emerged as a prominent and influential participant in AI, making significant contributions. Google AI is committed to performing impactful AI research to enhance societal well-being. It actively participates in the academic community by sharing its research outcomes through open-source projects and facilitating global collaborations [50].

Microsoft AI has the objective of facilitating the empowerment of people and organizations by democratizing AI [73]. Azure, the Microsoft cloud computing platform, plays a vital role within their AI ecosystem by facilitating the deployment and management of AI applications and solutions [73].

The selection of the parameters for the analysis is based on the insights obtained from prior scholarly research into the perceived impacts (positive and negative) of Generative Artificial Intelligence [14] and the human rights challenges in AI [1]. Thus, based on the concerns identified in [14] and [1], we have compiled a preliminary inventory of rights that could be undermined by the misuse of Generative AI or by insufficient protective measures. A comprehensive list can be found in the online appendix [13]. In this initial research, we have focused on five specific issues that will serve as parameters. In the future, we plan to broaden our analysis to include other criteria such as algorithmic transparency, liability for damages caused, etc.. With the above in mind, the five parameters chosen are as follows:

1) **Bias**. Following the definition given in [36], we define bias as "*an inclination of prejudice towards or against a person, object, or position*". The potential for undesirable biases in AI systems to exacerbate existing social inequities — or even generate new ones — has recently received considerable attention across a range of academic disciplines, from AI to SE to public policy, law, and ethics [17], [43], [121]. Bias reduction is related to the rights of Equality and Non-Discrimination.
2) **Misinformation**. Is fake or inaccurate information, while disinformation is deliberately created to deceive or manipulate [39]. Fighting dis/misinformation is linked to protecting the right to information, freedom of expression, and participation in public affairs. Depending on the misinformation content or intention, it may also affect rights such as health, non-discrimination, life, and personal security, among others. **Note:** In this study, we employ the term misinformation because, regardless of whether it is part of an orchestrated operation or not, false information has a variety of detrimental effects on society — some of the companies' strategies analysed focus on disinformation or misinformation.
3) **Hate Speech**. It is defined as a type of offensive language that uses stereotypes to express a hateful ideology [122]. It is further described as any communication that disparages a person or group based on a characteristic such as race, color, ethnicity, gender, sexual orientation, national origin, religion, etc. [85]. Fighting hate speech is related to protecting rights such as the right to equality and non-discrimination, freedom of thought and expression, security of person, public order and safety.
4) **Privacy**. There is a large discussion on its definition; while some consider it "*the state in which a person is not observed or disturbed, or to be free from public attention*" [30] others hold that "*is not simply an absence of information about us in the minds of others, rather it is the control we have over information about ourselves*" [84]. For the purposes of this research, we refer to privacy as the right to have one's data processed in a way that results compliant with the seven protection and accountability principles outlined in GDPR Article 5.1-2 [37]: *Lawfulness, fairness and transparency*, *Purpose limitation*, *Data minimization*, *Accuracy*, *Storage limitation*, *Integrity and confidentiality*, and *Accountability*. Privacy is a human right in itself and is related to others, such as the protection from Unlawful Interference with Privacy and Respect for Private and Family Life.
5) **Cybersecurity**. According to the European Union Agency for Cybersecurity (ENISA), "*shall refer to security of cyberspace, where cyberspace itself refers to the set of links and relationships between objects that are accessible through a generalised telecommunications network, and to the set of objects themselves where they*

present interfaces allowing their remote control, remote access to data, or their participation in control actions within that Cyberspace" [33]. Since data is mentioned in this definition too, and Cybersecurity itself is considered an implementable measure into GDPR [37], it is often difficult to mark a separation line between Privacy and Cybersecurity. That is why we considered also some specific data protection nuances in Section IV-E. Cybersecurity is related to rights such as the Security of Person and Privacy.

The rights listed above are enshrined in the *Universal Declaration of Human Rights* (UDHR), the *International Covenant on Civil and Political Rights* (ICCPR), and the *European Convention on Human Rights* (ECHR). We have included these rights as they are directly relevant to the risks identified in the literature. However, it is essential to remember that human rights are indivisible, interrelated, and interdependent [60]. Therefore, additional rights may be compromised if, for instance, biases in Generative AI persist.

D. Search strategy

For our search strategy, two researchers meticulously identified and extracted official information from the selected sources, such as the companies' official websites, official blogs, privacy policy documents, and other relevant official channels. This task was done through the Google search engine (https://www.google.com/). We performed the search in private-browsing mode after logging out from personal accounts and erasing all web cookies and history [99]. By including diverse formats of official information, we guarantee that collected data is reliable, up-to-date, and aligned with the companies' official stance on bias reduction, privacy protection, cybersecurity, combating hate speech, and addressing misinformation.

E. Eligibility Criteria and Data Extraction

Initially, two authors independently conducted the initial search; overall, they came up with **192 sources**. The selected sources met these criteria:

1) They had to come from official websites and qualify as primary sources (communications, reports, statements, and policies coming directly from the company's official websites).
2) For policies, only the most recent available update was included.
3) The resource must address at least one of the chosen parameters

We point out that we primarily searched for documents in the field of Generative AI, and when there were no specific documents on Generative AI, we focused on AI documents in general.

Then, a third author conducted a secondary filter on the initial **192 sources** to identify those that contained relevant information about each parameter (*bias reduction*, *privacy protection*, *cybersecurity*, *fighting hate speech*, and *addressing misinformation*); so, **152 sources**, including blog posts, privacy policies, reports, and guidelines, were finally obtained. The overall quantitative results are shown in Table I.

TABLE I
DOCUMENTS SELECTED IN THE DATA EXTRACTION PHASE

Company	Resources analyzed	Resources selected
OpenAI	42	30
Google AI	60	49
Meta AI Research	43	39
Microsoft AI	47	34
Total	**192**	**152**

F. Data analysis and synthesis

The qualitative analysis involved a comprehensive review of **152** diverse textual sources, encompassing official statements, communications, websites, reports, policies and terms of use, scientific articles, and blog posts, among others, obtained from the companies' websites under examination. For this purpose, we designed an **evaluative framework** for the selected parameters (bias, privacy, hate speech, misinformation, and cybersecurity). For each parameter, the framework emphasizes some specific indicators. To provide an example, for the parameter **Bias**, the indicators are *diversity* (internal, as an objective, and in practice), the existence of a *specialized task force*, *multi-sectorial cooperation* such as with governments, NGOs, and institutions specialized in Human Rights, *research, innovation & investment*, *findings disclosure and open access to mitigation resources (tools /methods)* (to fortify the AI ecosystem), *Transparency in ongoing efforts*, and *Voluntary audits/reports*. There are more indicators within the rest of the parameters (privacy, cybersecurity, hate speech, and misinformation). The evaluative framework is available in the online appendix [13]. Based on this framework, we identified strengths and potential areas for enhancement using a four-evaluation scale: (i) strengths, (ii) weaknesses, (iii) areas for improvement, and (iv) no information. Through this process, we discerned and recommended best practices already adopted by analyzed companies and those advisable for future implementation. An excerpt and simplified representation of our evaluative framework is presented in Table II; the entire framework with all parameters and the remaining indicators, the tables with document analysis for each company (OpenAI, Google AI, Meta AI Research, and Microsoft AI), together with all the resources retrieved and analyzed, are available in the online appendix [13].

