Charting Corporate Compass: The Influence of the UNESCO Recommendation on Ethics of AI

Segura Correa, Maria Alejandra Valens

Universiteit Leiden
The Netherlands

Bachelor thesis:

Charting Corporate Compass: The Influence of the UNESCO Recommendation on Ethics of AI

Bachelor Project: International Law, Use of Force, and Protection of Human Rights

BSc Political Science: International Relations and Organisations

Leiden University

Author: MAV Segura Correa (S3111962)

Supervisor: Dr. Müge Kinaciouglu

Second reader: Dr. Matthew di Giuseppe

Word count: 7733

Embargo statement: Public

24th May 2024

Acknowledgements

This thesis is dedicated to my parents whose love and support I am infinitely thankful and for being my long-distance companions in this experience: I could not have made it without you, thank you for your sacrifice, patience and most importantly, for believing in me.

To my friends, for becoming my extended family and encouraging me not give up: I will cherish our brunch days, our trips, and our long nights forever.

To my roommates, for waking me up to continue writing the thesis and sharing this last uni experience at homesita: We can finally sleep! And eat properly!

And finally, to Fabi, my little sister: I did it, and in the future, so will you. I missed you so much and one of my motivations for finishing was thinking on seeing you soon again. I adore you.

XoXo, Ale <3

Table of Contents

1. Introduction

Artificial Intelligence (AI) has rapidly emerged as a revolutionary technology that has transformed industries and redefined human interactions. AI refers to the capacity of computers and machines to replicate behaviors that resemble human intelligence (Lucey, 2022). This rapid development has challenged traditional regulatory frameworks; hence self-regulatory or 'soft law' approaches to govern AI design have emerged, since they are non-binding regulations without enforcement power (Taeihagh, 2021, p. 145). In this context, private tech companies play a crucial role in shaping AI regulations, addressing the processes and developments of AI, mostly following a risk-based perspective. Still, international regulations offer a human rights-based perspective, which lays out the legal basis for the ethical use of AI.

While the international community has yielded frameworks like the UNESCO Recommendation on the Ethics of Artificial Intelligence (Recommendation), the ways in which these principles are translated into corporate practices remain unclear. This paper aims to narrow down the gap by answering the question *In what ways does the UNESCO Recommendation on the Ethics of Artificial Intelligence impact corporate AI governance?*

To answer this question, this thesis first evaluates the different scholarly contributions on governance, in the AI and corporate domain, and the perspectives on adopting international guidelines or private self-regulation. Then, a theoretical approach is introduced, to construct the human rights-based approach (HRBA) and risk-based approach (RBA), within polycentric governance. Likewise, the conceptualisation defines what I mean when I refer to AI corporate governance. Furthermore, the research is conducted through qualitative content analysis (QCA) to explore the AI frameworks within the corporate domain. The findings are presentend as well as the discussion …

Since the aim is to explore the impact of the Recommendation on corporate AI frameworks, this paper argues that AI corporate governance is impacted in different areas such as safety and security, multi-stakeholder collaboration, ethical assessments and protection and promotion of human rights. Moreover, the research contributes to the ongoing debate on AI governance by examining the interaction between international guidelines and corporate practices. Understanding how the Recommendation shapes AI corporate governance, can draw attention to the approach companies use to develop their frameworks.

2. Literature Review

2.1. Governance

Governance is a longstanding term, particularly within international relations due to its political implications. Yet, it also expands to the social, corporate and technological fields. Some scholars approach corporate governance as the legal requirements for companies based on a shareholder-centric view. Cadbuy (1992) assessed corporate governance as "the system by which companies are directed and controlled" (as cited in Hilb, 2020, p. 852). Expanding on that idea, Ansell & Torfing (2022) maintained that corporate governance encompasses the institutionalised interaction among shareholders, management, boards of directors, employees, customers, institutions and the community involved in directing and controlling private companies (p. 2). Moreover, Mäntymäki et al. (2022) shifted the focus, contending that in addition to legal compliance, companies can include aspects that define desired behavior beyond what the law requires. Consequently, they can establish the rules by which to understand and enforce the desired behavior of their agents, while managing the relationship between shareholders and stakeholders in case of tensions (p. 605). The *G20/OECD Principles of Corporate Governance 2023* featured the characteristics described by the previously discussed authors (Mäntymäki et al. and Ansell & Torfing) asserting that corporate governance

involves a set of relationships between a company's management, board, shareholders and stakeholders. The OECD (2023) declared that it also provides the structure and systems through which companies are directed and its objectives are defined (p. 6). It is also important to note that as AI technologies progressed in contemporary society, scholars have also shed light on the implications of governance in the realm of AI. According to Dafoe (2018), AI governance focuses on the institutions and contexts in which AI is developed and deployed. The aim is to increase the likelihood that AI developers and users have the goals, incentives, perspectives, time, training, resources, support, and equipment required to do so for the benefit of society (p. 6). Supporting this view, Mucci & Stryker (2023) maintained that AI governance refers to the guardrails that guarantee safe and ethical AI systems and tools. It establishes frameworks, regulations, and standards to guide AI research, development, and application while ensuring safety, fairness, and respect for human rights. Mäntymäki et al. (2022) complied with Dafaoe and Mucci & Stryker, defining AI governance as the set of rules, practices, processes, and technological tools employed to ensure AI technologies align with strategies, objectives, and values of companies (p. 604).

2.2. AI, international/governmental regulation, and human rights

National governments and international organisations play pivotal roles in shaping the landscape of AI technologies. Together, these entities contribute to responsible and sustainable AI technologies. Some scholars favour implementing human rights into international frameworks in the early stages of AI development, like Rodrigues (2020). He asserted that whilst AI technologies interact with massive volumes of data, they might have overlapping effects posing legal and human rights issues. Hence, early considerations of the impact of AI on human rights, ethics, and societal values are crucial to mitigate such difficulties (p. 9). Furthermore, in the context of public administration and governance, specific public values are

relevant. These values include efficiency, effectiveness, accountability, transparency, and equity (Chen et al., 2023, p. 3). Su (2022) stated that adopting ethical AI practices requires a comprehensive understanding of these values, and international law can contribute to the organised coordination of national AI strategies and principles issued by private companies and public-private partnerships (p. 173). Likewise, Donahoe and Metzger (2019) pointed out that it is necessary to advocate for a global set of regulations based on human rights, otherwise, there is the risk of retaining negative effects. Additionally, if most members of the international community can agree on human rights-driven frameworks for AI governance, it will influence more states to adopt them (p. 123). Developing further from this view, Jones (2023) argued that international norms and procedures can be appropriately based on human rights so companies can build customer trust and reduce potential costs and time (p. 12). Following this perspective, the Universal Declaration of Human Rights (UDHR) must apply to AI. Despite the precise generation of AI systems, it is essential to ensure their alignment with human values to integrate human rights into business decisions (Risse, 2019, p. 9). Tzimas (2021) argued for a framework that critically evaluates the influence of human rights in the AI domain, guiding the enablement or restriction of certain AI developments and applications (p. 138). Accordingly, AI systems should empower human rights by allowing citizens to make informed decisions and fostering their fundamental rights (Gesley, 2020, p. 241). Opposing this view, Currie (n.d.) contended that it makes sense for a company to have a risk-based framework to analyse its own initiatives rather than relying on government regulation since it might inhibit its business model (p. 9). Furthermore, in its report, the Council of Europe (2024) revealed that national authorities produced 172 AI-related frameworks between 2010 and 2022, while international organisations produced 214 AI-related frameworks between 2015 and 2023. Interestingly, during this shorter time frame, international organisations have been more active than national authorities. Cihon, Schuett and Baum (2021) argued that the delayed

governmental responses to emerging technologies underscore the clear necessity for alternative stakeholders to engage in enhancing AI governance for the public interest (p. 21). While the existing literature explores the integration of human rights principles into AI regulations, it often emphasises the "musts" and "shoulds" rather than examining how human rights are translated into regulations within private companies. This research gap can be addressed by assessing the human rights perspective proposed by the Recommendation within private companies, given its adoption by member states.

