AI & SOCIETY (2022) 37:1663–1687
https://doi.org/10.1007/s00146-021-01267-0

OPEN FORUM

Operationalising AI ethics: how are companies bridging the gap between practice and principles? An exploratory study

Javier Camacho Ibáñez[1] · Mónica Villas Olmeda[2]

Received: 29 March 2021 / Accepted: 12 August 2021 / Published online: 31 August 2021

Abstract

Despite the increase in the research field of ethics in artificial intelligence, most efforts have focused on the debate about principles and guidelines for responsible AI, but not enough attention has been given to the "how" of applied ethics. This paper aims to advance the research exploring the gap between practice and principles in AI ethics by identifying how companies are applying those guidelines and principles in practice. Through a qualitative methodology based on 22 semi-structured interviews and two focus groups, the goal of the current study is to understand how companies approach ethical issues related to AI systems. A structured analysis of the transcripts brought out many actual practices and findings, which are presented around the following main research topics: ethics and principles, privacy, explainability, and fairness. The interviewees also raised issues of accountability and governance. Finally, some recommendations are suggested such as developing specific sector regulations, fostering a data-driven organisational culture, considering the algorithm's complete life cycle, developing and using a specific code of ethics, and providing specific training on ethical issues. Despite some obvious limitations, such as the type and number of companies interviewed, this work identifies real examples and direct priorities to advance the research exploring the gap between practice and principles in AI ethics, with a specific focus on Spanish companies.

Keywords Artificial intelligence · Ethics · Privacy · Explainability · Fairness · Principles

1 Introduction

AI ethics is a subfield of applied ethics focusing on the ethical issues raised in the development, deployment, and use of AI (European Commission 2019). AI ethics is generally concerned with how the AI industry can minimise ethical harms; less frequently mentioned is how the AI industry can maximise its potential benefits. This concern has already led to the development of ethical principles and guidelines (Winfield et al. 2019). In recent years, most of the countries and companies worldwide that are developing AI technologies have established and published ethical principles and guidelines to be followed by organisations in the development of AI systems.

This paper will seek to identify how companies apply those guidelines and principles and to further understand how they work to bridge the gap between practice and such principles and guidelines (Whittlestone et al. 2019). A qualitative methodology based on semi-structured interviews with several companies was designed. Through a detailed review of the existing literature on principles and a review of similar studies, the goal of the current study is to understand how the interviewed companies approach ethical issues related to AI systems and apply guidelines and principles, with particular attention to the topics of privacy, explainability, and fairness.

After a consistent review of the available literature on the subject and analysing eight similar studies, we find that there is still a gap between theory and practice in the field. The focus of this research is on four main questions regarding how companies apply AI ethics principles in practice, their perceptions about privacy, explainability and fairness issues, and how they try to approach them. Twenty-two companies were selected, and top- or senior-level managers with an

✉ Javier Camacho Ibáñez
jcamacho@comillas.edu

Mónica Villas Olmeda
mvillas9@alumnos.uned.es

[1] Universidad Pontificia Comillas, Madrid, Spain

[2] Universidad Nacional de Educación a Distancia (UNED), Madrid, Spain

average of over 16 years of experience in AI were interviewed. Two focus groups were also conducted at a second stage to refine the findings. A structured analysis of the transcripts identified many actual practices, insightful findings, and recommendations, which are detailed in Sect. 5. Despite some obvious limitations, such as the type and number of companies interviewed, the paper provides real examples and identifies direct priorities to advance understanding with regard to the gap between practice and principles in AI ethics. This is likely the only study to date with Spanish companies and provides insightful input given the potential social impact of AI applications and the recent EU regulations concerning AI ethics.

2 Objective and justification: from principles to practice

2.1 Analysis of the situation

This paper aims to advance the research exploring the gap between practice and principles in AI ethics. The starting point was a comprehensive analysis of the documentation of guidelines and principles reflecting that there have been many studies regarding ethical principles, codes, guidelines, and frameworks. Over 70 recommendations (Floridi 2019) were identified based on a combination of sources from academia, industry, and government (Morley et al. 2020). However, very few show the principles' practical implementation to focus on actionable items inside the organisations (Canca 2020; Floridi 2019; Hickok 2021; Mark and Anya 2019; Stahl et al. 2021; Vakkuri and Kemell 2019). There is a need to link AI principles, business processes, and data governance with a focus on specific procedures for both organisations and developers (Abrams et al. 2019; McNamara et al. 2018; Morley et al. 2020; Vakkuri and Kemell 2019). It is not an easy task to extract norms and requirements from high-level guidelines. Specific elements of technology, applications, the context of use and local regulations must be considered. In addition, a roadmap is required for how to move from principles to practice with a focus on ethical governance (Eitel-Porter 2020; Mittelstadt 2019; Winfield et al. 2019).

Consequently, despite the body of ethical guidelines and principles developed in recent years, almost none offer details on their practical deployment, which indicates a clear gap between those principles and their concrete application (Mittelstadt 2019). Most such efforts have been placed on the "what" of ethical AI, i.e. debates about principles and guidelines for a sort of responsible AI; however, not enough attention has been given to the "how" of applied ethics (Morley et al. 2020). Therefore, ethics operate at a considerable distance from the practices it seeks to govern (Morley et al. 2021). However, AI principles should not be mere words as they are intended to guide practice (Hagerty and Rubinov 2019). There is a need to delve into the specific practical approaches to sectors and applications. Different situations may require a different understanding of how to maintain ethical or moral control (Fabre et al. 2021).