G. Thematic analysis

Once selected the sources, using *Atlas.ti* [11] — a software for qualitative text analysis — we conducted a thematic analysis [20] on documents containing the most valuable information for each parameter. We selected the most representative document for each company and each parameter. The reason for selecting only one document is to ensure comparability across companies. For some parameters, not all companies provide more than one document, therefore varying document lengths could influence the coding results. Selecting only one document for our thematic analysis follows this rationale. For instance, certain "hate speech" documents from one company are shorter than those from other companies,

TABLE II
EXTRACT FROM THE EVALUATIVE FRAMEWORK

Parameter	Indicator	Rationale	Scale	Assessment Questions
BIAS	Internal Diversity	It allows multiple perspectives to be incorporated, contributing positively to research and innovation. Internal diversity is critical for identifying biases based on knowledge of the unique realities of different communities.	- Strength, - Weakness, - Need for Improvement - No Information	1. Is the team's composition involved in developing AI systems diverse, including members from different racial, ethnic, gender, and socioeconomic backgrounds? 2. Does the company promote diversity and inclusion in its hiring practices and workplace culture? [...]
HATE SPEECH	Exploring model weaknesses	It involves extensive testing, validation, and adversarial analysis to detect issues such as susceptibility to attacks, biases, or over-fitting. Furthermore, research into model vulnerabilities enables continuous improvement and innovation within AI companies, driving advances in model construction and algorithmic methodologies.	- Strength, - Weakness, - Need for Improvement - No Information	1. Has the company conducted thorough testing, validation, and adversarial analysis to detect weaknesses in AI models related to hate speech? 2. Is there evidence of ongoing efforts to identify issues such as susceptibility to attacks, biases, or overfitting in AI models for addressing hate speech? [...]
MISINFORMATION	Contribution to Media Literacy	AI companies can improve users' media literacy by providing easily accessible information, training programs, and educational materials. This indicator aims to assess the degree to which AI companies encourage responsible media usage and assist individuals in navigating an ever-more information-centric society with confidence and proficiency.	- Strength, - Weakness, - Need for Improvement - No Information	1. Does the company provide educational materials to enhance users' understanding of misinformation and responsible use of AI models? 2. Are there initiatives to inform users about the importance of verifying information and the potential consequences of spreading misinformation? [...]

Note: *The comprehensive evaluative framework can be found in the online appendix [13].*

thus we opted to select documents of equivalent length for all four company parameters. Following this rationale and the eligibility criterion 3 outlined in Section III-E, and after reviewing all relevant documents pertaining to each parameter, we selected the ones that encapsulate the most significant efforts from the companies across the selected parameters.

As an example, in the case of "misinformation", we opted for the document titled *"Here's how we're using AI to help detect misinformation"* [65] for Meta, as it provides the most relevant insights into Meta's efforts against misinformation. Since not all companies feature a dedicated section addressing how they combat misinformation, we selected, for instance, *"Forecasting potential misuses of language models for disinformation campaigns and how to reduce risk"* [86] from the Research section for OpenAI. This document serves as a basis for evaluating OpenAI's efforts in this domain. The selected texts for each company are: *1)* OpenAI [88], [92], [93], [86] [28], *2)*; Google AI [48], [51], [29], [46], [48], *3)*; Meta AI Research [66], [67], [108], [112], [65]; and *4)* Microsoft AI [72], [63], [54], [53], [75], resulting in **20 documents**, including texts from official websites, privacy policies and other texts.

After employing Artificial Intelligence-based coding in *Atlas.ti* and reviewing and discarding codes that were either repetitive or did not provide significant information for content analysis (codes like *Artificial Intelligence*, *Technology*, *Intelligent Systems*, *Machine Learning*, etc.), we obtained **28 codes**. The complete list with all codes retrieved is available in the online appendix [13]. We omitted codes such as "Hate Speech," "Misinformation," "Bias," "Security," and "Privacy". Since we selected texts based on these parameters, these codes led to redundancy.

Figure 2 reveals that the most prominent themes overall are *"Ethics"*, *"Data Management"*, *"Representation: Diversity"*, and *"Fairness"*, which reflect the companies' commitment to the development of responsible and diverse AI, as indicated by the code *"Representation: Diversity"*.

Other noteworthy findings include *"Collaboration: Partnerships"*, *"Research"*, and *"Safety and Compliance"*, which hold significance, particularly when considering they come from companies' efforts in combating misinformation, hate speech, and bias.

Additional relevant codes encompass *"User-Centric design"*, *"Adaptability"*, *"Technology impact"*, *"Transparency"*, *"Accountability"*, and *"Technology: Education"*, implying a more people-centric approach and a focus on transparency and accountability.

Upon analyzing each company separately, it is possible to observe that the most recurrent themes for Google are *"Ethics"*, *"Data Management"* and *"Fairness"*, with *"Collaboration: Partnerships"*, *"Research"*, and *"User-centered design"* being of slightly lesser prominence. On the other hand, for OpenAI, the most prevalent theme is *"Data Management"*, followed by *"Ethics"* and *"Collaboration: Partnerships"*. Furthermore, *"Research"*, *"Accountability"*, *"Accuracy"*, and *"Transparency"* are also notable, albeit to a lesser extent.

Meta, on its part, places a greater emphasis on *"Fairness"*, *"Representation: Diversity"*, and *"Research"*, while *"DeepFakes Detection" "Adaptability"*, and *"Accountability"* hold slightly less weight.

Finally, Microsoft exhibits a distinct focus on *"Representation: Diversity"*, *"Safety and Compliance"*, *"Ethics"*, and *"Fairness"* with *"DeepFakes Detection"*, *"Accountability"*, and *"Transparency"* being comparatively less emphasized.

Overall, the thematic analysis conducted in this study reveals that Meta AI Research, OpenAI, Microsoft AI and Google AI prioritize *ethics*, *diversity*,*collaboration*, *data management*, and *research* as crucial elements of Generative AI development. The subsequent section provides more detailed insights into each company's measures, substantiating the concerns identified in this thematic analysis.

IV. COMPARATIVE ANALYSIS RESULTS

This section discusses the most important results of our analysis in terms of strengths, weaknesses, and best practices for *bias* reduction, *privacy* protection, *cybersecurity*, fighting *hate speech*, and addressing *misinformation* by the four

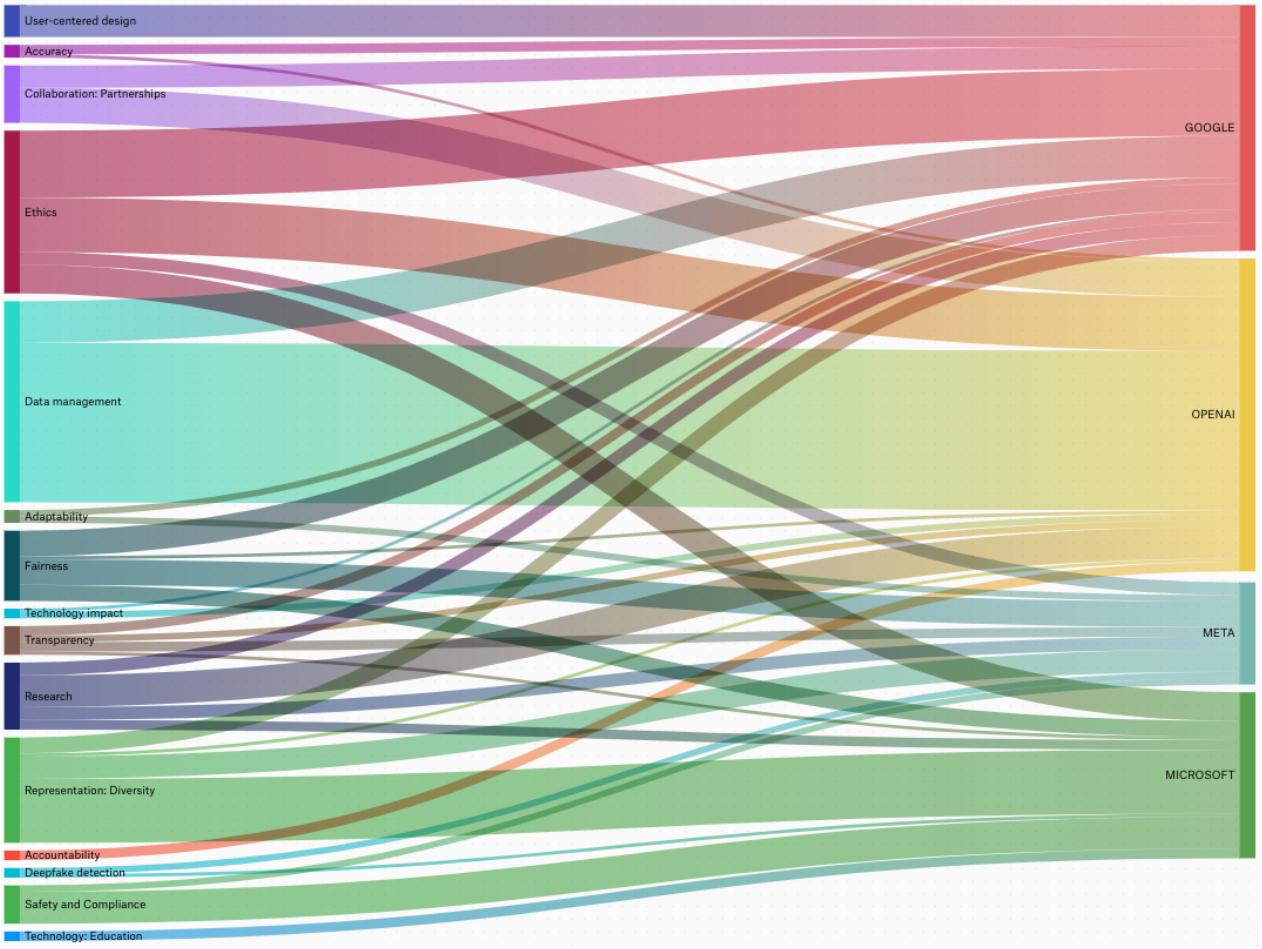

Fig. 2. Thematic analysis: Sankey diagram of the most recurrent codes

analyzed companies, based on our evaluative framework, its indicators and scales. All the analysis details, together with a more detailed discussion, can be found in the online appendix [13].