2.3. AI, corporate regulation and private companies

Many tech companies advocate for self-regulation or market-driven governance as the best solution to AI and associated technologies. Su (2022) argued that both industry and government stakeholders recognise that the law is slow to adapt and can sometimes hinder innovation. Consequently, there is a prevailing belief that the industry is in the best position to develop the standards and rules that will guide innovation, considering public welfare and minimising the risks (p. 169). Building upon this view, Jones (2023) maintained that it is in the best interest of tech companies to produce academic papers about their research, even though discussion initiatives led by corporations often fail to address the ethical dimensions of AI and the significance of integrating—or interacting with—human rights principles (p. 11). Supporting Jones, Cihon et al. (2021) argued that private companies are the major players – if not the main – in AI research, development, and deployment. For instance, in the United States in 2018, 50% or more academic papers focused on AI, were published by corporate-affiliated researchers (p. 1). In that sense, Erman & Furendal (2024) mentioned that there are two overlapping ways in which tech companies significantly influence the emerging global governance of AI. First, they have epistemic authority, where they are not only developers but have the potential to shape public opinion and policy decisions (p. 4). Ulnicane et al. (2021)

did not share this view. They stated that the current oligopoly of a small number of large companies is one of the reasons for problems such as a lack of consideration of societal needs and concerns (p. 171). While the literature explores the concept of self-regulation in AI governance by tech companies, a critical gap remains regarding how these companies are incorporating international regulations within their AI governance frameworks. The literature gap can be assessed by investigating how tech companies promote safe AI practices.

3. Theoretical and Conceptual Framework

3.1. Polycentric governance theory

The polycentric governance theory holds that the state cannot perform all the complex duties required to address the most pressing social challenges on its own and, as a result, must work with other actors to leverage its capabilities (Ruggie, 2014, pp. 8-9). The Bloomington School proposes that this approach has three characteristics: (1) multiple centres of semiautonomous decision-making – there is no single decision-making centre with ultimate authority, but rather multiples; (2) the existence of a single set of rules – whether institutionally or culturally enforced; and (3) the existence of a spontaneous social order – as a result of evolutionary competition between different ideas, methods, and ways of life (Xue, 2024, p. 225). Additionally, Aguerre et al. (2024) argued that polycentrism exposes several power centres and relationships in digital data governance. This idea encompasses formal and informal structures, multiple levels (local to global), and several sectors (governmental, commercial, civil society, technological, academic). From a polycentric perspective, multiple disciplines bring together a growing range of insights facilitating interdisciplinary discussions about the rules and regulatory processes around AI (p. 3-9). In the context of AI corporate governance, the Recommendation exemplifies a potential set of shared rules, establishing principles for the development, deployment and use of AI that can be culturally enforced. Concurrently,

exploring how it influences the behavior of various actors, in this case private tech companies, it emphasises the different power centers surrounding AI governance. For that reason, considering a polycentric approach, the paper will focus on examining how companies shared the regulations outlined by UNESCO. Based on this, we can expect the Recommendation to serve as a potential set of shared rules for tech companies, even though there may be different power dynamics at play.

3.1.1. Human-rights based approach

As shown in the literature review, many scholars argue for a greater integration of human rights in the legal guidelines to govern AI, called a human rights-based approach – referred to as HRBA. This approach, as advocated by Donahoe & Metzger (2019), can achieve what the newly developing ethical frameworks seek to accomplish. They suggest four features to enable this: First, AI governance should prioritise individuals by focusing assessments on their impact. Then, address a wide variety of AI-related issues, both procedural and substantive. Additionally, outlines the roles of governments and the private sector in safeguarding human rights. Finally, reach a global consensus to guarantee universal application. This paradigm highlights the need to protect human dignity and place individuals as the central focus of governance to ensure AI benefits them rather than harms them (p. 119). Aligning to this perspective Yeung, Howes & Pogrebna (2020) found that terms like transparency, fairness, and explainability are often discussed when assessing the effect of AI technology indicating ethical considerations (p. 80). Furthermore, they argued that HRBA relies on independent regulatory entities with investigative and enforcement capabilities to exert oversight over AI technologies (p. 101) and ensure these technologies follow ethical guidelines. This approach is achieved when, in the case of this research, tech companies pursue ethical norms and principles in their AI frameworks, hence, prioritising the protection of human

rights. Based on HRBA we can expect to see a focus on protecting human rights, this might involve ensuring transparency, fairness, and explainability in AI systems, along with mechanisms for independent oversight to ensure companies adhere to ethical principles.

3.1.2. Risk-based approach

Simultaneously, scholars advocating for a self-regulatory approach argue for a risk-based approach – referred to as RBA. Wirtz et al. (2022) stressed that the complexity and rapid development of AI necessitates RBA that is integrative, flexible, and adaptive, which can be best realised through collaboration among all relevant stakeholders including governments, the tech industry, NGOs, and academia (p. 9). According to Mahler (2022), it may be appropriate for achieving proportionality and avoiding regulatory overreach (p. 267). To achieve this, according to Lütge et al. (2022), AI systems can be classified as minimal or no risk, limited risk, high-risk or prohibited (p. 1). In addition, Malgieri and Kamath (2023) sustained that the RBA aims to achieve several specific objectives: Foster innovation and competitive potential within the private sector, establish clear and enforceable liability obligations and safety measures, enable precise allocation of compliance responsibilities and potential liabilities, mitigate the impact of biased or discriminatory outcomes resulting from algorithm-based decisions and ensure the continuity of existing data protection and privacy principles (p. 22). Therefore, for the purpose of this research, the RBA is accomplished when tech companies opt for prioritising the implications of AI developments in terms of risk-assessment and collaboration with different sectors. Elaborating on the RBA, we can expect we can expect companies to adopt a flexible and adaptable approach to mitigate AI risks and ensure safety. AI corporate frameworks will likely balance innovation with the need for safety considerations and interaction among stakeholders.

3.2. Conceptualisation

3.2.1. AI Corporate Governance

In the literature review, I explored various perspectives on governance, including AI governance and corporate governance. However, the focus of this research lies specifically on AI corporate governance. I ellaborate further taking the concept of Mäntymäki et al. (2022), which cited Schneider et al., defining AI governance for companies as rules, practices, and processes used to ensure that AI technologies align with the strategies and objectives of corporations (p. 604). Moreover, I extend beyond the corporative benefits of AI governance to include the values that tech companies adhere to. As a result, my definition of AI corporate governance encompasses the rules, principles and procedures employed by tech companies to guarantee the responsible development and deployment of AI technologies. This definition of AI corporate governance diverges from existing literature on governance, since it does not approach governance as just legal processes and interactions between stakeholder. It is important to mention that, in this paper, the terms "AI corporate governance" and "AI corporate frameworks" will be used interchangeably.

3.3. Hypotheses

After carefully examining the relevant theories and concepts, building upon the HRBA and RBA, I formulated four hypotheses that I believe are pertinent for assessing the influence of the Recommendation on AI corporate governance:

Based on the HRBA,

H1a: When private tech companies adhere to the UNESCO Recommendation on the Ethics of Artificial Intelligence, their frameworks are more likely to promote and protect human rights.

H1b: When private tech companies adhere to the UNESCO Recommendation on the Ethics of Artificial Intelligence, their frameworks consider ethical assessments.

Based on the RBA,

H2a: When private tech companies adhere to the risk-based approach outlined in the UNESCO Recommendation on the Ethics of AI, their frameworks assess potential risks associated with AI deployment and actively mitigate them.

H2b: When private tech companies adhere to the UNESCO Recommendation on the Ethics of Artificial Intelligence, their frameworks foster collaboration and cooperation with diverse stakeholders and sectors of society.

Based on polycentric governance,

H3: When AI corporate frameworks align with the UNESCO Recommendation principles on the Ethics of AI, then it positions the Recommendation as a global shared set of rules

Testing these hypotheses will help me to gain a deeper understanding of how the UNESCO Recommendation shapes AI corporate governance and potentially predict some of the key findings. At the same time, it is important to note that these hypotheses were developed prior conducting the analysis.

4. Research Design and Methodological Approach

This thesis investigates how the Recommendation shapes AI corporate governance through a qualitative approach using content analysis. Since I aim to analyse written documents, according to Halperin & Heat (2020), content analysis is an appropriate method. Particularly, the recording unit in this research is sentences since it can provide factual insights from the documents.