Regarding proper AI ethics implementation, tensions in society may be produced for the stakeholders involved such as industry actors, developers, academics, government officials, and end-users (Floridi et al. 2018; Whittlestone et al. 2019). For example, a data-driven society might offer advantages, but doubts may emerge concerning data privacy protection. Therefore, it is crucial to continue to approach all stakeholders in society affected by AI ethics. There have been several initiatives to address this issue such as AI HLEG 2019 in Europe, PANNELFIT (Participatory approach of a new ethical and legal framework for ICT), SHERPA (Shaping the Ethical dimension of Smart Information System), and SHIENA (Stakeholder-informed ethics for new technologies with high socioeconomic and human rights impact). These initiatives have successfully addressed social issues (Daly et al. 2020; Fernow et al. 2019). The case of the European AI regulation published in April also served as a good baseline. However, AI ethics is progressing differently in different countries or regions such as Australia, China, India, the US and the European Union (Daly et al. 2020). Therefore, the operationalisation of AI principles and impacts might be uneven from a geographical perspective, adding another variable to this complex issue (McDuie-Ra and Gulson 2020).

There are also some cases where ethics seems to be integrated into companies or institutions merely as a marketing strategy (Bietti 2020; Canca 2020; Charisi et al. 2017). However, companies need to face the ethical consequences of AI development. Several recent examples in the market have not considered AI ethics (Amazon[1] or Cambridge Analytica[2]) that, apart from potential legal consequences, have caused major reputational issues. Therefore, companies must deal with these ethical AI challenges, which can be summarised as affecting the areas of compliance, governance, and transparency (Eitel-Porter 2020; Hagendorff 2020).

While organisational-level policies and guidelines may aim to direct development work, many decisions are left to individual developers. Developers working with AI need to identify the ethical dimension in the systems they build. They should not consider only their datasets and how they are selected (Gebru et al. 2018) but also how they process, use and trace them (de Bruin and Floridi 2017) and how their algorithms are coded or implemented (Kitchin 2016).

[1] See Dastin (2018).

[2] See Confessore (2018).

Table 1 Similar studies Source: developed by the authors

References	Interviewees	Sector/country	Key findings
Vakkuri et al. (2019)	Eight interviews with software developers, data analysts	Health care/Finland	Transparency Lack of tools
Vakkuri (2020)	249 surveyed	Software development companies/ US/Finland/others	Focus on ART (accountability, responsibility and transparency)
Kevin and Ana (2019)	One interview	Telco/Scandinavia	Privacy
Mark and Anya (2019)	16 interviews	Four sectors/Nordics and The Netherlands	Need guidance and policies applied to Smart cities Privacy and transparency
Rothenberger et al. (2019)	Seven interviews and 51 surveyed	ICT industry	Transparency, responsibility, data privacy, bias, robustness
Stahl et al. (2021)	42 interviews	Eight sectors/Netherlands, Germany, UK, and Finland	Organisational awareness, privacy, human oversight (trust and accountability) Training in ethics Additional stakeholders
Taylor and Dencik (2020)	Eight interviews academia, NGO and industry	Telecom and data/the UK, the Netherlands, and Germany	Responsibility from individual to collective ethics
Orr and Davis (2020)	21 interviews	Australia	Distributed responsibility

Developers are generally aware of the importance of these issues and try to implement a certain degree of transparency in these systems. However, there are still very few market tools for implementing ethics (Eitel-Porter 2020; McNamara et al. 2018; Morley et al. 2021; Vakkuri 2020). One of the most recent developments is the responsible AI toolkit by PwC.[3] It consists of different frameworks and tools to facilitate the understanding of AI ethics maturity in a company. In addition to the need for different tools and frameworks, AI ethics requires legal enforceability, an issue that remains unresolved in the majority of countries around the world (Daly et al. 2020). Without legal enforceability, ethical guidelines per se do not tend to change professional behaviour in technical communities (Hagendorff 2020; Mittelstadt 2019). A developer-centric approach to ethics in AI is relevant (Vakkuri and Kemell 2019); however, it is also necessary to obtain an understanding of the perceptions and needs of other stakeholders (Orr and Davis 2020).

2.2 Similar studies

Some studies have followed a similar approach to understand how companies try to bridge principles and practice (see Table 1). Vakkuri et al.(2019) focuses on five companies and the various technical profiles of those who work with AI. The issues covered are transparency of development, accountability and responsibility. The main conclusion is that the gap between guidelines and practice is real. AI developers bear most of the responsibility for AI ethics; in some cases, they do not have the proper knowledge to apply these concepts, nor do they possess the right tools or knowledge of the necessary processes. More recently, Vakkuri (2020) surveyed 249 participants, mainly in Finland and the US, to understand the implementation of AI ethics in different companies, with a focus on the technical profiles of these enterprises. The study's conclusion was that AI ethics principles are not yet fully deployed in the industry. The report mentions a lack of regulation, except for GDPR, and highlights the need to include AI ethics checks in the course of development. The study concludes with recommendations that might be helpful such as following a systematic approach to prevent AI ethics issues or improving transparency.

Kevin and Ana (2019) focus on customer relationship management (CRM) systems, highlighting the company's responsibilities in data collection and how those data are used. It also refers to the potential limitation of the GDPR in terms of developing specific algorithms in Europe that can be done without any issues in other regions. Mark and Anya (2019) conducted a study with four organisations regarding smart cities, where consent, transparency and data ownership are again highlighted as the primary ethical considerations. With regard to privacy, when considering the nature of a smart city application, privacy is one of the fundamental principles to bear in mind in developing these kinds of solutions. Transparency is also considered to be a critical component in generating citizens' trust.

Rothenberger et al. (2019) carry out seven qualitative interviews to rank principles inside companies and obtain the following: responsibility, data privacy, transparency, and

[3] https://www.pwc.com/gx/en/issues/data-and-analytics/artificial-intelligence/what-is-responsible-ai.html.