A. Bias Mitigation

Strengths: In the current landscape, the selected AI companies are handling bias through various approaches. Internally, they seek representation within their teams, particularly among staff dedicated to addressing these issues. For instance, OpenAI maintains a team of reviewers from different backgrounds who follow specific guidelines [88], [95]. Similarly, Microsoft AI has established a team dedicated to *Fairness, Accountability, Transparency, and Ethics* (FATE) in Artificial Intelligence. This group is tasked with researching the social implications of artificial intelligence to foster accountability across various AI models [40]. Google AI, for its part, has committed to increasing its internal representation ratios. This incorporates a specific objective of tripling the leadership representation percentage from 30% to 60% by 2025 [47].

On the other hand, these companies also engage in collaborative projects with academia and experts to enhance methodologies and tools for bias mitigation. For instance, Meta AI Research has established a multidisciplinary team that has pioneered a novel approach for assessing a broad spectrum of biases in NLP models [66]. This approach goes beyond considerations of race, gender, and ethnicity and is referred to as the *"demographic text perturber"*.

On the research and development (R&D) front, OpenAI has also considered *"Constitutional AI"*, which involves using AI to oversee other AIs. The process does not include human labeling, which has a risk of human-induced biases but instead employs a list of principles and norms that should help in creating a safer model [12]. Microsoft AI has conducted a study titled *"Performance and Bias in Human-AI Teamwork in Hiring"*, investigating the collaborative performance and potential biases that may arise when humans work alongside AI in the hiring process [97]. Meta AI Research has also created datasets and other sources related to the "*Improving methods for measuring the fairness of AI systems*" project [96], which are available on their GitHub repository[1].

Finally, some companies test their models to identify the probability of generating bias. Google AI, for example, performs exhaustive adversarial testing of final AI systems to detect and correct unfair outcomes. Notably, one of their research outcomes involves the development of the regularization technique *MinDiff for ML Fairness* [101].

[1]https://github.com/facebookresearch/ResponsibleNLP

Other notable efforts include undergoing a Civil Rights Audit, as seen in the case of Meta AI Research in the United States, which demonstrates their commitment to gaining a deeper understanding of bias and advancing their services towards equity [6].

Weakness: Despite the notable efforts by the four companies, there is a prevailing presence of Western perspectives in AI research and tools across all of them. Additionally, companies like OpenAI have acknowledged that implementing and testing bias mitigation measures has primarily focused on English content [89]. In other cases, such as Microsoft AI, it is unclear how they address bias in Generative AI, including whether there is a diverse group of reviewers and if they adhere to specific guidelines. Although there is a brief mention of a strategy for mitigating bias in Generative AI in the Microsoft's blog [18]. The above remains a recurring challenge in this and other parameters, especially for companies offering a diverse array of services. Furthermore, it is imperative to further enhance and promote academic collaboration in research embracing a multidisciplinary and geographically varied approach.

Best practice: *1)* Actively engage with non-governmental organizations and AI ethics and human rights research centres from diverse backgrounds to enhance the approach to bias mitigation. *2)* The research team, research lines, and tools in this field should encompass geographic, linguistic, social, and cultural diversity.

These recommendations draw inspiration from the practices of leading organizations such as OpenAI, which works to ensure diversity in its data labelling team, Microsoft AI's establishment of a specialized team on this matter, FATE, and Google AI's commitment to increasing representation indices within its internal team.

B. Addressing Misinformation

Strengths: In terms of combating misinformation, the AI industry collaborates with experts to deepen their understanding and better address it. For example, OpenAI's team has explored the potential applications of Generative AI in generating misinformation and proposed strategies to counteract it. Its article, *"Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations"* delves into the dimensions, actors, and shifts in the dynamics of disinformation stemming from the use of Generative AI [44]. Microsoft AI, on its part, has conducted research on disinformation campaigns in the United States, France, Poland, Germany, and Ukraine across various social media platforms.

The advancement of technology is also a vital component. Microsoft AI has made significant financial investments in technology such as the *Video Authenticator*, which is expressly aimed at identifying potentially misleading images and videos online, particularly deepfakes [53]. In addition, Microsoft AI has implemented supplementary technologies to authenticate the authenticity of news material [24]. Google AI has also introduced *Microsoft AccountGuard*, a dedicated cybersecurity tool tailored for political leaders, political parties, non-governmental organizations, and think tanks using *Office 365*. Each of these initiatives constitutes a crucial component of their *Defending Democracy program* [24].

Companies such as Meta AI and Google AI have a comprehensive strategy that includes a wide range of projects: their services include social media, news platforms, search engines, and other related platforms. Meta AI has created dedicated hubs that center around specific subjects, such as the *COVID-19 Information Centre*, the *Climate Science Information Centre*, and the *US Voting 2020 information* [107].

Regarding technological development, Meta AI has also organized the *Kaggle DeepFake Detection* competition, leading to the creation of the *DeepFake Detection Challenge* dataset. This dataset is currently the most extensive publicly available collection of face-swapping videos, comprising over 100,000 clips [32]. Meanwhile, Google AI and YouTube actively explore using artificial intelligence to detect synthetic content in real time. This project involves specialists from all over the world [46]. Google AI is also attuned to the impact of synthetic content, especially in critical contexts like elections. To address this, they have initiated projects like Jigsaw, which assesses risks to open societies and develops technology for scalable solutions [46].

Companies are also actively collaborating with governments. For instance, Google AI is extensively cooperating with the European Union to counter disinformation, particularly in regions like Central and Eastern Europe [61]. This collaborative effort involves policy adjustments, the enhancement of product features, support for fact-checkers, and engagement with academics from around the globe. Additionally, Google AI has played a role in strengthening the *"EU Code of Practice on Disinformation"* [61].

Weakness: Most of the anti-misinformation and disinformation efforts are concentrated in Europe or the global north. Additionally, it is crucial to involve as well non-governmental organizations and other pertinent entities in the study of the impact and potential of disinformation. Despite significant strides in researching disinformation campaigns in diverse countries, there is a pressing need for greater collaboration with institutions and academics from other nations. Given that countering misinformation and disinformation demands an understanding of the local, cultural, and social contexts in which false narratives spread, diversity becomes an indispensable prerequisite.

Best Practice: *1)* Collaborate with academia and specialized centres to deepen understanding of the scope and consequences of misinformation, involving experts from diverse regions, *2)* Expand collaborative networks for election security and false content identification across several locations, especially in unstable contexts and during critical periods, *3)* Produce reports on work in the field of addressing misinformation, *4)* Generate comprehensive reports on the progress made in combating misinformation, *5)* Promote transparency by publishing reports on content removal requests, *6)* Promote media literacy to raise awareness among the audience about the capabilities of Generative AI and their potentially harmful effects in terms of misinformation.