4.1. Case selection

To select relevant cases for this research, I first examined the Council of Europe's report "AI Initiatives" (2024) to identify tech companies that issued AI-related. Then, considering the independent variable is the Recommendation, issued in late 2021, I focused on companies that published or updated AI-related frameworks after this timeframe. This approach ensures that frameworks could potentially incorporate the Recommendation's principles. Following a purposive sampling strategy, Google and Microsoft were selected since they are the most important leaders in the AI domain, particularly Generative AI (Linacre, 2024). At the same time, both companies released AI frameworks after 2022, respecting the timeframe previously established. While selecting smaller companies can also provide valuable insights, it might not be as relevant and representative as choosing large tech companies.

4.2. Data and Sources

Given the variety of research papers, reports and frameworks available, the source selection process focused on narrowing down the results to ensure the data is relevant to assess the influence of the Recommendation on AI corporate governance. To achieve this, official documents from Google and Microsoft were chosen based on their titles, since they explicitly contain words alluding to governance such as "governing," "standards," "framework," and "principles." On this basis, four sources were deemed most relevant for analysis: *Google's AI Principles Progress Update 2023* (Google, 2024), *Secure AI Framework Approach: A Quick Guide to Implementing the Secure AI Framework (SAIF)* (Google, 2023), Microsoft's *Governing AI: A Blueprint for the Future* (Microsoft, 2023) and *Microsoft Responsible AI Standard, v2* (Microsoft, 2023). While a broader range of AI-related documents were available, these four sources offer a suitable for analyzing AI corporate governance in the context of the Recommendation.

4.3. Coding frame

Drawing upon the HRBA and RBA conceptualisation, the coding frame takes a mixed approach, since it also includes the principles obtained from the Recommendation (UNESCO, 2021). This concept and data-driven perspective helps to get a more clear and specific classification of the dimensions for identifying relevant elements during the coding process. In that sense, the categories are divided into 2 dimensions: HRBA and RBA, taken from the conceptualization section. Simultaneously, the subcategories and indicators are drawn upon the policy areas covered by the Recommendation.

The first category is HRBA, which has been divided into 3 subcategories:

a. Sustanability:

Continuous assessment of the human, social, cultural economic and environmental impact of AI technologies should therefore be carried out with full cognizance of the implications of AI technologies for sustainability as a set of constantly evolving goals (p. 21). This subcategory can be identified when the sentence discusses the impact of AI at the human, social, cultural, economic and environmental level. Another indicator is when sustainable initiatives are explicitly mentioned. The assigned code is HRS.

b. Protection and promotion of human rights:

i. *Privacy*

The right to privacy protects human dignity, human autonomy and human agency, and should be respected, protected and promoted throughout the life cycle of AI systems (p. 21). This subcategory can be identified when the sentence mentions respecting privacy. The assigned code is HRP.

ii. Data protection

Data protection frameworks and any related mechanisms should take reference from international data protection principles and standards concerning the collection, use and disclosure of personal data and exercise of their rights by data subjects while ensuring a legitimate aim and a valid legal basis for the processing of personal data, including informed consent (p. 21). This subcategory can be identified in three ways. First, when the sentence mentions data protection principles. Second, when the sentence mentions appropriate measures to collect and process personal/sensitive data. Third, when the sentence mentions informed consent. The assigned code is HRDP.

c. Ethical asessement

i. Oversight

The Recommendation stated that human oversight refers not only to individual human oversight, but also to inclusive public oversight. As a rule, decisions should not be ceded to AI systems. This subcategory can be identified in two ways – when sentences mention human oversight and they mention the use of independent (external or internal) AI Ethics Committees/Entities. The assigned code is HRO.

ii. Diversity and inclusiveness

AI should be available and accessible to all, taking into consideration the specific needs of different age groups, cultural systems, different language groups, persons with disabilities, girls and women, and disadvantaged, marginalized and vulnerable people or people in vulnerable situations (p. 20). This subcategory can be identified when the sentence mentions representation of minorities and marginalised groups (disabled individuals, women, ethnic groups). Another indicator is when sentences mention the promotion of inclusive initiatives (language, gender, equality, battling stereotypes). The assigned code is HRDI.

iii. Transparency

People should be fully informed when a decision is informed by or is made on the basis of AI algorithms. Individuals should be able to understand how each stage of an AI system is put in place, appropriate to the context and sensitivity of the AI system (p. 22). There are two indicators to identify this subcategory. The first is when sentences mention transparency requirements. The second is when sentences mention informing users they are interacting with AI. The assigned code is HRT.

The second category is RBA, which has been divided into 3 subcategories:

a. Safety and security

The UNESCO (2021) established that unwanted harms (safety risks), as well as vulnerabilities to attack (security risks) should be addressed, prevented and eliminated throughout the life cycle of AI systems to ensure human, environmental and ecosystem safety and security. At the same time, AI processes should not exceed what is necessary to achieve legitimate aims and are context-appropriate (p. 20). This subcategory can be identified when sentences mention risk/harm prevention and mitigation; also, when sentences mention safety and security measures – also referred to as impact/risk assessments. Since proportionality is included within this subcategory, the last indicator for safety and security is when sentences mention proportional measures. The assigned code is RSS

b. Awareness and literacy

Public awareness stands for the promotion of understanding AI through open and accessible methods such as education, civic engagement, training, media and information literacy to ensure effective public participation. This ensures all members of society to take informed decisions about their use of AI systems (p. 23). There are three indicators to help us identify this. The first, sentences mention informed decision. The second, sentences mention

promoting AI education. The third, sentences mention accessible and public information. The assigned code is RAL.

c. ***Multi-stakeholder collaboration***

The Recommendation also stated that within the realm of AI, stakeholders include but are not limited to governments, intergovernmental organisations, the tech industry, civil society, researchers and academia, media, NGOs, etc. The adoption of open standards and interoperability to facilitate collaboration should be in place to allow for meaningful participation of marginalised groups, communities and individuals (p. 23). To identify this, we can follow four indicators. When sentences mention collaboration among different sectors (academia, NGOs, governments, companies); when sentences mention international cooperation and when sentences mention collaborative research. The assigned code is RMSSP.

The coding frame matrix can be found in the Appendix, section 8.1.

5. Analysis

The Recommendation is a global agreement adopted by all 193 Member States on the 23rd of November 2021. It provides a framework for the ethical governance of AI and emphasises the protection of human rights, covering extensive policy areas, including data governance, environment, gender, education, research, and health. Since it has the potential to be influencial in the private sector, this analysis draws upon the assumption that tech companies take the Recommendation as a baseline for developing their frameworks. As established in the methodology, I will be looking at the sentences to analyse whether the impact is on following an HRBA or an RBA, and the specific areas in which this impact is manifested. Additionally, for a more detailed textual representation of the findings, consult the coding text matrices attached to the Appendix, specifically sections 8.2.1 and 8.2.2.

5.1. Google AI frameworks

5.1.1. Secure AI Framework Approach: A quick guide to implementing the Secure AI Framework (SAIF)

The *Secure AI Framework Approach: A quick guide to implementing the Secure AI Framework (SAIF)* was issued in 2023 as a conceptual framework for secure AI systems and intended to provide high-level practical considerations on how organisations could go about building the SAIF approach into their existing or new adoptions of AI (Google, 2023, p. 2). The analysis is based on the possible areas of impact of the Recommendation following an HRBA and an RBA.

Under HRBA, I seek to identify considerations for data protection, oversight, sustainability, privacy, diversity and inclusiveness and transparency. While content related to the latter four indicators wasn't found in the SAIF framework, the document addressed data protection and oversight. For instance, oversight was emphasised throughout the framework. First, Google (2023) argued that human oversight is necessary to ensure that AI systems are used ethically and responsibly because they can be biased or make mistakes (p. 7). At the same time, the document mentions that it is essential to keep humans in the loop to oversee relevant AI systems, processes, and decisions (p. 8). This aligns with the ethical assessment element recognised by the Recommendation, following a HRBA, even though transparency and diversity and inclusiveness are not included. Besides, data protection is touched upon by the SAIF framework since it mentions using data security controls to protect the data AI systems use to train and operate (p. 4). Even though this topic was less discussed, this finding aligns with the promotion and protection of human rights as established by the indicator for data protection, while privacy is missing.