Table 2 Studies on guidelines and principles in AI ethics Source: developed by the authors

References	Set of principles
Jobin et al. (2019)	Transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, sustainability, dignity, solidarity
Greene et al. (2019)	Universal concerns objectively measured, expert oversight, values-driven determinism, design as the locus of ethical scrutiny, better building, stakeholder-driven legitimacy and machine translation
European Commission (2019)	Human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, environmental and societal well-being and accountability
Fjeld et al. (2020)	Privacy, accountability, safety and security, transparency and explainability, fairness and non-discrimination, human control of technology, professional responsibility, promotion of human values
Morley et al. (2020)	Justice, autonomy, explicability, fairness, robustness, security, safety, transparency, accountability
Eitel-Porter (2020)	Fairness, accountability, transparency, explainability, privacy

robustness. Stahl et al. (2021) develop 10 case studies with 42 interviews and note that companies are aware of the issue but have typically developed limited mitigation strategies focusing only on organisational awareness, analysis of privacy, ethics training and human oversight.

In a series of eight interviews, Taylor and Dencik (2020) seeks to determine companies' definitions of data ethics. The study concludes by identifying the need to reframe the question, include additional stakeholders in future analysis, and raise the topic of moving ethics from an individual to a collective anchoring for ethics. Finally, Orr and Davis (2020) performed 21 interviews with AI practitioners in Australia. The findings highlight that practitioners play a crucial role in AI ethics but that it should be more of a collective task.

Based on the analysis of these studies, there is a need to continue researching the gap between principles and practices focusing on concrete cases and specific scenarios. Furthermore, no studies consider companies in Spain.

3 Theoretical framework and research questions

Several studies have reviewed the various initiatives, guidelines, and principles (Brundage et al. 2020; Fjeld et al. 2020; Greene et al. 2019; Whittlestone et al. 2019). Through research from the private sector, government, civil society and other stakeholders, those studies have selected AI principles documents from relevant sources, regions and sectors. There seems to be a common language of ethical concern (Greene et al. 2019) and some degree of global convergence around a set of ethical principles. Table 2 presents a non-exhaustive summary of some of these commonly agreed principles.[4] There is still an open discussion about how those principles should be understood in terms of culture, linguistics, geography and organisational context (Hickok 2021; Rothenberger et al. 2019). AI principles are likely to be translated linguistically and culturally in different ways in different regions, as might happen with concepts of "fairness" or "privacy" (Hagerty and Rubinov 2019).

As shown in Table 2, there is some sort of convergence towards a set of overarching principles such as transparency, explainability, accountability or privacy (Daly et al. 2020). For this research, we have chosen three main principles: privacy, explainability and fairness. This is because more than 90% of the documents mention these principles (Fjeld et al. 2020), and they are explicitly mentioned in the AI HLEG EU documents (European Commission 2019). This factor was also considered to be relevant in the selection of the principles as this study was conducted in Europe.

AI systems should respect individuals' **privacy**, both in the use of data to develop technological systems and to provide people with control over their data and the decisions they make with it. The individual should control their data usage, the context in which their data are processed, and how it is activated. **Explainability** articulates the requirements that AI systems should implement to allow for oversight, which include translating operations into intelligible outputs and providing information about where, when, and how they are used. **Fairness** and non-discrimination call for AI systems to be designed and used to avoid biases and prejudices and promote inclusivity. Particular attention should be given to people who are particularly vulnerable to profiling that may adversely affect their rights and control or expose them to discrimination or stigmatisation, for example, due to their financial, social or health-related conditions (Fjeld et al. 2020; Data Ethics 2021).

Based on this framework, we developed the following research questions:

Q1. How do companies try to apply the existing guidelines or principles for responsible AI?
Q2. How do companies consider privacy, and what specific measures (if any) do they take with regard to this issue?

[4] For an updated list of principles, please check https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/.

Q3. How do companies consider explainability, and what specific measures (if any) do they take with regard to this issue?
Q4. How do companies consider potential biases and fairness, and what specific measures (if any) do they take with regard to this issue?

4 Methods

4.1 Justification of the methodology

A qualitative methodology approach may be appropriate given the gap between theories or principles and their practical application (Canca 2020). AI ethics is a complex topic (Edwards et al. 2018). Therefore, it is important to understand the practical needs, perceptions and realities of different stakeholders (Orr and Davis 2020). Therefore, a qualitative methodology based on interviews and focus groups was developed for this study's exploratory purpose. Similar studies in the field have also used interviews (Govia 2020; Mark and Anya 2019; Orr and Davis 2020; Rothenberger et al. 2019; Taylor and Dencik 2020; Vakkuri et al. 2019; Watson et al. 2021).

The combined use of interviews and focus groups reinforces each participant's usefulness. Interview results can provide helpful information for developing focus group guidance; similarly, focus groups can enrich interview results by facilitating discussion and interaction on shortlisted topics (Morgan 1997).

4.2 Development of the scripts

An interview script should consist of understandable and short questions that convey the research's central topics (Vallés 2009). Based on the theoretical framework described above, an interview script was developed grouping the topics around the main research questions. The goal was to ask the interviewees about some recent projects that they consider to have an ethical impact; their specific knowledge of the principles and guidelines; specific ethics milestones in project development; and their concerns regarding privacy, explainability, or fairness. We opted for a semi-structured focused interview model to help the interviewees contribute their perspectives on the issues raised (Marshall and Rossman 1995). The interview script can be found in Appendix 1.