Some of these measures are based on the work of companies such as Google AI and their collaboration with the Jigsaw

project, Meta AI Research's contributions to the AI community with the *DeepFake Detection Challenge* dataset, OpenAI's collaboration with academia, and Microsoft's global focus on combating Misinformation

C. Tackling Hate Speech

Strengths: Regarding the fight against hate speech, the AI Industry's efforts encompass producing reports on the likelihood of their models generating hate speech content. For instance, OpenAI's *GPT-4 Technical Report* [87] explicitly recognizes particular content can harm marginalized communities, engender hostile online environments, and, in extreme cases, escalate into actual violence and discrimination. The report also provides additional information regarding the team's efforts to resolve the potentially harmful applications of their models [87].

On the other hand, efforts also involve exploring how Generative AI can be employed to combat hate speech. For example, OpenAI researches how GPT-3 can be utilized for hate speech identification [28]. Through these efforts, the OpenAI team observed that the model demonstrates a precision ranging from 48% to 69% in detecting racist and sexist comments. Furthermore, with a few-shot learning approach and instruction in the message, the model's accuracy can reach up to 78% [28]. Additionally, Microsoft AI has developed *(De)ToxiGen*, a dataset tailored for training content moderation tools to detect implicitly harmful language more effectively [52]. *ToxiGen* as an algorithmic mechanism that creates adversarial situations between a Generative AI and a specific content moderation tool [54]. This process actively involves the content classifier, addressing the unique vulnerabilities that each moderation tool may exhibit based on the type of data it was trained on.

Furthermore, organizations like Microsoft AI assign their AI Red Team also to address the issue of hate speech [62]. This team not only concentrates on identifying security vulnerabilities but also on identifying potentially harmful content, rendering it a crucial component for Responsible AI [62].

In the realm of innovation, Meta AI Research has pioneered a new field of learning called *"Reinforced Integrity Optimizer (RIO)"*, which directly utilizes real-world online data from their production systems to optimize the AI models that detect hate speech instead of relying on a static, offline dataset [65]. *RIO* operates throughout the entire machine learning development lifecycle, from data sampling to A/B testing [65].

Cooperative initiatives are also a crucial element in countering hate speech. Google AI, Meta AI Research, Microsoft AI, Twitter, and the European Commission have collaborated to tackle this problem in Europe [58].

Overall, companies like Google AI and Meta AI Research have a broader scope in this regard, as they have dealt with hate speech across their multiple services, which has afforded them invaluable experience.

Weakness: Among the identified weaknesses in this area is the need for greater collaboration with human rights institutions and ethical AI-focused organizations to take a comprehensive stance on the issue. Additionally, companies must publish content removal request reports to promote accountability and transparency.

In some cases, given the diverse nature of the services offered, information on how companies handle hate speech primarily pertains to their social media, search engines, or news platforms, as well as the use of artificial intelligence for its detection. Therefore, there is a need for greater precision regarding the efforts to tackle hate speech on their Generative-AI-based services. Additionally, diversification of research lines and collaborative partnerships is required.

Best practices: *1)* Actively conduct reviews of Generative-AI-based services to assess the likelihood of hate speech generation, *2)* Collaborate with specialized institutions and research centers in ethics and human rights to comprehensively capture the complex dimensions of hate speech, *3)* Share research findings with the AI community, with particular support for startups and less experienced companies in this domain, *4)* Collaborate with organizations and governments worldwide to counteract hate speech.

These recommendations are based on the efforts of companies such as Google AI and its AI Red Team efforts to evaluate models beyond security concerns, as well as Meta AI Research and OpenAI, who actively collaborate and share their research findings.

D. Protecting Privacy

Strengths: Regarding privacy protection, some companies inform users that their interactions with Generative-AI-based services do not contribute to the training of their models. This practice was observed at OpenAI and Microsoft AI [110] [77].

Another area of action is educating and assisting users in understanding the importance of privacy protection and how to exercise their data protection rights. For instance, OpenAI provides instructions in the *"Advice and Answers from the OpenAI Team"* [57] section on how users can effectively handle the data produced from their interactions with the models. In addition, the company offers comprehensive tutorials for its web interface and iOS and Android apps [57]. The website also provides additional information regarding OpenAI's data management methods [57], and the *Security Portal* [94] offers thorough details about the company's adherence to diverse data privacy standards.

Some companies have special portals/centers to facilitate user control over their data, such as *Microsoft's Privacy Dashboard* [78]. Additionally, Microsoft AI shares its certifications, regulations, and compliance standards, including *ISO 22301*, a management standard that outlines requirements for business continuity management systems. It aims to safeguard against and respond to disruptive incidents [56]; SOC, which is a set of audit reports for service organizations, validating internal controls over provided information systems [10], and *General Data Protection Regulation* (GDPR), an extensive data privacy regulation established by the European Union, designed to safeguard the personal data of EU citizens and residents [37], via its *Service Trust Portal* [76].

Meta AI Research has formed a specialized team called the *"Red Privacy Team"* to focus specifically on privacy-related concerns [68]. This team actively collaborates and consults with policymakers and domain specialists. Furthermore, the

team partakes in various collaborative initiatives such as workshops, in-depth discussions, financial assistance negotiations, and participation in conferences [68].

In the realm of research, Microsoft AI has developed proposals such as meta-frameworks for differentially private fine-tuning of large-scale pre-trained language models [126]. Their experiments have demonstrated that differentially private adaptations of these approaches outperform private algorithms in three key dimensions: utility, privacy, and computational and memory costs of private training [126].

On the other hand, Meta AI Research has instituted a rigorous protocol and established a robust cryptographic framework, in collaboration with entities like *Osis Labs*, to advance *Secure Multi-Party Computation* (SMPC) methodologies [8]. This cryptographic component operates within a specified field, enabling a comprehensive analysis of encrypted data [8].

Google AI, on its part, has spearheaded the *Randomized Aggregatable Privacy-Preserving Ordinal Response* (RAPPOR) project [34] and developed a People + AI Guidebook in 2018 [45], that is constantly updated and provides guidance for data security and privacy [45]. The guidebook resulted from the collaboration of a multidisciplinary group for the design of a human-centred AI [49]. By its part RAPPOR introduces an innovative approach to gather software statistics, prioritizing user security and error identification, and enhancing the overall user experience by implementing differential privacy [34]. Moreover, Google AI leverages open-source frameworks like *TensorFlow*, making substantial strides in safeguarding privacy [115]. *TensorFlow* is engineered to be compatible with federated learning, enabling developers to oversee dataset access and usage throughout the training and inference processes [115].

Weakness: One notable weakness in the practices of companies like Microsoft AI, Google AI, and Meta AI Research, as well as others in the AI industry, is the need for more explicit communication with consumers regarding their potential interactions with Generative AI when using their diverse services. Additionally, there needs to be more clarity in providing relevant privacy regulations applicable to their Generative-AI-based services. Privacy policies and data management tools should also be made easily accessible to all segments of the population.

In some companies that offer multiple services, it is challenging to locate policies and methods for exercising data control concerning Generative-AI-based services on their data management centers/portals. In addition, organizations such as OpenAI, Google AI, and Microsoft AI must invest in data protection educational resources, which could be delivered in a variety of media formats, such as videos, audio content, infographics, and informative training modules. Sometimes users are not even aware that companies have tools to facilitate control over their data, therefore such tools should be publicized more.

Best Practices: *1)* Promote privacy awareness among the general public by disseminating educational materials on privacy. *2)* Foster ongoing collaboration with diverse academia, research centers, organizations, and pertinent government entities in this field. *3)* Ensure transparent communication regarding interactions with Generative AI, and *4)* Provide users with accessible and straightforward information on applicable privacy policies, along with mechanisms for exercising control over their data.

Some of these measures are inspired by the efforts of Google and its commitment towards multidisciplinary in research [45]. OpenAI and Meta AI Research, show great progress in this area. Meta AI has established a Privacy Team dedicated to addressing privacy concerns. Furthermore, Meta AI Research actively engages in various activities, including workshops, talks, and conferences on privacy. Additionally, Meta AI Research stands out for providing straightforward, user-friendly, and easily understandable audiovisual content on this matter for its audience [2].

E. Cybersecurity

Strengths: Regarding cybersecurity, some companies have Transparency Centers/Portals where they share information about certifications and security measures. For instance, OpenAI has a *Security Portal* that offers comprehensive details about the company's adherence to various regulatory standards in cybersecurity [94]. These standards include (i) The *California Consumer Privacy Act* (CCPA) [23]; (ii) The GDPR; (iii) The *Service Organization Control 2* (SOC2) [9]; and (iv) The *Service Organization Control 3* (SOC3)[94].