Moving to the RBA dimension, the analysis seeks to identify safety and security, multi-stakeholder collaboration, and awareness and literacy. In this case, while the latter was not

present, the other two subcategories were successfully recognised. *Safety and security* emerged as the most prominent area within the SAIF framework. The document emphasizes a plan for detecting and responding to security incidents and mitigating the risks of AI systems making harmful or biased decisions (p. 6). Additionally, it advocates for understanding the specific risks associated with AI models in use and implementing security controls to mitigate those risks along with having clear roles and responsibilities (p. 9). This focus on risk assessment and mitigation aligns with the indicators established for *safety and security*. While less emphasised, this framework acknowledges the composition of a team including stakeholders across multiple organizations (p. 3).” This aligns with the Recommendation's call for multi-stakeholder engagement in AI governance based on one of the indicators proposed for identifying *multi-stakeholder collaboration.*

Drawing upon the findings, the *SAIF* framework reflects both dimensions, HRBA and RBA, outlined in the Recommendation. It prioritises *ethical assessment* and *protection and promotion of human righ*ts from a HRBA, and *multistakeholder collaboration* and *safety and security* from a RBA. Overall, the findings suggest the Recommendation has impacted Google’s AI corporate governance approach to safety and security measures, establishing appropriate oversight entities, collaborating with different actors and sectors for AI development, and ensuring the protection of users’ data.

5.1.2. AI Principles Progress Update 2023

The *AI Principles Progress Update 2023* was issued in 2024 as the 5th edition of Google’s annual AI Principles progress report, to provide input into how they implement the core principles into practice (Google, 2024, p. 6). This analysis examines the aforementioned framework through the lens of HRBA and RBA, following the indicators established in the coding frame.

Opposite to the first analysed document, the Google's *AI Principles Progress Update 2023* demonstrates strong alignment with HRBA principles as suggested by the Recommendation. Firstly, *sustainability* was identified in several instances. Google (2024) mentions considering a broad range of social and economic factors (p. 3), alongside environmental sustainability, when reviewing AI systems. This demonstrates alignment with the first subcategory, sustainability, as approached by the Recommendation, using the appropriate indicators. Besides, the framework emphasises privacy protections for giving users choice and control over their private data (p. 34). This relates to the *privacy* indicator. Moreover, in a less frequent manner, Google explicitly commits to protecting personal information through security controls, representing the *data protection* subcategory. These findings align with the protection and promotion of human rights since there are references to both privacy and data protection. At the same time, the framework focuses on developing techniques to build more inclusive models for people from diverse backgrounds. To achieve this, it references community-based research efforts, focusing on historically marginalised communities or groups of people who may experience unfair outcomes of AI (p. 19). Besides, Google asserts "methods to account for rater diversity (p. 20)." These findings position *diversity and inclusiveness* as the most prominent subcategory of HRBA. In addition, there are allusions to *oversight* mentioning appropriate human direction and control (p. 4). Finalising with the HRBA category, *transparency* is also mentioned. The document is very explicit when addressing transparency documentation, releasing technical reports and sharing standards on model transparency (p. 13–15). This means that the framework successfully demonstrates alignment with the ethical assessment subcategory as approached by the Recommendation following the corresponding indicators.

Similar to the SAIF framework, the *AI Principles Update Report 2023* shows that the Recommendation impacted all the areas referring to the RBA since all the subcategories were

identified. *Safety and security* is easily identifiable within this framework due to the repetitive color coding in the text. Google underscors the continuous development and application of strong safety and security practices (p. 3), highlighting that AI Principles ethics reviews and impact assessments are part of a larger safety testing and security reviews (p. 9). Furthermore, it states that Google's AI Principles would guide how to limit harmful outcomes (p. 25). This shows the impact of the Recommendation on AI corporate governance. Instances such as participating in external community engagement (p. 21), and "collaborating with underrepresented groups in the international community (p. 22)," illustrate the representation of *multi-stakeholder collaboration* within the framework. This demonstrates commitment to promoting collaboration across groups, as outlined by UNESCO. While not as pronounced as the other subcategories, *awareness and literacy* is included. The framework focuses on communicating essential practices, information literacy to support AI knowledge and informing users when they are engaging with AI systems. This reflects alignment with the third subcategory of RBA under the indicators based on the Recommendation.

Altogether, the analysis reveals that Google's *AI Principles Progress Update 2023* takes both a HRBA – with a special focus on *ethical assessment* and pro*tection and promotion of human rights*, and a RBA – following all three safety and security, awareness and literacy and multi-stakeholder collaboration, which reflects the translation of the Recommendation principles into AI corporate governance. Similarly to the SAIF framework, the evidence suggests the Recommendation has impacted Google's AI corporate governance approach to data protection, privacy, diversity and inclusiveness, transparency, safety and security, oversight and multi-stakeholder collaboration.

5.2. Microsoft AI frameworks

5.2.1. Microsoft Responsible AI Standard, v2

The Microsoft Responsible AI Standard, v2 was issued in 2022 and it is the second edition as the product of a multi-year effort to define product development requirements for responsible AI (Microsoft, 2022, p. 3). This analysis examines how the Recommendation has impacted Microsoft's AI corporate frameworks in the context of the *Microsoft Responsible AI Standard, v2* following the coding frame.

Starting with the identification of HRBA indicators, the sustainability subcategory was lacking. This means that contrary to what the Recommendation suggested, sustainability is not considered within this framework, hence it had no impact on this field. Still, the other two subcategories were represented. Alluding to *privacy*, Microsoft AI systems adhered to protecting privacy following the Microsoft Privacy Standard (p. 26). The framework stresses the definition and documentation of procedures for the collection and processing of data (p. 7). In the meantime, *data protection* is also mentioned once. Even if these indicators are slightly discussed, it suffices to give an insight and argue that it reflects the implications according to UNESCO to demonstrate alignment with the protection and promotion of human rights subcategory. Simultaneously, the term *oversight* was identified but does not align with the indicators since it lacks detail. This means that, on this basis, an actual impact cannot be assured due to the limitations the findings present. Accordingly, for the *diversity and inclusiveness* subcategory, plenty of evidence is found highlighting Microsoft's efforts to design AI systems to provide a quality of service for identified demographic groups, including marginalised groups and to evaluate all data sets to assess the inclusiveness of those groups and close gaps (p. 13). It also mentions that Microsoft AI systems are designed to be inclusive in accordance with the Microsoft Accessibility Standards (p. 27). This category is the most represented

subcategory under HRBA, showing a noticeable impact of the Recommendation on this area. Correspondingly, records from this framework include *transparency.* The framework explicitly states that Microsoft AI systems are designed to inform people they are interacting with an AI tool (p. 12). Basing the latest findings, the subcategory ethical assessment is conceded, following the indicators established following the Recommendation principles, fulfilling oversight, transparency and diversity and inclusiveness.

Following the RBA, evidence points that the *Microsoft Responsible AI Standard, v2* framework includes all three subcategories under this approach. This framework includes *multi-stakeholder collaboration* by mentioning that the tech industry, academia, civil society, and government need to collaborate to advance the state-of-the-art and learn from one another (p. 3). This shows a compliance with the principle outlined by UNESCO, thereby showing its impact on multi-stakeholder collaboration. Concurrently, *awareness and literacy* is covered in this framework. Evidence shows the scope covered to provide documentation to customers describing the AI system's intended uses, demonstrating the system fits the purpose for each intended use (p. 6). Also, it compromises to inform about the capabilities and limitations of AI systems and publish documents to understand them (p. 11). This means the framework adopts the principle of *awareness and literacy* according to the indicators and demonstrates the Recommendation's impact on AI corporate governance. Moreover, evidence indicates *safety and security* since the framework constantly highlights the importance of impact assessments. This aligns with the indicators to measure the impact on this subcategory, thus, demonstrating the impact on this domain.

Reflecting on the findings, the *Microsoft Responsible AI Standard, v2* framework reveals compliance with the HRBA and the RBA. This indicates the impact of the Recommendation on the *protection and promotion of human rights* and *ethical assessment*, as well as in *safety and security*, *awareness and literacy* and *multi-stakeholder collaboration* - key aspects of both

categories analysed. Altogether, the framework evidences the efforts of Microsoft to integrate the principles outlined in the Recommendation, which suggests the impact it had on AI corporate governance in the fields of *data protection*, *privacy*, *diversity and inclusiveness*, *transparency*, *multi-stakeholder collaboration*, and *safety and security*.