The main goal of a focus group is to help interpret and contrast interview results by facilitating debate and interaction on the selected topics, furthered by the creation of an adequate trustworthy climate (Krueger and Casey 2000). Therefore, the script for the focus groups was developed after the interviews were completed and analysed, and a short introduction document was shared with the participants. The questions used during the discussion groups were open and easy to formulate; they were presented with a conversational tone and kept simple (Krueger and Casey 2000). The focus group script can be found in Appendix 2.

4.3 Selection of the participants

Given the objective pursued in this project, we searched for companies developing artificial intelligence projects in Spain to interview senior executives or technical officers with responsibility and senior experience. These profiles usually work as connecting hubs between stakeholders such as customers, business and technical teams, other areas within the company, and technology providers (Orr and Davis 2020). Therefore, we considered that they might provide a valuable perspective for this study. The interviewed participants had an average of 16.8 years of experience in artificial intelligence (see Table 3).

A total of 140 companies were identified using public information.[5] These companies were researched, and some were discarded because of the lack of contact data or because they were merely an agency of an international company with no actual activity or product development in Spain. We finally contacted 44 companies according to their type of activity and size with the aim of sampling corporate companies and start-ups. After verifying the person to be interviewed and their initial willingness to collaborate in the investigation, we sent a letter of introduction and a request for participation explaining the purpose of the investigation, the intended content of the interview, the conditions of anonymity and confidentiality, the estimated duration, and a range of dates.

A total of 22 live interviews were carried out between May and September 2020 (50% of the contacted companies). This number of interviews is aligned with similar studies (Orr and Davis 2020; Rothenberger et al. 2019; Taylor and Dencik 2020; Watson et al. 2021).

The following figures (Figs. 1 and 2) show the distribution of companies interviewed by the number of employees and type of activity.

Once the interviews were analysed and the first draft for discussion was elaborated, two focus groups were held with ten interviewed participants to contrast the findings and gather further feedback.

4.4 Quality criteria

To ensure the research quality as much as possible, elements of credibility, reliability and ethical criteria were applied.

[5] For example, https://mapa.estrategiaia.es/mapa, that refers to an updated list of organisations related to AI in Spain.

Table 3 List of companies interviewed Source: developed by the authors

ID	Role of person(s) interviewed	Years of experience in AI	Sector/activity	Number of employees*
1	CEO	10	Product development	Less than 25
2	AI senior product manager	20	Product development	Between 25 and 100
3	Business analyst	8	Technology consultancy services	Less than 25
4	Head AI product development	18	Technology consultancy services	More than 2000
5	Founder and CEO	10	Product development	Less than 25
6	Lead data architect	10	Banking	Between 500 and 2000
7	AI strategist advisor	24	Product development	Between 25 and 100
8	AI leader	15	Technology consultancy services	More than 2000
9	Chief data officer	15	Energy and utilities	Between 500 and 2000
10	Business development	14	Product development	Between 25 and 100
11	Director of AI	14	Technology consultancy services	Between 100 and 500
12	AI consulting director	25	Technology consultancy services	Less than 25
13	AI senior product manager	8	Product development	Between 25 and 100
14	Senior AI tech lead	12	Technology consultancy services	Between 100 and 500
15	Director of AI	25	Technology consultancy services	Between 100 and 500
16	AI senior product manager	20	Product development	Less than 25
17	COO	20	Product development	Between 25 and 100
18	CEO	15	Product development	Between 100 and 500
19	Head of AI	20	Technology consultancy/cybersecurity	Between 500 and 2000
20	Head of AI	20	Technology consultancy services	More than 2000
21	General manager Europe	20	Insurtech	Between 100 and 500
22	Head of advance analytics	20	Banking	More than 2000

*Note on the number of employees: this number should be considered with caution as it does not necessarily indicate, especially in large companies, the number of people engaged in AI-related tasks and projects

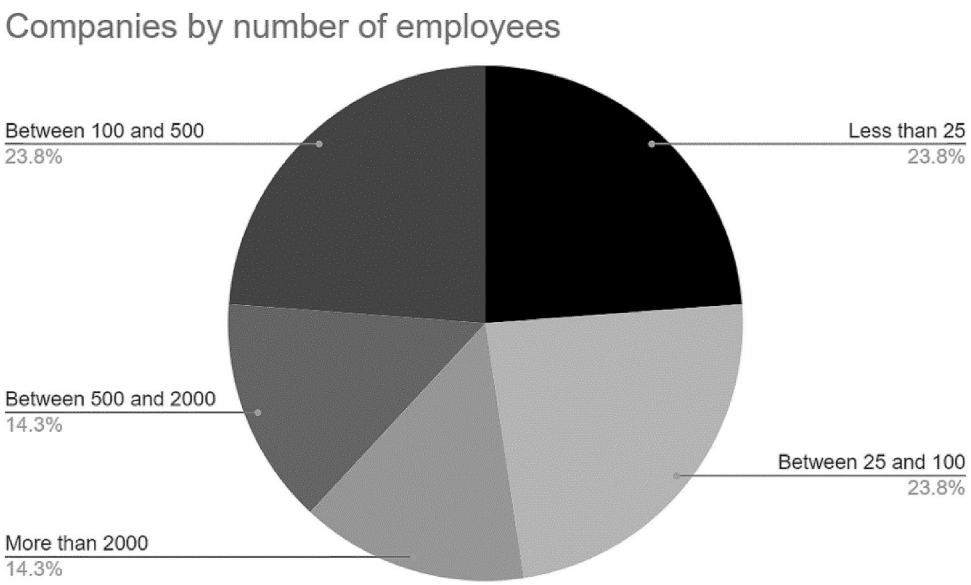


Fig. 1 Interviewed companies by the number of employees Source: developed by the authors