On the other hand, a notable strength of Google AI lies in its *Secure AI Framework* (SAIF), which embodies an open and collaborative approach to cybersecurity [51]. *SAIF* derives its strength from Google AI's extensive expertise in cybersecurity, so it is a valuable resource that other companies can adopt.

Google AI has two robust components in the cybersecurity field: the *Mandiant* and *TAG teams*. These teams handle cyber activity associated with AI systems and consistently address global cyber threats [51]. Google AI includes an active Red Team as well [38].

On the other hand, OpenAI has opened a grant through its cybersecurity program. This grant is intended to improve and assess the effectiveness of AI-driven cybersecurity capabilities, while also advancing the discourse surrounding advanced cybersecurity measures [109]. Additionally, OpenAI is presently establishing an external interdisciplinary Red Team [91]. This collaborative initiative involves working with individual experts, academic institutions, and civil society organizations to bolster the security of its models [91].

In line with this, Microsoft AI offers guidance for forming Red Teams to assess Generative-AI-based models [72], and Microsoft AI has also collaborated with industry peers and academics to release the *"Adversarial Machine Learning Threat Matrix"* [62] contributing further to collective efforts in this domain. On the other hand, *Microsoft's Cyber Defense Operations Center* is equipped with state-of-the-art technology and staffed by experts from around the globe [74]. Furthermore, Microsoft AI operates a *Digital Crimes Unit* that collaborates with security institutions worldwide [71].

In the sphere of research, Meta AI Research prioritises transparency by sharing its research findings on global security issues with the AI community, thereby making a significant

contribution [69]. In addition, Meta AI Research publishes regular integrity reports that provide information on its efforts to combat cyberthreats [69]. These reports include a variety of actions, such as counter-malware campaigns and strategies to combat global adversarial networks. This also includes influence operations and cyber espionage [108].

Weaknesses: Given that there is insufficient publicly available information to assess the degree to which companies comply with our Cybersecurity Indicators (available in the online appendix [13]), the only weakness to report is the predominance of a Western perspective on tools and research in cybersecurity.

Note on Cybersecurity evaluation It is important to note that the analysis of cybersecurity measures in this study is limited to publicly available information on the criteria and policies in place at each company, which makes identifying weaknesses challenging. Although the degree of information accessibility may not entirely reflect the robustness of cybersecurity measures, it helps to improve the public's perception of a company's preparedness. We understand that due to the sensitive nature of this information, some data must remain confidential; however, companies like OpenAI provide access to reports, audits, and analyses upon request, and Google AI shares best practices based on its extensive experience. Therefore, openness and transparency improve the perception of a company's cybersecurity strength.

Best practice: *1)* Prioritize accessibility of data control centers, with a particular focus on ensuring inclusivity for individuals with disabilities. *2)* Sustain partnerships with research institutions, regulatory bodies, and emerging startups and continue sharing lessons learned, findings, and datasets to fortify the cybersecurity of the AI ecosystem. *3)* Whenever feasible, proactively disseminate information regarding cybersecurity compliance certifications. *4)* Geographically, linguistically and socially diversify research and tool development efforts in cybersecurity.

This recommendation draws inspiration from the practices of various leading companies. For instance OpenAI *Security Portal* [94] offers comprehensive information on certifications, internal policies, infrastructure, updates, and cybersecurity ratings. Microsoft AI guidance on establishing Red Teams for evaluating Generative AI, and Meta AI Research distinguishes itself through its contribution to a secure AI ecosystem, exemplified by its collaborative research projects and sharing research findings. Finally, Google AI has also distinguished itself in this field by developing SAIF.

F. On The Importance of Good Practices for Responsible AI

Addressing the challenges of bias reduction, privacy protection, cybersecurity, and fighting hate speech and misinformation is critical for AI companies. It not only ensures responsible industry practices but also prevents, addresses, and mitigates the potential impacts of these models on human rights.

Companies like Microsoft AI, OpenAI, Google AI, and Meta AI Research bear substantial responsibility for societal well-being and progress. Each of their actions or inactions can have highly detrimental effects on specific populations, particularly vulnerable ones. Consider, for instance, the proliferation of hate speech in unstable contexts; under these conditions, the use of these models could compromise the safety and integrity of hundreds of people.

Nevertheless, AI companies are not solely accountable for addressing the multiple aspects of this challenge. Regulators, decision-makers, academia, non-governmental organizations, media, and society all share the duty to guide, nurture, and reinforce a Responsible AI ecosystem.

V. Discussion of Results

RQ1. What Responsible AI practices, in terms of *risk mitigation and human rights protection strategies/actions*, do leading AI companies like OpenAI, Google, Microsoft, and Meta employ when developing AI models, and in particular Generative AI models? The measures and efforts related to human rights are strongly influenced by the diversity of services and industry experience in AI. For instance, companies like Google AI and Meta AI Research stand out in combating misinformation due to the array of products they offer, such as social media platforms and free news aggregator services. This grants them influence but also entails a significant responsibility in addressing issues like misinformation and hate speech. On the other hand, OpenAI, by providing Generative-AI-based services and having a narrower range of products, can contribute to protecting human rights, such as privacy, through the research of novel techniques for privacy preservation. Meanwhile, Microsoft AI's experience and services enable it to bolster cybersecurity techniques, particularly valuable in contexts like elections to protect the systems from hacking.

Furthermore, OpenAI and Microsoft AI, for instance, exhibit a proactive involvement in pioneering research, forging partnerships with leading institutes and research centres to push the boundaries of AI capabilities. In contrast, Google AI and Meta AI Research place a pronounced emphasis on fostering collaboration with governmental bodies.

RQ2. What are companies' *strengths*, and *weaknesses* in risk mitigation and human rights protection actions? Strengths include in certain cases, accessibility to information certifications compliance, access to findings from various research endeavors, as well as databases, and the ability to exercise control over personal data.

Additionally, there is a growing awareness of the need for interdisciplinary collaboration, which will undoubtedly contribute to fortifying protection mechanisms. Another strength lies in their active and continuous engagement in research and development, a growing trend towards investigating the social impact of technologies, as well as the creation of specialized teams in collaboration with experts in the field of human rights.

However, there are notable weaknesses to address. One significant area is the limited cultural, social, and linguistic diversity. This deficit results in a potential bias towards Western perspectives in research lines and tools for addressing bias, privacy, and cybersecurity, combating hate speech, and countering misinformation. Furthermore, it is vital to broaden

and vary the collaboration with organizations, institutions, and governments across various regions of the globe. On the other hand, it is also necessary for data control tools to be accessible to all population sectors, especially individuals with disabilities.

In a broader sense, there is a lack of engagement with non-governmental organizations, civil associations, and institutes dedicated to AI ethics and human rights protection.

RQ3. Which best practices can be adopted to enhance risk mitigation and human rights safeguard within the AI industry, with particular reference to Generative AI development?

Based on the findings of this research, it is recommended:

1) Actively engage with non-governmental organizations and AI ethics and human rights research centers from diverse backgrounds to enhance the approach to bias mitigation.
2) The research team, research lines, and tools on bias should encompass geographic, linguistic, social, and cultural diversity.
3) Collaborate with academia and specialized centres to deepen understanding of the scope and consequences of misinformation, involving experts from diverse regions.
4) Expand collaborative networks for election security and false content identification across several locations, especially in unstable contexts and during critical periods.
5) Generate reports outlining efforts in the field of combating misinformation.
6) Promote transparency by publishing reports on content removal requests.
7) Promote media literacy to raise awareness among the audience about the capabilities of Generative AI and their potentially harmful effects in terms of misinformation.
8) Actively conduct reviews of Generative AI to assess the likelihood of hate speech generation.
9) Collaborate with specialized institutions and research centres in ethics and human rights to comprehensively capture the complex dimensions of hate speech.
10) Share research findings with the AI community, with particular support for startups and less experienced companies on hate speech.
11) Collaborate with organizations and governments worldwide to counteract hate speech.
12) Promote privacy awareness among the general public by disseminating educational materials on privacy.
13) Foster ongoing collaboration with diverse academia, research centres, organizations, and pertinent government entities on privacy protection.
14) Ensure transparent communication regarding interactions with Generative AI models.
15) Provide users with accessible and straightforward information on applicable privacy policies, along with mechanisms for exercising control over their data.
16) Prioritize accessibility of data control centres with a particular focus on ensuring inclusivity for individuals with disabilities.
17) Sustain partnerships with research institutions, regulatory bodies, and emerging startups and continue sharing lessons learned, findings, and datasets to fortify the cybersecurity of the AI ecosystem.
18) Whenever feasible, proactively disseminate information regarding cybersecurity compliance certifications.