5.2.2. Governing AI: A Blueprint for the Future

The *Governing AI: A Blueprint for the Future* framework was issued in 2023, it builds on the work we have done and will continue to do to advance responsible AI through company culture to forge a responsible future for artificial intelligence (Microsoft, 2023, p. 8). This analysis studies this framework, following a HRBA and a RBA to identify the impact of the Recommendation on it.

After examination the document, no records of *data protection* or *privacy* were found. This indicates that Microsoft has not focused on protection and promotion of human rights, as suggested by the Recommendation, hence there was no impact on this area. Nevertheless, *sustainability* is identified within framework since it emphasises a sociotechnical lens to develop AI systems, considering the cultural, political, and societal factors of AI (p. 36). This means that as recommended by UNESCO, this framework includes *sustainability*, recognising an area of impact. At the same time, evidence shows that under the HRBA, Microsoft considers *oversight*. Especially, the framework explains the existence of the Environmental, Social, and Public Policy Committee of the Microsoft Board to provide oversight over AI systems (p. 31). On top of that, they created a program for ongoing review called Sensitive Uses (p. 33). Also, *diversity and inclusiveness* and *transparency* are specified with limited findings. This can be shown in the continuous attempts of Microsofts to champion diversity and inclusion at all levels of AI programs by recruiting and retaining a diverse, dynamic, and engaged employee community (p. 37). At the same time, they provide transparency documentation in the form of

Transparency Notes (p. 39). These findings strictly address the importance of ethical assessments within the framework, evidencing the impact of the Recommendation on this area.

Shifting onto the RBA dimension, all three subcategories are identified. *Multi-stakeholder collaboration* is demonstrated when Microsoft mentioned supporting and collaborating with a multistakeholder group, including representatives across academia (p. 25) and bringing the public and private sectors together to use AI as a tool to improve the world (p. 26). This highlights the Recommendation impact on AI corporate governance by promoting *multi-stakeholder collaboration*. Then, evidence for *awareness and literacy* is revealed in Microsoft's support of broad educational initiatives to make information about AI technologies and responsible AI practices available to legislators, judges, and lawyers (p. 19) and the development of a national registry of high-risk AI systems that is open for inspection so that individuals can learn where and how those systems are used (p. 23). This points out the impact of the Recommendation on AI frameworks at the *awareness and literacy* level. Lastly, *safety and security*, records most mentions overall. Microsoft established this subcategory as a standard goal, aiming at minimising the time to fix AI system failures by identifying potential harms using iterative red teaming to mitigate them (p. 33). In parallel, it seeks to regulate licensed AI deployment through pre-release safety and security requirements, with post-deployment safety and security monitoring and protection (p. 18). The alignment with this subcategory, shows the impact produced by the Recommendation, to prevent and mitigate unwanted harm.

Based on the evidence, the *Governing AI: A Blueprint for the Future* framework fails to address the promotion and protection of human rights areas, as described by the HRBA dimension. Despite this, it includes other relevant aspects, such as sustainability and ethical assessments. At the same time, focusing on the RBA, this framework prioritises safety and security, awareness and literacy and multi-stakeholder collaboration. All in all, the findings reflect the

impact the Recommendation had on AI corporate governance, within the Microsoft frameworks.

6. Discussion and further considerations

The discussion builds upon the findings presented in section 5 and it aims to address the research question *In what ways has the UNESCO Recommendation on the Ethics of Artificial Intelligence impacted corporate AI governance?* by answering to the hypotheses proposed in section 3.3. In general, the analysis of Google's AI Principles Progress Update 2023 (Google, 2024) and Secure AI Framework Approach (SAIF) (Google, 2023), alongside Microsoft's Governing AI: A Blueprint for the Future (Microsoft, 2023) and Microsoft Responsible AI Standard, v2 (Microsoft, 2023), reveals both convergence and divergence in how these companies approach AI governance frameworks, yet it demonstrates that the Recommendation has impacted these frameworks in specific areas.

On that account, all four sources addressed *safety and security* and *multi-stakeholder collaboration,* key components of the RBA dimension. This demonstrates the shared industry focus on mitigating potential risks associated with AI deployment and fostering collaboration across sectors, thus revealing a clear impact of the Recommendation on AI governance frameworks on *safety and security* and *multi-stakeholder* aspects. This finding supports H2a and H2b by demonstrating that Google and Microsoft prioritise risk assessment and mitigation, as well as collaboration across different sectors.

In addition, Google's frameworks emphasise HRBA principles, as they particularly highlight ethical assessment – prioritising transparency, oversight and diversity and inclusiveness – and protection and promotion of human rights – referring to privacy and data protection. Microsoft's frameworks also consider *ethical assessment* from a *transparency* and *oversight* perspective and *data protection* within *protection and promotion of human rights*, under the

HRBA. Integrating these findings, one can assume that under the influence of the Recommendation, private tech companies promote and protect human rights and consider ethical assessments – supporting H1a and H1b. Nonetheless, it is important to mention that both, Google and Microsoft, are aligned with *sustainability* – an aspect recognized due to its incorporation in the Recommendation, rather than in the theories. This also demonstrates another area of impact, considering the HRBA.

Altogether, both companies demonstrate that the Recommendation has impacted their AI frameworks, influencing companies to address various aspects of responsible AI development and deployment in the areas proposed by the Recommendation such as sustainability, *data protection*, privacy, oversight, transparency, diversity and inclusiveness, safety and security, awareness and literacy and multi-stakeholder collaboration. These findings suggest that the Recommendation can be considered a shared set of rules since private tech companies incorporated the principles described by UNESCO in their AI corporate frameworks, showing that it is influential – supporting H3.

Digging deeper into the sentences, it is also important to consider that many indicators were mixed together, i.e, one sentence encapsulated two or more indicators. For instance, "We focus on identifying societal harms to the diversity of user communities impacted by our models (Google, 2024, p. 21) – one can identify both diversity and inclusiveness and safety and security," "Data security controls can be used to protect the data that AI systems use to train and operate (Google, 2023, p. 4) – one can identify data protection and awareness and literacy," "Microsoft will release an annual transparency report to inform the public about its policies, systems, progress, and performance in managing AI responsibly and safely (Microsoft, 2023, p. 23) – one can identify both transparency and awareness and literacy." These are only a few instances in which indicators overlap with each other, however, it was common to find

sentences under two or more subcategories from different dimensions, which posed a challenge for organising the data. This led to identifying a limitation within the method of analysis. The current focus is on analysing sentences within the data, following content analysis. To provide a more nuanced understanding of how these companies approach the principles outlined in the Recommendation – since most of the indicators are implicit – a discourse analysis approach could be more appropriate in future research to gather the meaning behind the data. In this way, one can get a more comprehensive understanding since the findings presented in this research are limited by specific indicators and the context of the data could not be evaluated. Considering not all indicators were identified in all four frameworks – besides safety and security, oversight and multi-stakeholder – questions for further research remain: What motivates private companies to adopt these principles? Why do they decide to emphasise these specific categories more than others? Another question arises from an identified area of convergence between Microsoft and Google: Red Teams. These are internal groups in charge of identifying and exploiting vulnerabilities in AI systems as a form of internal oversight. Future research could investigate the specific guidelines and operating procedures employed by these teams within tech companies, and how they compare to the oversight provided by external bodies.

7. Conclusions

This research reflects the diverse ways in which the UNESCO Recommendation on the Ethics of Artificial Intelligence has impacted AI corporate governance since it was issued. It aims to address a critical gap in the existing literature. Current literature addressing corporate AI corporate governance focuses on what they ought to implement following internal law rather than examining how private tech companies translate international regulations. To fill this gap, this thesis used a qualitative content analysis to examine frameworks issued by two different

leading tech companies, Google and Microsoft. This analysis revealed a wide range of areas in which the Recommendation has impacted AI corporate governance. Supporting the hypotheses based on RBA, all four frameworks addressed safety and security and multi-stakeholder collaboration. Furthermore, both companies addressed sustainability, a principle not considered at first but identified through the analysis process. These findings suggest that the Recommendation is acting as a shared set of rules for AI development and deployment, focusing on mitigating risks, fostering collaboration, and considering environmental and social impacts, therefore, following the expectations of RBA. At the same time, ethical assessment and promotion and protection of human rights were noticeable areas of impact, even though individually addressed in the frameworks, this can also be translated for the collective. This finding supports the expectations of HRBA. To conclude, this research demonstrates that the Recommendation has had impact on corporate AI governance frameworks in different areas, including safety and security, ethical assessments, protection and promotion of human rights, sustainability, awareness and literacy, and multi-stakeholder collaboration. Evidence supports the impact of the Recommendation on AI corporate governance. However, the extent to which Google and Microsoft address the various principles outlined in the Recommendation differs, for which further research is needed to explore the motivations behind these preferences and delve deeper into the nuances of how these principles are implemented within corporate AI governance practices. Since, the findings are drawn upon a small sample size, the generalizability and reliability of the findings might be limited. For that reason, future research could benefit from larger and more diverse case studies and sources, that could potentially reveal different results, or accept the argument presented in this thesis. All in all, this signifies a step in analysing the influence of international regulations on AI corporate frameworks.