Advanced notice was given with regard to the research goal and process. The selection criteria were based on knowledge of the subject by the interviewees. We tried to anticipate some of the possible obstacles during the interviews and focus groups, such as the ambiguity or excessive scope of some of the issues raised or the bias of induced responses for social convenience. As much as possible, the questions were formulated precisely and concretely,

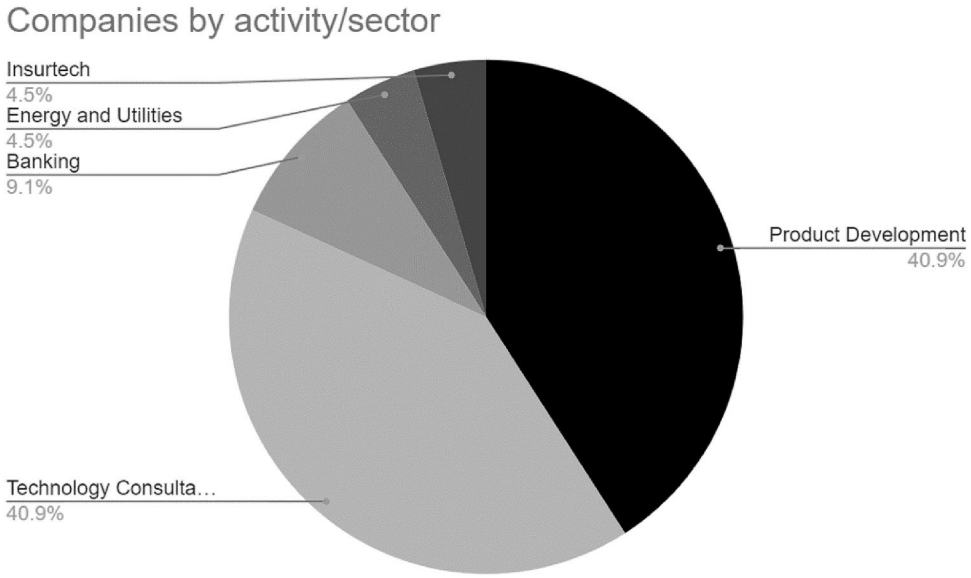


Fig. 2 Interviewed companies by activity/sector Source: developed by the authors

concerning events experienced by the interviewee and avoiding direct questions, to limit answers based on opinions, beliefs, or social desirability effects. There was also one pilot interview to ensure comprehension of the questions, timing, and possible issues.

Once the interviews were completed, transcripts were sent to the interviewees requesting acknowledgement and comments or amendments and confirming that their content would not be published in full. For this reason, the transcripts are not included in this paper. During the focus group sessions, the participants could comment on the different findings and citations. During that stage, there was no contradiction to the discussion notes.

Interrater reliability calculations were not performed as only one person encoded the data; however, the coding phase was performed iteratively. A sort of methodological triangulation was obtained since the results of the interviews were used and discussed with the participants of the discussion groups. The interviews were analysed, which produced a first coding of the terms found, and these results were used to prepare the script of the discussion groups to confirm these results. Subsequently, with the transcripts and notes of the groups, a second coding and analysis were carried out, which included a review of the codes and groups of codes previously identified. A second researcher oversaw the whole process.

The project is based on the fundamental criteria of informed consent, confidentiality of information, and the anonymity of both the participant and the company.

4.5 Registration and data analysis

The interviews and the focus groups were recorded on video (Google Meet). Automatic transcription was subsequently performed using the Sonix.ai provider, which was reviewed between two researchers, to obtain a verbatim transcription. Additionally, notes taken during interviews were reviewed.

For data analysis, the ATLAS.ti 9 software was used. The use of specific software for qualitative analysis helps store and organise data, make connections between text fragments, conceptualise different elements, plot a graph of topics and codes, and store notes (Creswell 2012). The analysis was carried out considering the main objective of the research (Krueger and Casey 2000). The process was divided into three stages: exploration, encoding and data reduction, and interpretation (Marshall and Rossman 1995).

The versions of the transcripts sent, accepted, or modified with the participants' comments were used for the exploration phase. In the exploratory phase, the data were read several times to promote familiarity with the content of the transcripts and add comments according to the field notes collected. Codes for analysis also began to be outlined, depending on the issues raised in the responses, always in consideration of the theoretical framework and the research questions.

Coding refers to the process of assigning categories, concepts, or codes to segments of information that are of interest to the research objectives. Defining, selecting, grouping, and reflecting on the codes was carried out iteratively. A first version was established using relevant quotations regarding the research questions from the transcribed interviews. Once the first coding was done, the codes were reviewed to find definitions and similar codes, which were combined, grouped, or simplified. We iterated this process until all relevant quotations fit into an existing code or group of codes and no other categories could be identified (Braun and Clarke 2006). Networked graphical representations of code groups

were also drawn to help visualise interrelated concepts (see Appendix 4).

The concept of saturation can be defined as "the point at which gathering more data about a theoretical construct reveals no new properties nor yields any further theoretical insights" (Bryant and Charmaz 2007, p. 611). It is usually difficult to assess (Mason 2010); however, we consider both code saturation and theme saturation to have been reached in our study. Obtaining a certain degree of homogeneity in the participants and identifying concrete topics contribute to achieving saturation with a relatively small sample (Hennink et al. 2016).

Coding was an iterative and collaborative process between the authors, who jointly identified the relevant topics and carried out the exploratory phase. One author coded the complete set of interviews while the other author double-checked the coding and the corresponding quotations. This type of iterative and reflective process contributes to the consistency of the findings (Krefting 1991). Once the iterative process was completed, 194 codes and 11 code groups were obtained (see Appendix 3).