VI. Threats to validity

In this section, we deal with the threats to the validity [123] of our study. Threats to validity refer to factors that may compromise the accuracy and reliability of the study's findings.

Construct validity. The entire comparative analysis was executed manually by the authors, introducing the possibility of subjective judgment. In order to address this concern, we followed the negotiated agreement technique [114] between two of the authors, which was achieved after a careful examination of a number of comments.

Internal validity. Our analysis was restricted to the publicly available documents, as we did not have access to the internal documents and policies that each company formulated. We point out that we did not contact companies to request additional documents. Nevertheless, this threat may be mitigated by recognizing that even if these internal documents existed, they would remain inaccessible to the general public. It is crucial to clarify that the lack of access to this information does not imply that the organization is not progressing in that direction. Nonetheless, it does result in a lack of awareness among the general public regarding the policies and actions the company is implementing in the domain of Responsible AI. Instead, these initiatives should be publicized, as they contribute to promoting awareness among users in the various areas under examination.

Generalizability – Transferability. We know the companies' sample size may not be sufficient or may be skewed in a way that affects the generalizability of the results. Anyway, we address this threat by conducting this study on four companies that, at the present moment, are among the most active players in the field of Generative AI. We are aware that some best practices can be applied with difficulty by small companies, e.g. startups. Therefore, we plan in the future to broaden the analysis on small-medium size companies as well.

VII. Conclusion and Future Work

In this work, we have investigated how the pursuit of human rights might shape responsible AI development and deployment. We conducted our analysis on five parameters, which we mapped with their respective human rights, with the aim of understanding what actions four leading AI companies are putting in place in these contexts. Then, we recommend a set of best practices that can be applied in the AI industry to foster Responsible AI in the light of the human rights perspective.

Based on our investigation, we have found that certain organizations demonstrate exceptional performance in many areas, such as offering easily available security reports, empowering users to manage their data, promoting multidisciplinary collaboration, establishing expert groups to research the social impact of their technologies, prioritizing R&D of technologies

and methodologies for privacy preservation and bias reduction and sharing research findings, and information with the wider community. However, there is still a long way to go regarding linguistic, cultural, and geographic diversity in research lines, tools, and collaborative efforts.

AI companies could enhance the understanding of their impact on human rights and at the same time strengthen their efforts towards responsible and trustworthy AI by collaboratively working with diverse, renowned organizations and research centers dedicated to AI ethics, responsible AI, and digital rights (such as the Center for Responsible AI [25], Institute for Ethics in Artificial Intelligence [55], CLAIRE [27], Responsible AI Institute [103], The Alan Turing Institute [116], Allen Institute for Artificial Intelligence [117], among others).

Currently, the AI industry must be mindful of its commitment to society and the protection of freedoms and rights, as the use of these models is likely to increase significantly in the future, requiring constant cooperation for a reflexive, collaborative, and responsible technological advancement.

Future research may expand our human rights analysis to include additional parameters, such as algorithmic transparency and liability for damages caused, among others, and a diverse set of companies. In addition, as an attempt to promote diversity in this field, we plan to explore the approaches of AI companies outside the Western context.

VIII. Acknowledgements

This study has been partially supported by the following projects: SSA (Secure Safe Apulia - Regional Security Center, Codice Progetto 6ESURE5) and KEIRETSU (Codice Progetto V9UFIL5) funded by "Regolamento regionale della Puglia per gli aiuti in esenzione n. 17 del 30/09/2014 (BURP n. 139 suppl. del 06/10/2014) TITOLO II CAPO 1 DEL REGOLAMENTO GENERALE "Avviso per la presentazione dei progetti promossi da Grandi Imprese ai sensi dell'articolo 17 del Regolamento"; SERICS (PE00000014) under the MUR National Recovery and Resilience Plan funded by the European Union – NextGenerationEU.

References

[1] Legal and human rights issues of AI: Gaps, challenges and vulnerabilities. Journal of Responsible Technology, 4:100005, 2020.

[2] Facebook privacy policy, 2023. https://www.facebook.com/privacy/policy/.

[3] Access Now. Artificial intelligence, 2023. https://www.accessnow.org/issue/artificial-intelligence/.

[4] AI Now Institute. Home, 2023. https://ainowinstitute.org/.

[5] E. Aizenberg and J. van den Hoven. Designing for human rights in ai. Big Data & Society, 7(2), 2020.

[6] R. Alao, M. Bogen, J. Miao, I. Mironov, and T. Jonathan. How meta is working to assess fairness in relation to race in the u.s. across its products and systems, 2021. https://ai.meta.com/research/publications/\how-meta-is-working-to-assess-fairness-in-relation-to-race-in-the-us\-across-its-products-and-systems/.

[7] Amnesty International. Amnesty international, 2023. https://www.amnesty.org/en/.

[8] E. Arcaute and R. Austin. Assessing fairness of our products while protecting people's privacy, 2022. https://ai.meta.com/blog/assessing-fairness-of-our-products-while-protecting-peoples-privacy/.

[9] Association of International Certified Professional Accountants. SOC 2® - SOC for Service Organizations: Trust Services Criteria, n.d. Retrieved October 5, 2023.

[10] Association of International Certified Professional Accountants. System and organization controls: SOC suite of services, n.d.

[11] ATLAS.ti. ATLAS.ti, 2023. https://atlasti.com.

[12] Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional ai: Harmlessness from ai feedback. 2022.

[13] M. T. Baldassarre, D. Caivano, D. Gigante, B. F. Nieto, and A. Ragone. Online appendix, 2023. https://figshare.com/s/57ef5620007720ce5657.

[14] Maria Teresa Baldassarre, Danilo Caivano, Belen Fernandez Nieto, Davide Gigante, and Azzurra Ragone. The social impact of generative ai: An analysis on chatgpt. In Proceedings of the 2023 ACM Conference on Information Technology for Social Good, pages 363–373. CM Conference on Information Technology for Social Good, 2023.

[15] Teresa Baldassarre, Nicola Boffoli, Danilo Caivano, and Giuseppe Visaggio. Managing software process improvement (spi) through statistical process control (spc). Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 3009:30 – 46, 2004.

[16] Vita Santa Barletta, Danilo Caivano, Domenico Gigante, and Azzurra Ragone. A rapid review of responsible ai frameworks: How to guide the development of ethical ai. EASE '23, page 358–367, New York, NY, USA, 2023. Association for Computing Machinery.

[17] Solon Barocas and Andrew D. Selbst. Big data's disparate impact. California Law Review, 104:671, 2016.

[18] S. Bird. Deploy large language models responsibly with azure ai, 2023. https://techcommunity.microsoft.com/t5/ai-machine-learning-blog/deploy-large-language-models-responsibly-with-azure-ai/ba-p/3876792.

[19] Robert Boyd and Robert J. Holton. Technology, innovation, employment and power: Does robotics and artificial intelligence really mean social transformation? Journal of Sociology, 54(3):331–345, 2018.

[20] Virginia Braun and Victoria Clarke. Thematic analysis., pages 57–71. 01 2012.

[21] Miles Brundage. Artificial intelligence and responsible innovation. In Vincent C. Müller, editor, Fundamental Issues of Artificial Intelligence, pages 543–554. Springer International Publishing, 2016.

[22] Nicolas Bueno and Claire Bright. Implementing human rights due diligence through corporate civil liability. Social Science Research Network, 2020.

[23] California State Legislature. California consumer privacy act (ccpa), 2018. https://oag.ca.gov/privacy/ccpa.

[24] CEE Multi-Country News Center. Securing democracy in the digital age, 2021.

[25] Center for Responsible AI. We believe in fair, explainable and sustainable ai, 2023. https://centerforresponsible.ai/.

[26] CEPS and et al. Study to support an impact assessment of regulatory requirements for artificial intelligence in europe: Final report. Technical report, Publications Office of the European Union, 2021.