8. Reference list

Aguerre, C., Campbell-Verduyn, M., & Scholte, J. A. (2024). Introduction: Polycentric Perspectives on Digital Data Governance. In C. Aguerre, M. Campbell-Verduyn, & J. A. Scholte (Eds.), *Global Digital Data Governance: Polycentric Perspectives*. Routledge. https://doi.org/10.4324/9781003388418

Ansell, C., & Torfing, J. (2022). *Handbook on Theories of Governance* (C. Ansell & J. Torfing, Eds.; 2nd ed.). Edward Elgar Publishing. https://doi.org/10.4337/9781800371972

Bullock, J., Chen, Y., Himmelreich, J., Hudson, V. M., Korinek, A., Young, M., & Zhang, B. (2022). Introduction. In *The Oxford Handbook of AI Governance*. https://doi.org/10.1093/oxfordhb/9780197579329.013.1

Chen, Y.-C., Ahn, M. J., & Wang, Y.-F. (2023). Artificial Intelligence and Public Values: Value Impacts and Governance in the Public Sector. *Sustainability*, *15*(6), 4796. https://doi.org/10.3390/su15064796

Cihon, P., Schuett, J., & Baum, S. D. (2021). Corporate Governance of Artificial Intelligence in the Public Interest. *Information*, *12*(275). https://doi.org/10.3390/info12070275

Council of Europe. (2024). *AI initiatives*. https://www.coe.int/en/web/artificial-intelligence/national-initiatives

Currie, N. (n.d.). *Risk based approaches to Artificial Intelligence*. Crowe. Retrieved December 10, 2020, from https://www.crowe.com/-/media/Crowe/LLP/folio-pdf/Risk-Approaches-to-AI.pdf

Dafoe, A. (2018). *AI Governance: A Research Agenda*. Future of Humanity Institute. fhi.ox.ac.uk/govaiagenda

Donahoe, E., & Metzger, M. M. (2019). Artificial Intelligence and Human Rights. *Journal of Democracy*, *30*(2), 115–126. https://doi.org/10.1353/jod.2019.0029

Erman, E., & Furendal, M. (2024). The democratization of global AI governance and the role of tech companies. *Nature Machine Intelligence*, *6*(3), 246–248. https://doi.org/10.1038/s42256-024-00811-z

Fukuyama, F. (2013). *What Is Governance?* Center for Global Development. http://www.cgdev.org/content/publications/detail/1426906

Gesley, J. (2020). Legal and Ethical Framework for AI in Europe: Summary of Remarks. *Proceedings of the ASIL Annual Meeting*, *114*, 241. Cambridge University Press. https://doi.org/10.1017/amp.2021.46

Google. (2023). *Secure AI Framework Approach A quick guide to implementing the Secure AI Framework (SAIF)*. https://kstatic.googleusercontent.com/files/00e270b1cccb1f37302462a162c171d86f293a84de54036e0021e2fe0253cf05623bae2a62751b0840667bc6c8412fd70f45c9485972dc370be8394fae922d31

Google. (2024). AI Principles Progress Update 2023. In *Google AI*. Google. https://ai.google/static/documents/ai-principles-2023-progress-update.pdf

Jones, K. (2023). AI governance and human rights: Resetting the relationship. *Royal Institute of International Affairs*. https://doi.org/10.55317/9781784135492

Linacre, S. (2024, February 6). *IBM leads Google and Microsoft as race to next generation AI heats up*. Digital Science. https://www.digital-science.com/news/ifi-claims-report-race-to-next-generation-ai-heats-up/

Lucey, S. (2022). What's artificial intelligence and what benefits does it deliver? In *What's artificial intelligence?* Australian Strategic Policy Institute. https://www.jstor.org/stable/resrep40313.5

Lütge, C., Hohma, E., Boch, A., Poszler, F., & Corrigan, C. (2022). White Paper On a Risk-Based Assessment Approach to AI Ethics Governance. *IEAI*. https://doi.org/10.13140/RG.2.2.13586.94406

Mahler, T. (2022). Between Risk Management and Proportionality: The Risk-Based Approach in the EU's Artificial Intelligence Act Proposal. *The Swedish Law and Informatics Research Institute*, 247–270. Nordic Yearbook of Law and Informatics. https://doi.org/10.53292/208f5901.38a67238

Malgieri, G., & Kamath, G. (2023). *Generative AI: Global Governance and the Risk- based approch*. Centre on Regulation in Europe.

Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022). Defining organizational AI governance. *AI and Ethics*. https://doi.org/10.1007/s43681-022-00143-x

Microsoft. (2022). *Microsoft Responsible AI Standard, v2* . Microsoft. https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2022/06/Microsoft-Responsible-AI-Standard-v2-General-Requirements-3.pdf

Microsoft. (2023). *Governing AI: A Blueprint for the Future*. Microsfot. https://query.prod.cms.rt.microsoft.com/cms/api/am/binary/RW14Gtw

Mucci, T., & Stryker, C. (2023, November 28). *What is AI governance?* IBM. https://www.ibm.com/topics/ai-governance

OECD. (2023). *G20/OECD Principles of Corporate Governance 2023*. OECD Pubishing. https://doi.org/10.1787/ed750b30-en

Risse, M. (2019). *Human Rights and Artificial Intelligence: An Urgently Needed Agenda*. *41*(1). https://doi.org/10.12957/publicum.2018.35098

Rodrigues, R. (2020). Legal and Human Rights Issues of AI: Gaps, Challenges and Vulnerabilities. *Journal of Responsible Technology*, *4*. ScienceDirect. https://doi.org/10.1016/j.jrt.2020.100005

Ruggie, J. G. (2014). Global Governance and "New Governance Theory": Lessons from Business and Human Rights. *Global Governance: A Review of Multilateralism and International Organizations*, *20*(1), 5–17. https://doi.org/10.1163/19426720-02001002

Su, A. (2022). The Promise and Perils of International Human Rights Law for AI Governance. *Law, Technology and Humans*, *4*(1). https://doi.org/10.5204/lthj.2332

Taeihagh, A. (2021). Governance of artificial intelligence. *Policy and Society*, *40*(2). https://doi.org/10.1080/14494035.2021.1928377

Tzimas, T. (2021). AI and Human Rights. In *Law, governance and technology series*. Springer International Publishing. https://doi.org/10.1007/978-3-030-78585-7_6

Ulnicane, I., Knight, W., Leach, T., Stahl, B. C., & Wanjiku, W.-G. (2021). Framing governance for a contested emerging technology:insights from AI policy. *Policy and Society*, *40*(2), 171. https://doi.org/10.1080/14494035.2020.1855800

UNESCO. (2021). *Recommendation on the ethics of artificial intelligence*. United Nations Educational, Scientific and Cultural Organization.