Once the processes of classification and purification of the codes and groups were finished, we proceeded with the interpretation phase to examine each group in detail, analysing the fragments belonging to each of them and the codes of each group, to find common patterns, relate the data, and elaborate the interpretations. Networked graphic representations of groups and codes were also created, which is very useful in identifying the interrelationship of concepts (see Appendix 4 for the network diagrams).

5 Results and discussion

This section summarises the main findings resulting from the qualitative analysis. The findings (F) are related to each of the research questions:

Q1. How do companies try to apply the existing guidelines or principles for responsible AI?
Q2. How do companies consider privacy, and what specific measures (if any) do they take with regard to this issue?
Q3. How do companies consider explainability, and what specific measures (if any) do they take with regard to this issue?
Q4. How do companies consider potential biases and fairness, and what specific measures (if any) do they take with regard to this issue?

One additional category is proposed that is derived from the analysis: accountability and governance. Additionally, at the end of this section, some recommendations (R) derived from the study are suggested.

Table 4 Main findings on ethics and principles Source: developed by the authors

F1. Companies are more concerned about regulation than ethics
F2. Each sector and application might have a different ethical impact
F3. The gap between practices and principles is recognised
F4. Effect on employment is an ethical issue raised by some stakeholders

5.1 Ethics and principles

The goal of the first research question was to determine how companies apply the guidelines for responsible or ethical AI. The main findings are summarised in Table 4 below:

F1. Companies are more concerned about regulation than ethics.

From the point of view of companies, regulatory concerns are more relevant than ethics. Companies, in general, have low interest in ethical questions and are much more concerned about business goals and business performance. The other concern for customers is potential reputational risks.

Quote #11[6]
There is no perceived ethical concern, yes for regulatory concern […]
I have not found any RFP case that says, "this algorithm has to be explainable or has to avoid some type X of bias, for example, gender bias" I have not found it.
Quote #2
The questions companies ask me about have more to do with their use cases or business cases and how to comply with the regulations in force.
Quote #12
I would say that there is almost more interest from a reputational perspective and from a perspective of realising that this is a problem that increasingly impacts public opinion. Then it might be of interest to the company.

This concern about regulation seems to some extent to support the approach of "enforcing" AI ethics through oversight by institutions to monitor regulatory requirements (Yeung et al. 2019), and the European Union recently proposed regulation.[7]

F2. Each sector and application might have a different ethical impact.

[6] The number after "Quote" refers to the company ID (Table 3) from which the actual quote has been drawn.

[7] https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1623335154975&uri=CELEX%3A52021PC0206.

However, it is very relevant to contextualise the sector and type of application when establishing AI ethical use criteria.

Quote #12
So, I think what is missing is a regulatory environment with a different approach, where not everything is the same. Not all artificial intelligence, not all algorithms, not all use cases are the same, much less.
Quote #20
I think it depends quite a bit on the industry's level of maturity from my experience. Some industries have direct involvement with more evident ethical principles such as financial services or insurance.

In general, technological consultancy and product development companies perceive the ethical dimension of projects. These organisations are aware of the opportunities and risks and can visualise the ethical dimension of applications. These companies mention some intrinsic principles for ethical behaviour and ethically responsible action, such as respect for individuals' self-determination, justice, or trust, reinforcing awareness about the ethical dimension within their projects.

Quote #14
Every time we decide that we want to start exploring a particular front, we, and any company, should check if what is being developed may have some ethical impact and mitigate it. Yes, especially when defining strategic decisions, let's say, new products, new lines of work, etc. It is important to keep that in mind.
Quote #3
I am worried about that invasion, that personal invasion, even to the subconscious, of no longer knowing whether the decisions are yours or not. […] But companies that deal with personal data should know that they are people who are behind them and that you have to respect them ... behind each [individual's] personal data, there is a person.

F3. The gap between practices and principles is recognised.

A gap between principles and their implementation, or vice versa, is recognised. Expectations about the operability of principles are not very high. The widespread view is that it is necessary to define principles by sector, type of application or project.

Quote #22
They sent two hundred pages of what it should be today from the European Union, but then in reality, what can be applied? What is the reality of companies, and what is practical?
Quote #14
I think we read them all because they are coming out. There are many in the "stratosphere". That is when you read the principles and say, how do I translate them in practice? It gets more complicated.
Quote #7
There is much practice but few principles.

It is also worth noting that the knowledge of the principles is motivated more by a matter of personal concern or experience than by formal communication at the organisational level. In most cases, the knowledge of the principles is generic as they are identified as those of the European Union (in 5 interviews) and others mentioned in the interviews such as Asilomar, Montreal, Nilsson, and Google.

Quote #3
[…] the typical ones, I know them but on a personal level and [based on] curiosity.
Quote #11
The White Paper of the European Commission on Artificial Intelligence, well, I know it in my case, but it is also a bit because of my profile and past.

F4. Effect on employment is an ethical issue raised by some stakeholders.

An additional issue that consultants and developers often encounter is ethical questions about the impact on present or future employment. This issue goes beyond most of the guidelines and principles, but it is important to be considered from a social perspective and its implications.

Quote #12
Artificial intelligence and work, how artificial intelligence changes the future of employment, how that has to do with these evolutions that we are experiencing, the culture of leisure, the working days of three or four days, the distribution of work, the basic income.
Quote #18
We are sometimes asked the question of, well, you're using artificial intelligence to destroy jobs […] because instead of 200 people, maybe we will need 100 or 50. How do you feel as a company, participating in the process of, say, job destruction?
Quote #2
We are then selling that we reduce the operational cost that then does not come back to those who have taken this job away from them.

Our study confirms the gap between principles and practices, such as other research. It points out the leading role of regulation over ethics and the general lack of knowledge about the principles within the organisations, which might indicate their lack of effectiveness. As also mentioned by

Table 5 Main findings about privacy Source: developed by the authors

F5. Privacy is the primary concern for companies
F6. Anonymization is, among other tools, the most frequently used

some studies (Mark and Anya 2019), there is a need to develop specific recommendations for specific sectors.