[27] CLAIRE. Claire - confederation of laboratories for artificial intelligence research in europe, 2020. https://claire-ai.org/.

[28] Andrew Collins and Ryan Alexander. Gpt3-hate-speech-detection, 2022. https://github.com/kelichiu/GPT3-hate-speech-detection.

[29] E. Collins and Z. Ghahramani. Lamda: our breakthrough conversation technology, 2021. https://blog.google/technology/ai/lamda/.

[30] J. Curzon, T. A. Kosa, R. Akalu, and K. El-Khatib. Privacy and artificial intelligence. IEEE Transactions on Artificial Intelligence, 2(2):96–108, 2021.

[31] P. de Laat. Companies committed to responsible ai: From principles towards implementation and regulation? Philosophy & Technology, 34:1135–1193, 2021.

[32] Brian Dolhansky, Joel Bitton, Brett Pflaum, Jing Lu, Ryan Howes, Meng Wang, and Chiori Ferrer. The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397, 2020.

[33] ENISA. Definition of cybersecurity - gaps and overlaps in standardisation, 2016-07-01. https://www.enisa.europa.eu/publications/definition-of-cybersecurity.

[34] Úlfar Erlingsson. Learning statistics with privacy, aided by the flip of a coin, 2014.
[35] European Center for Not-for-Profit Law Stichting. Home, 2023. https://ecnl.org/.
[36] European Commission. Ethics guidelines for trustworthy ai, 2019.
[37] European Parliament and Council of the European Union. Regulation (EU) 2016/679 of the European Parliament and of the Council, 2016-05-04.
[38] D. Fabian. Google's ai red team: the ethical hackers making ai safer, 2023. https://blog.google/technology/safety-security/googles-ai-red-team-the-ethical-hackers-making-ai-safer/.
[39] Don Fallis. What is disinformation? Library Trends, 63(3):401–426, 2015.
[40] FATE: Fairness, Accountability, Transparency & Ethics in AI. Microsoft research, n.d. https://www.microsoft.com/en-us/research/theme/fate/.
[41] S. Feldstein. The global expansion of ai surveillance, 2019.
[42] Irene Gabriel, Margaret Mitchell, Timnit Gebru, and Vinod Prabhakaran. A human rights approach to responsible ai, 2022.
[43] Jean I. Garcia-Gathright, Aaron Springer, and Henriette Cramer. Assessing and addressing algorithmic bias - but before we get there. ArXiv, abs/1809.03332, 2018.
[44] Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. Generative language models and automated influence operations: Emerging threats and potential mitigations, 2023.
[45] Google. People + AI Guidebook. https://design.google/ai-guidebook. Accedido 4 de marzo de 2024.
[46] Google. How google fights disinformation, 2019. https://blog.google/documents/37/How_Google_Fights_Disinformation.pdf/.
[47] Google. Building a sense of belonging at google and beyond, n.d. https://about.google/belonging/.
[48] Google. Google responsible AI practices, no date. https://ai.google/responsibility/responsible-ai-practices/.
[49] People + Google. Updating the people + ai guidebook in the age of generative ai, noviembre 2023.
[50] Google AI. Research, n.d. https://ai.google/discover/research/.
[51] R. Hansen and P. Venables. Introducing google's secure ai framework, 2023. https://blog.google/technology/safety-security/introducing-googles-secure-ai-framework/.
[52] Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Debanjan Ray, and Ece Kamar. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. 2022.
[53] Eric Horvitz. New steps to combat disinformation, 2020.
[54] Adam Hughes. (de)toxigen: Leveraging large language models to build more robust hate speech detection tools. 2022.
[55] Institute for Ethics in Artificial Intelligence. Ai ethics, n.d. https://www.ieai.sot.tum.de/.
[56] ISO. Iso 22301:2019. https://www.iso.org/standard/75106.html, June 5 2023.
[57] J. J. Data controls faq — openai help center, 2023. https://help.openai.com/en/articles/7730893-data-controls-faq.
[58] L. Junius. Our commitment to fighting illegal hate speech online, 2016. https://blog.google/around-the-globe/google-europe/our-commitment-to-fighting-illegal-hate_39/.
[59] Fernando Kamei, Igo Wiese, Cristine Lima, Ivano Polato, Victor Nepomuceno, Wesley Ferreira, Marcio Ribeiro, Carlos Pena, Bruno Cartaxo, Gustavo Pinto, and Sérgio Soares. Grey literature in software engineering: A critical review. Information and Software Technology, 138:106609, 2021.
[60] Ingeborg Elisabeth Koch. Human Rights as Indivisible Rights: The Protection of Socio-economic Demands under the European Convention on Human Rights. Martinus Nijhoff Publishers, 2009.
[61] A. Kroeber-Riel. Our latest commitments to countering disinformation in central and eastern europe, May 4 2023. https://blog.google/around-the-globe/google-europe/latest-disinformation-commitments-in-cee/.
[62] R. S. S. Kumar. Microsoft ai red team building future of safer ai, 2023.
[63] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pages 1–14, 2020.
[64] L. McGregor, D. Murray, and V. Ng. International human rights law as a framework for algorithmic accountability. International & Comparative Law Quarterly, 68(2):309–343, 2019.
[65] Meta. Here's how we're using ai to help detect misinformation, 2020.
[66] Meta. Introducing two new datasets to help measure fairness and mitigate AI bias, 2022. https://ai.meta.com/blog/measure-fairness-and-mitigate-ai-bias/.
[67] Meta. How meta uses information for generative ai models, 2023. https://www.facebook.com/privacy/genai/.
[68] Meta. We have a responsibility to protect people's privacy and give them control to make their own choices, 2023. https://about.meta.com/privacy-progress/.
[69] Meta. Security, n.d. https://transparency.fb.com/es-la/metasecurity/.
[70] Meta AI. About meta ai, 2023. https://ai.meta.com/about/.
[71] Microsoft. Digital crimes unit: Leading the fight against cybercrime – on the issues, 2023. https://news.microsoft.com/on-the-issues/2022/05/03/how-microsofts-digital-crimes-unit-fights-cybercrime/.
[72] Microsoft. Introduction to red teaming large language models (llms) - azure openai service, 2023. https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/red-teaming.
[73] Microsoft. Microsoft AI, 2023. https://news.microsoft.com/ai/.
[74] Microsoft. Microsoft cyber defense operations center (cdoc), 2023. https://www.microsoft.com/en-us/msrc/cdoc.
[75] Microsoft. Microsoft privacy statement – microsoft privacy, 2023. https://privacy.microsoft.com/en-us/privacystatement.
[76] Microsoft. Service trust portal, 2023. https://servicetrust.microsoft.com/.
[77] Microsoft. Unleash your productivity with ai and microsoft 365 copilot - microsoft support, 2023.
[78] Microsoft. View your data on the privacy dashboard - microsoft support, 2023.
[79] Dan Milmo. Chatgpt reaches 100 million users two months after launch. The Guardian, February 2 2023. https://www.theguardian.com/technology/2023/feb/02/chatgpt-100-million-users-open-ai-fastest-growing-app.
[80] Benjamin Mueller. How much will the artificial intelligence act cost europe?, 2021.
[81] David Nersessian and Robert Mancha. From automation to autonomy: Legal and ethical responsibility gaps in artificial intelligence innovation. SSRN Scholarly Paper, 2020.
[82] AP News. Google suspends gemini ai chatbot's ability to generate pictures of people, 2024. https://apnews.com/article/google-gemini-ai-chatbot-image-generation-1bd45f1e67dfe0f88e5419a6efe3e06f.
[83] BBC News. Tay: Microsoft issues apology over racist chatbot fiasco, 2016. https://www.bbc.com/news/technology-35902104.
[84] Helen F. Nissenbaum. Privacy in Context: Technology, Policy, and the Integrity of Social Life. Stanford Law Books, 2010.
[85] John T. Nockleby. Hate speech. In Leonard W. Levy and Kenneth L. et al. Karst, editors, Encyclopedia of the American Constitution, pages 1277–1279. Macmillan, 2nd edition, 2000.
[86] OpenAI. Forecasting potential misuses of language models for disinformation campaigns and how to reduce risk, 2023. https://openai.com/research/forecasting-misuse.
[87] OpenAI. Gpt-4 technical report. Technical report, 2023.
[88] OpenAI. How should ai systems behave, and who should decide?, 2023. https://openai.com/blog/how-should-ai-systems-behave.
[89] OpenAI. Is chatgpt biased? — openai help center, 2023. https://help.openai.com/en/articles/8313359-is-chatgpt-biased.
[90] OpenAI. Openai, 2023.
[91] OpenAI. Openai red teaming network, 2023. https://openai.com/blog/red-teaming-network.
[92] OpenAI. Privacy policy, 2023. https://openai.com/policies/privacy-policy.
[93] OpenAI. Security & privacy, 2023. https://openai.com/security.
[94] OpenAI. OpenAI — security portal, n.d. https://trust.openai.com/.
[95] OpenAI. Openai platform, n.d. https://platform.openai.com.
[96] Zoe Papakipos and Joanna Bitton. AugLy: A new data augmentation library to help build more robust AI models, 2021.
[97] Alexander Peng, Besmira Nushi, Emre Kiciman, Kathleen Inkpen, and Ece Kamar. Investigations of performance and bias in human-ai teamwork in hiring. ArXiv.Org, February 2022.
[98] Matjaž Perc, Mehmet Ozer, and Janja Hojnik. Social and juristic challenges of artificial intelligence. Palgrave Communications, 5(1):1–7, 2019.
[99] Jan Piasecki, Marcin Waligora, and Vilius Dranseika. Google search as an additional source in systematic reviews. Science and Engineering Ethics, 24, 12 2017.
[100] II Price, W. Nicholson. Artificial intelligence in health care: Applications and legal implications. The SciTech Lawyer, 14(1), 2017.