Wirtz, B. W., Weyerer, J. C., & Kehl, I. (2022). Governance of artificial intelligence: A risk and guideline-based integrative framework. *Government Information Quarterly*, *39*(4), 101685. https://doi.org/10.1016/j.giq.2022.101685

Xue, J. H. (2024). Polycentric Theory Diffusion and AI Governance. In C. Aguerre, M. Campbell-Verduyn, & J. A. Scholte (Eds.), *Global Digital Data Governance: Polycentric Perspectives*. Routledge. https://doi.org/10.4324/9781003388418

Yeung, K., Howes, A., & Pogrebna, G. (2020). AI Governance by Human Rights–Centered Design, Deliberation, and Oversight. In M. D. Dubber, F. Pasquale, & S. Das (Eds.), *The Oxford Handbook of Ethics of AI* (pp. 75–106). Oxford Academic Books. https://doi.org/10.1093/oxfordhb/9780190067397.013.5

9. Appendix

9.1. Coding frame matrix

Categories	Subcategories		Indicators	Code
Human rights-based approach	*Sustainabiliy*		The sentence discusses the impact of AI at the human, social, cultural, economic and environmental level The sentence mentions pursuing sustainable initiatives	HRS
	Protection and promotion of human rights	Right to Privacy	The sentence mentions respecting privacy	HRP
		Data Protection	The sentence mentions data protection The sentence mentions appropriate measures to collect and process personal/sensitive data The sentence mentions informed consent	HRDP
	Ethical assessment	Human oversight and determination	The sentence mentions human oversight The sentence mentions external or internal AI Ethics Committees/Entities	HRO
		Diversity and inclusiveness	The sentence mentions mentions representation of minorities and marginalised groups (disabled individuals, women, ethnic groups)	HRDI

			The sentence mentions the promotion of inclusive initiatives (language, gender, equality, battling stereotypes)	
		Transparency	The sentence mentions transparency requirements The sentence mentions informing users they are interacting with AI	HRT
Risk-based approach	Safety and security		The sentence mention risk/harm prevention and mitigation The sentence mentions safety and security measures (impact/risk assessments, reviews,) The sentence mentions proportinal measures adapted to different contexts	RSS
	Awareness and literacy		The sentence mentions informed decision The sentece mentions promoting AI education The sentence mentions accessible/public information	RAL
	Multi-stakeholder collaboration		The sentence mentions collaboration among different sectors (academia, NGOs, governments, companies)	RMSC

	The sentence mentions international cooperation The sentence mentions collaborative research

9.2. Coding text matrix

9.2.1. Google: "AI Principles Progress Update 2023"

Examples	**Code**
"As we consider potential development and use of AI technologies, we will take into account a broad range of social and economic factors, and will proceed where we believe that the overall likely benefits substantially exceed the foreseeable risks and downsides (p. 3)."	HRS
"We will seek to avoid unjust impacts on people, particularly those related to sensitive characteristics such as race, ethnicity, gender, nationality, income, sexual orientation, ability, and political or religious belief (p. 3)."	HRDI
"We will continue to develop and apply strong safety and security practices to avoid unintended results that create risks of harm (p. 3)."	RSS
"Our AI technologies will be subject to appropriate human direction and control (p. 4)."	HRO
"We will give opportunity for notice and consent, encourage architectures with privacy safeguards, and provide appropriate transparency and control over the use of data (p. 4)."	HRT HRP
"We will work with a range of stakeholders to promote thoughtful leadership in this area, drawing on scientifically rigorous and multidisciplinary approaches (p. 4)."	RMSSP

“Where there is a material risk of harm, we will proceed only where we believe that the benefits substantially outweigh the risks, and will incorporate appropriate safety constraints (p. 5).”	RSS
“Our AI Principles ethics reviews and impact assessments are part of a larger, end-to-end pre-launch process that includes technical safety testing and standard privacy and security reviews (p. 9).”	RSS
“Our risk assessment framework seeks to identify, measure, and analyze risks throughout the product development lifecycle. AI Principles reviews map these risks to appropriate mitigations and interventions, drawing upon our best practices from our cross-company enterprise risk management efforts.	RSS
“AI Principles reviews assess a range of harms, taking into account impacts ranging from unfair biases and stereotypes, poor product experiences, and social harms such as the spread of misinformation (p. 11).”	HRDI RSS
“In addition, as we’ve reported in detail in our 2022 AI Principles Progress Update, we engage external experts to conduct human rights impact assessments as appropriate (p. 11).”	RMSC RSS
“To guide product teams internally, we’ve established a framework to define the types of harmful content that we do not permit our models to generate. It also guides how we protect personal identifiable information (such as Social Security Numbers) (p. 11).”	RSS HRDP
“These harms can then be mitigated with the use of responsible datasets, classifiers and filters, and in-model mitigations such as fine tuning, reasoning, few-shot prompting, data augmentation, and controlled decoding to address potential harms proactively (p. 12).”	RSS

“Our second essential practice, adversarial testing, refers to systematic evaluation of a model by providing malicious or inadvertently harmful inputs across a range of scenarios to identify and mitigate potential safety and fairness risks (p. 12).”	RSS
“In addition to adversarial testing for safety and fairness, we’ve also established a dedicated Google AI Red Team focused on testing AI models and products for security, privacy and abuse risks (p. 13).	RSS
“Maintaining transparency documentation for developers, governments, and policy leaders is also key (p. 13).”	HRT
“This can mean releasing detailed technical reports or model or data cards that appropriately make public essential information based on our internal documentation of safety and other model evaluation details (p. 13).”	HRT
These transparency artifacts are more than communication vehicles; they can offer guidance for AI researchers, deployers, and downstream developers on the responsible use of the model (p. 13).”	HRT RAL
“By sharing the common risks that we find in our AI Principles reviews, we can offer transparency into our emerging best practices to mitigate these risks. These range from the technical, such as SynthID or About this image, tools we developed this year that can help identify mis- and dis-information when generative AI tools are used by malicious actors, to explainability techniques such as increasing explanatory information throughout the AI product, not just at the moment of decision. (p. 15).”	HRT RSS RAL
“We’re also researching the security benefits and risks of our largest model in the Gemini family of generative models. This has included scoping new	RSS RMSC

evaluation techniques, as well as joining relevant external fora, such as the UK's new Biosecurity Leadership Council (p. 15)."	
"We're committed to reporting specific capabilities, limitations, risks, and mitigations we've applied into our generative AI- powered systems, and contributing to shared industry standards on model transparency (p. 15)."	RSS HRT
"This year, we're piloting a transparency artifact specifically for the integration of research generative AI models into AI-powered systems. This artifact is called a generative AI system card. It builds upon our work of designing widely referenced and adopted transparency artifacts such as model and data cards (p. 15)."	HRT
"A key part of our ML work involves developing techniques to build models that are more inclusive (p. 19)."	HRDI
"We're developing methodologies to build models for people from a diversity of backgrounds (p. 19)."	HRDI
"We've made the Monk Skin Tone Examples (MST-E) dataset publicly available to enable AI practitioners everywhere to create more consistent, inclusive, and meaningful skin tone annotations as they create computer vision products that work well for all skin tones (p. 19)."	HRDI
"We have developed methods to account for rater diversity, and in the recent past, we've shared responsible practices for data enrichment sourcing (p. 20)."	HRDI
"We focus on identifying societal harms to the diversity of user communities impacted by our models (p. 21)."	RSS HRDI
"We also participate in external community engagement to identify "unknown unknowns" and to seed the data generation process (p. 21)."	RMSC

“To provide the high-quality human input required to seed the scaled processes, we partner with groups such as the Equitable AI Research Round Table (EARR), and with our internal ethics teams to ensure that we are representing the diversity of communities who use our models (p. 21).”	RMSC HRO HRDI
“We continue to expand our reach in terms of collaborating with underrepresented groups (p. 22).”	RMSC
“We’re committed to a global approach, so we gather feedback by collaborating with the international research community (p. 22).”	RMSC
“Analysis of language styles, including query length, query similarity, and diversity of language styles (p. 23).”	HRDI
“Multi-disciplinary AI research can help address society-level, shared challenges from forecasting hunger to predicting diseases to improving productivity (p. 25).”	RMSC HRS
“We identify primary and secondary indicators of impact that we optimized through our collaborations with stakeholders (p. 26).”	RMSC
“And we expanded our ongoing work in information literacy to support AI literacy (p. 28).”	RAL
“To address international frameworks and guidance for safe, secure, and trustworthy AI, we’re prioritizing cybersecurity safeguards (p. 30).”	RSS
“By making generative AI in Search first available through Search Labs, we were transparent that the technology was still in an early phase (p. 32).”	HRT
“We also try to let users know when they are engaging with a new generative AI technology and document how a generative AI service or product works (p. 32).”	RAL