5.2 Privacy

Regarding the second research question, the goal was to understand companies' perception of privacy and the tools they use to cope with it. The main findings are summarised in Table 5 below.

F5. Privacy is the primary concern for companies.

Interestingly, one of the main concerns of organisations is privacy, which is understood as the control and protection of "their own" data and information or as a concept more related to information security.

Quote #19
And one of the first questions is privacy; that is, these algorithms that you are presenting, where are they going to be run? What will their information requirements be?
Quote #5
[…] privacy, is always the "main topic" to address
Quote #13
Privacy is an issue that we do review, always.
Quote #14
My perception is that companies do take great care of their information, to the point that they often prefer not to generate value from information [rather] than to expose their information to a risk of leakage.

On the other hand, the concern for privacy from the users' point of view is mainly motivated by regulation, mainly the GDPR in projects carried out in Europe. Companies are aware of the importance and responsibility of using the personal data of users or customers and about the different criteria and regulations to be considered in different regions.

Quote #5
The first thing is the GDPR.
Quote #11
[…] there you have the data of people, their addresses, you even have precious information, about when they are at home or not, private data, and making proper use of them is essential.
Quote #22
We learned from GDPR to have sensitive data in a separate sensitive database that can only be accessed by several users or with special permissions.
Quote #18
On the other hand, already recently, California has launched its own regulations inspired by GDPR, and other states within the United States are already working along similar lines.

F6. Anonymization is, among other tools, the most frequently used.

Organisations use various tools to manage privacy issues, the foremost being anonymization. However, in many cases, they are aware of the risks of this technique. Other privacy tools, such as synthetic data, are also known and used, and federated learning is being explored, but the concept of differential privacy is not well known.

Quote #10
About anonymised data coming out of a site, that's common practice. Ten years ago, it wasn't, you took a database without a problem, and now they won't come out unless anonymised.
Quote #8
We work with anonymised data because they either take care of anonymising the data previously or help them do that anonymization with software. From there, they keep their keys, and we at no time can draw or triangulate who the individuals are.
Quote #22
We use anonymization in general in data when we use it for training.
Quote #6
There are both sanitized datasets, that is, obfuscation, anonymization, or tokenisation of sensitive fields. If you cannot use real datasets, synthetic ones are created. You're trying to use a series of statistical rules that meet the statistical distributions of the different variables to simulate the original dataset.
Quote #6
[…] because it is known that certain information can reverse the pseudo anonymization that is performed.
Quote #19
Synthetic data generation is one of the issues that help us have enough useful data for customers.
Quote #22
The federated training part we have been taking a look at, more than anything for internal things.
Quote #16
The other option is to have them in a federated data system.
Quote #3

Table 6 Main findings about explainability Source: developed by the authors

F7. Interest in explainability is directly related to the impact on business
F8. Explainability is perceived as a complex feature, with a specific focus on transparency and traceability

Differential privacy? No, I do not [know it].

The previous research has also covered the topic of privacy, linking it to GDPR and highlighting the differences between customers inside and outside Europe (Kevin and Ana 2019; Mark and Anya 2019; Rothenberger et al. 2019; Stahl and Wright 2018). However, our study delves into the various techniques to mitigate privacy issues and discusses privacy from the companies' point of view instead of from the user's. Our study also points out the extended use of anonymization despite some concerns and limitations.

5.3 Explainability

Explainability is one of the topics that has received more attention recently in the research community (Mittelstadt et al. 2019). As per the findings in Table 6 below, the interest from the companies is there, but only when correlated with the potential impact on the business. In addition, explainability is still a complex issue to be addressed.

F7. Interest in explainability is directly related to the impact on business.

Customers are more interested in results than in the how or why. Only in those cases that might significantly impact their business is the customer interested in knowing why.

Quote #8
[…] the client does not want to know why things happen but only wants to know the prediction.
Quote #4
The customer is not asking the machine to explain why they have obtained this data point instead of another one. It is an issue that has not been internalised yet.
Quote #1
Explainability is something they had not even questioned.
Quote #17
In the end, retailers have a lot in the game depending on what we are doing because it is not at the store level. It is that maybe you are moving like a thousand stores, and why is that happening? People sometimes ask.

F8. Explainability is perceived as a complex feature, with a specific focus on transparency and traceability.

Explainability is also perceived as a challenge, full of complexity, and an area of growing interest.

Quote #4
Although we are going to start doing it, now we're not using any explanatory mechanism.
Quote #10
But explainability for us is not possible. Today, we are not able to provide it.
Quote #22
There's generally little transparency everywhere because it is hard to make that transparent to the customer [...] I think it is still challenging to give that security and transparency.
Quote #17
Our algorithms are not black-box precisely to be able to explain. We did not go to straight machine learning and let the algorithm deliver on its own. No, why is this coming out?
Quote #20
If you tell me that an ethical principle is to be able to reproduce the decision that an algorithm made four years ago with a particular training dataset, I need to see how I can recover at that point in time and at that point in responsibility, who were the actors involved.
Quote #16
Two factors that have been emphasised a lot, the ability to know with which elements the model makes the decision and then the provenance systems. I want to know this model backwards to ensure that the data correspond to the same characteristics as the subject being treated.
Quote #7
[…] we give the possibility to automate the whole process, something that not only guarantees repeatability but also totally guarantees traceability, because every time a modification is made to a dataset in a data source, automatically changes the unique and unrepeatable record that identifies each of those elements within the process. So, we are able, whenever certain decision-making or a certain prediction has been reached, to pull back and know exactly what version of the model it is and the specific version of each dataset that led to the generation of that model and, therefore, the prediction that we are putting into production.