[101] Flavien Prost and Alex Beutel. Mitigating unfair bias in ML models with the MinDiff framework, 2020.
[102] F. A. Raso, H. Hilligoss, V. Krishnamurthy, C. Bavitz, and L. Kim. Artificial intelligence & human rights: Opportunities & risks. SSRN Scholarly Paper, (3259344), 2018.
[103] Responsible AI Institute. Responsible artificial intelligence institute, n.d. https://www.responsible.ai.
[104] C. Rikap. Same end by different means: Google, amazon, microsoft and facebook's strategies to dominate artificial intelligence. SSRN Scholarly Paper, (4472222), 2023.
[105] M. Risse. Human rights and artificial intelligence: An urgently needed agenda. Human Rights Quarterly, 41(1), 2019.
[106] R. Rodrigues. Legal and human rights issues of ai: Gaps, challenges and vulnerabilities. Journal of Responsible Technology, 4:100005, 2020.
[107] G. Rosen. How we're tackling misinformation across our apps, 2021. https://about.fb.com/news/2021/03/how-were-tackling-misinformation-across-our-apps/.
[108] G. Rosen. Meta's q1 2023 security reports: Protecting people and businesses, 2023. https://about.fb.com/news/2023/05/metas-q1-2023-security-reports/.
[109] B. Rotsted, G. Sastry, H. Nguyen, G. Bernadett-Shapiro, and J. Parish. Openai cybersecurity grant program, 2023. https://openai.com/blog/openai-cybersecurity-grant-program.
[110] M. Schade. How your data is used to improve model performance — openai help center, 2023.
[111] D. Schönberger. Deep copyright: Up- and downstream questions related to artificial intelligence (ai) and machine learning (ml). Zeitschrift für geistiges Eigentum, 10(1):35, 2018.
[112] Mike Schroepfer. How ai is getting better at detecting hate speech, 2023.
[113] E. C. Schwarz. Human vs. machine: framework of responsibilities and duties of transnational corporations for respecting human rights in the use of artificial intelligence. Columbia Journal of Transnational Law, 58(1):232–278, 2019.
[114] Davide Spadini, Maurício Aniche, Margaret-Anne Storey, Magiel Bruntink, and Alberto Bacchelli. When testing meets code review: Why and how developers review tests. In Proceedings of the 40th International Conference on Software Engineering, ICSE '18, page 677–687, New York, NY, USA, 2018. Association for Computing Machinery.
[115] TensorFlow. Tensorflow federated, n.d. Retrieved October 4, 2023.
[116] The Alan Turing Institute. The alan turing institute, 2023. https://www.turing.ac.uk/.
[117] The Allen Institute for Artificial Intelligence. Allen institute for ai, 2023. https://allenai.org/.
[118] V. Türk. Artificial intelligence must be grounded in human rights, says high commissioner, 2023.
[119] United Nations. Peace, dignity and equality on a healthy planet. https://www.un.org/en/.
[120] United Nations. Guiding Principles on Business and Human Rights. United Nations, 2011.
[121] Michael Veale, Max Van Kleek, and Reuben Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, page 1–14, New York, NY, USA, 2018. Association for Computing Machinery.
[122] William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In Proceedings of the Second Workshop on Language in Social Media, pages 19–26, Montréal, Canada, June 2012. Association for Computational Linguistics.
[123] Claes Wohlin, Per Runeson, Martin Höst, Magnus C. Ohlsson, and Björn Regnell. Experimentation in software engineering. In The Kluwer International Series in Software Engineering, 2000.
[124] Karen Yeung. A study of the implications of advanced digital technologies (including ai systems) for the concept of responsibility within a human rights framework. Technical report, Council of Europe, 2019.
[125] The New york Times. Apple card investigated after gender discrimination complaints, 2019. https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html.
[126] Da Yu, Sidharth Naik, Adam Backurs, Suvam Gopi, Hakan A Inan, Gautam Kamath, Jaidev Kulkarni, Yoon Tae Lee, Alberto Manoel, Lukas Wutschitz, Sergey Yekhanin, and Hongyi Zhang. Differentially private fine-tuning of language models. arXiv preprint arXiv:2110.06500, 2022.

Maria Teresa Baldassarre is an Associate Professor at the Department of Computer Science, University of Bari "A. Moro", Bari, Italy. Her research interests include empirical software engineering, human factors in software engineering, and software measurement. Baldassare received her Ph.D. from the University of Bari "A. Moro". She is the representative of the University of Bari in the International Software Engineering Research Network and an associate editor of Decision Support Systems.

Danilo Caivano is currently a Full Professor of software engineering and project management with the Department of Computer Science, University of Bari "A. Moro", and a Consultant for companies and organizations especially in the field of research and development projects. He is also the Head of the SERLAB Research Laboratory and the Director of the short master in cyber security. He contributed to the creation of The Hack Space, Cyber Security Laboratory, University of Bari. He is also a member of the Board of Director of the Southern Italy Chapter Project Management Institute, the Co-Ordinator of the PMI-SIC Academy, and a member of the Technical Scientific Committee of the Apulian Information Technology District and the IT Strategic Steering Committee.

Berenice Fernandez Nieto Berenice has a BA and an MA in International Relations from the National Autonomous University of Mexico (UNAM) (Mexico). She also holds an International Master's in Security, Intelligence, and Strategic Studies (IMSISS) from the University of Glasgow (United Kingdom), Charles University (Czech Republic), and the University of Trento (Italy). From 2019 to 2022, she worked as a researcher at Data-Pop Alliance (Mexico). She is a Ph.D. student in the National Cybersecurity Programme at the University of Bari "A. Moro" (Italy) and IMT School of Advanced Studies Lucca (Italy) since 2022. Her research interests include digital rights, data governance, and cybersecurity.

Domenico Gigante Domenico Gigante received the Bachelor degree in Computer Science and M.S. degree in Cybersecurity from the University of Bari "A. Moro", Italy, in 2019 and 2021, respectively. Since 2021, he has been an industrial PhD student at the Department of Computer Science, University of Bari "A. Moro"; his fellowship is funded by SER&Practices Srl, where he currently works. His research interests include Trustworthy AI, secure software engineering, and empirical software engineering.

Azzurra Ragone is an Assistant Professor at the Department of Computer Science, University of Bari "A. Moro". Her current research interests include Trustworthy AI, Responsible AI engineering, Software Engineering for AI, and Human factors in software engineering. She worked as a researcher for several years at the Polytechnic of Bari, the University of Trento, the University of Michigan (USA) and the University of Milan Bicocca. She has published her research works in more than 90 national and international workshops and conferences as well as in international journals.