“Our foundational privacy protections for giving users choice and control over their private data applies to generative AI (p. 34).”	HRP
“We’re committed to protecting your personal information (p. 34).”	HRDP
“Our most novel models are developed with scientific rigor and transparency (p. 34).”.	HRT
“In addition, we evaluate against multiple criteria and, as appropriate, with external reviews (p. 34).”	HRO
“Our AI Principles guide how we limit harms for people (p. 25).”	RSS
We engage in broad-based efforts — across government, companies, universities, and more — to help translate technological breakthroughs into widespread benefits, while mitigating risks (p. 36).”	RMSC RSS
“We outlined a three-pillared approach for governments to collaborate with the private sector, academia, and other stakeholders to develop shared standards, protocols, and governance so we can boldly realize and maximize AI’s potential for more people around the world (p. 37).”	RMSC

9.2.2. Google: “Secure AI Framework Approach: A quick guide to implementing the Secure AI Framework (SAIF)” –

Examples	**Code**
“This expands the composition of the team to include stakeholders across multiple organizations (p. 3).”	RMSC
“Existing security controls across the security domains apply to AI systems in a number of ways (p. 4).”.	RSS
For example, data security controls can be used to protect the data that AI systems use to train and operate (p. 4).”	RSS HRDP

“Organizations that use AI systems must have a plan for detecting and responding to security incidents and mitigate the risks of AI systems making harmful or biased decisions (p. 6).”	RSS
“This is because AI systems can be biased or make mistakes, and human oversight is necessary to ensure that AI systems are used ethically and responsibly (p. 7).”	HRO
“Red Team exercises are a security testing method where a team of ethical hackers attempts to exploit vulnerabilities in an organization's systems and applications. This can help organizations identify and mitigate security risks in their AI systems before they can be exploited by malicious actors (p. 8).”	RSS
“At the same time, it is essential to keep humans in the loop to oversee relevant AI systems, processes, and decisions (p. 8).”	HRO
“This means identifying all AI models in use, understanding the specific risks associated with each model, and implementing security controls to mitigate those risks along with having clear roles and responsibilities (p. 9).”	RSS
“Perform a risk assessment that considers organisational use of AI (p. 9).”	RSS
“This means understanding the specific risks associated with each AI use case and implementing security measures to mitigate those risks (p. 10).”	RSS

9.2.3. Microsoft: “Governing AI: A Blueprint for the Future”

Examples	**Code**
“Regulate through pre-release safety and security requirements, then license deployment for permitted uses in a licensed Al data center with post-deployment safety and security monitoring and protection (p. 18).”	RSS

“License for training and deployment of powerful AI models based on security protections, export control compliance, and safety protocols to ensure human control over autonomous systems that manage critical infrastructure (p. 18).”	RSS HRO
“Third, we will support broad educational initiatives to make information about AI technologies and responsible AI practices available to legislators, judges, and lawyers (p. 19).”	RAL
“To achieve safety and security objectives, we envision licensing requirements such as advance notification of large training runs, comprehensive risk assessments focused on identifying dangerous or breakthrough capabilities, extensive prerelease testing by internal and external experts, and multiple checkpoints along the way (p. 21).”	RSS RAL HRO
“Deployments of models will need to be controlled based on the assessed level of risk and evaluations of how well- placed users, regulators, and other stakeholders are to manage residual risks (p. 21).”	RSS RMSC
“Microsoft will release an annual transparency report to inform the public about its policies, systems, progress, and performance in managing AI responsibly and safely (p. 23).”	HRT RAL
“Microsoft will support the development of a national registry of high-risk AI systems that is open for inspection so that members of the public can learn where and how those systems are in use p. 23).”	RAL
“Microsoft will commit that it will continue to ensure that our AI systems are designed to inform the public when they are interacting with an AI system and that the system’s capabilities and limitations are communicated clearly (p. 23).”	HRT

“We believe that transparency is important not only through broad reports and registries, but in specific scenarios and for the users of specific AI systems (p. 24).”	HRT
Microsoft will continue to build AI systems designed to support informed decision making by the people who use them (p. 24).”	RAL
“We take a holistic approach to transparency, which includes not only user interface features that inform people that they are interacting with an AI system, but also educational materials, such as the new Bing primer, and detailed documentation of a system’s capabilities and limitations, such as the Azure OpenAI Service Transparency Note (p. 24).”	HRT RAL
“This is the need to provide broad access to AI resources for academic research and the nonprofit community (p. 24).”	RAL RMSC
“We will collaborate with the National Science Foundation to explore participation in a pilot project to inform efforts to stand up the National AI Research Resource, including by facilitating independent academic research relating to the safety of AI systems (p. 25).”	RMSC RSS
“Microsoft would welcome the opportunity to develop such practices by supporting and collaborating with a multistakeholder group, including representatives across the academic community (p. 25).”	RMSC
“At the highest level, the Environmental, Social, and Public Policy Committee of the Microsoft Board provides oversight of our responsible AI program (p. 31).”	HRO
“They break down a broad principle like accountability into definitive outcomes, such as ensuring AI systems are subject to impact assessments, data governance, and human oversight (p. 32).”	RSS HRO

“We then ask teams to measure the prevalence of those harms and mitigate them by testing and implementing various tools and established strategies (p. 33).”	RSS
“Our Sensitive Uses program provides an additional layer of oversight for teams working on higher-risk use cases of our AI systems (p. 33).”	HRO
“For particularly high-impact or novel-use cases, we elevate the project for review and advice from our Sensitive Uses Panel, which is a group of Microsoft experts spanning engineering, research, human rights, policy, legal, and customer-facing organizations from around the world (p. 34).”	HRO RMSC
“We ask teams who develop and use AI systems to look at technology through a sociotechnical lens. This means we consider the complex cultural, political, and societal factors of AI as they show up in different deployment context (p. 36).”	HRS
“As part of this commitment, we provide transparency documentation for our platform AI services in the form of Transparency Notes to empower our customers to deploy their systems responsibly (p. 39).”	HRT
Transparency Notes communicate in clear, everyday language the purposes, capabilities, and limitations of AI systems so our customers can understand when and how to deploy our platform technologies (p. 39).”	RAL

9.2.4. Microsoft: “Microsoft Responsible AI Standard, v2”

Examples	**Code**
“We believe that industry, academia, civil society, and government need to collaborate to advance the state-of-the-art and learn from one another (p. 3).”	RMSC
“Microsoft AI systems are assessed using Impact Assessments (p. 4).”	RSS

“Microsoft AI systems are reviewed to identify systems that may have a significant adverse impact on people, organizations, and society, and additional oversight and requirements are applied to those systems (p. 5).”	RSS HRO
“Provide documentation to customers which describes the system’s: 1) intended uses, and 2) evidence that the system is fit for purpose for each intended use (p. 6).”	RAL
“When the system is a platform service made available to external customers or partners, include this information in the required Transparency note (p. 6).”	RAL HRT
“Define and document procedures for the collection and processing of data, to include annotation, labelling, cleaning, enrichment, and aggregation, where relevant (p. 7).”	HRDP
“Document these stakeholders and their oversight and control responsibilities using the Impact Assessment template (p. 8).”	HRO RSS
“Publish documentation for the system so that stakeholders can understand the system (p. 11).”	RAL
“Microsoft AI systems are designed to inform people that they are interacting with an AI system or are using a system that generates or manipulates image, audio, or video content that could falsely appear to be authentic (p. 12).”	HRT
“Microsoft AI systems are designed to provide a similar quality of service for identified demographic groups, including marginalized groups (p. 13).”	HRDI
“Use Analysis Platform to understand the representation of identified demographic groups in data sets that you plan to use for training and evaluating your system, respecting privacy controls for working with sensitive data (p. 15).”	HRP

“Microsoft AI systems that allocate resources or opportunities in essential domains are designed to do so in a manner that minimizes disparities in outcomes for identified demographic groups, including marginalized groups (p. 16).”	HRDI
“Determine and document the operational factors, including quality of system input, use, and operational context that are critical to manage for reliable and safe use of the system in its deployed context (p. 21).”	RAL RSS
“Microsoft AI systems are designed to protect privacy in accordance with the Microsoft Privacy Standard (p. 26).”	HRP
“Microsoft AI systems are designed to be secure in accordance with the Microsoft Security Policy (p. 26).”	RSS
“Microsoft AI systems are designed to be inclusive in accordance with the Microsoft Accessibility Standards (p. 27).”	HRDI