Various tools are used, such as LIME, SHAP, local and global explanations or textual explanations. The benefits of symbolic AI are highlighted in addressing the issue of explainability.

Quote #7

Table 7 Main findings about fairness. Source: developed by the authors

F9. Bias is an important issue, but it is application dependent
F10. Bias might be caused by the data or by the programmers

We use a composition of LIME and SHAP.
Quote #22
For example, in the case of recoveries, we have been using LIME.
Quote #7
[…] regardless of the algorithm we use, we will give both global explanations and local explanations of the results.
Quote #19
There are lines of work studying how to explain to people or how to get that layer of interpretability of the models that are being generated
Quote 18
In the symbolic system, any decision you make is then traceable and explainable.

Our conclusions are aligned with the previous research, both in terms of how companies try to accommodate ethics to business objectives (Stahl and Wright 2018) and highlighting transparency as one of the critical aspects (Rothenberger et al. 2019; Vakkuri et al. 2019; Vakkuri 2020). Our study contributes with a specific approach to some of the tools used and highlights the issue's complexity.

5.4 Fairness

The last research question was related to how companies consider potential biases and the concept of fairness and the specific measures that they take with regard to this issue. The findings are shown in Table 7 below.

F9. Bias is an important issue, but it is application dependent.

Bias is a significant and common problem although it depends on the application. Sometimes the problem of bias in data is understood in terms of unbalanced datasets, not as "prejudice". It is also important for companies to differentiate between bias and the objective criteria needed to make a decision.

Quote #11
[…] you have to be very careful, especially when applying bias, especially in artificial intelligence models. I think that is the most important thing.
Quote #19
It is widespread to find situations when we talk about bias, which is one of the most recurring problems. It is perhaps the false perception that if you do not consider certain kinds of information that may already have a bias in terms of income level, gender, etc., then your algorithms and models will be more neutral. Logically that is not the case. There is a lot of impregnation of different variables that collect information or collect previous behaviours or stereotypes that may be conditioned by gender or other conditions such as race or purchasing power. And then? Although that information does not intervene directly, you can have the function of information absorbed by other variables. Even if they are not directly linked to these variables, which would already generate a bias. And in the end, in some way, they're perpetuating that inertia in the future.
Quote #4
At the moment, we are attacking very primary problems where there is no bias.
Quote #14
I have never encountered a problem with biases. In terms of how the data are distributed, yes. And I think there are quite a few tools to help the models work correctly, even in unbalanced datasets, with very limited classes.
Quote #22
Although, indeed, that it's not just about removing variables […]. We try to see which smaller populations come out with biases and then try to retrain or even eliminate that model.
Quote #20
Sometimes there is some confusion in what are objective criteria and what are ethical criteria., companies are there to make money […] What it is about is to embed these models of ethics in such a way that there is no additional bias or exclusion to which reality itself by objective criteria can mark on access to a service.
Quote #10
In certain products like biometrics, biases are absolute because, in fact, I'm looking for biases. I'm looking for a difference.

F10. Bias might be caused by the data or by the programmers.

Most companies consider the bias to be in the data, and in some cases, the source of the data is targeted as a source of possible biases. However, some companies also consider the bias to be in the systems or introduced by the programmers who develop the system. Diversity in the development team is a viable way to reduce these biases. Previous studies have not covered the topic of bias in depth (Rothenberger et al. 2019).

Quote #11
But in the end, biases are often a topic of the data themselves, without the algorithm.

Table 8 Main findings about accountability and governance Source: developed by the authors

F11. Two critical issues for accountability are the quality of data and the monitoring of operations
F12.Regarding decision support systems, it is very relevant to establish a close collaboration with companies or field experts
F13. There is a lack of formalisation of procedures and policies

Quote #7
Those of us working at ML are pretty aware, in general, of the risk that either the model or the data will be biased. In my particular opinion, and I believe that here we generally agree by an overwhelming majority, the bias is not so much in the algorithm or the model; I would almost say that there is no bias there; if anything, there may be bias rather as long as the data scientist prefers using a certain type of algorithm, rather than bias, I think we should talk about "preference". I think that bias is more in the data and how the data are analysed to conclude which family of algorithms or characteristics may be most interesting when analysing our problem when obliged to simplify. Suppose I use any technique of main component analysis, for example. In that case, I am making a simplification that may be conditioned by individual preferences that the analyst may have when deciding what to get rid of.
Quote #21
In the end, everything is biased. That is, having data without bias is challenging.
Quote #2
Unconsciously, the bias is put by the one that develops the algorithm: the person.
Quote #15
With a service we were developing, who had been given a man's name. And then they had made example avatars and were all avatars of a gentleman with a moustache and so on. There were like several, but everyone looked, all coming out of the same pattern. But going a step further, because this is like some kind of virtual employee, I am also analysing who's looking at your doubts and so on, and there was nothing related to maternity leave, for example. Topics that have to do directly with women. […] There, you realise how biased you are trying to assemble a virtual assistant for a company, and you will not only find people of 20–30 years. Some people who are 50 and some people have children, people who do not.

Although it is known that there are tools to correct possible biases, they are not being used frequently.

Quote #6
Fairness tools are there, and we are evaluating them. The state of the art within the group is "under evaluation".

5.5 Accountability and governance

Two topics were raised during the interviews: accountability and governance. Accountability may be understood as the organisation's capability to ensure quality, responsibility, and protection over the use of data, the related algorithms, and the results. Accountability is an integral part of all aspects of data processing, and efforts are being made to reduce the risks for the individual and mitigate social and ethical implications (Data Ethics 2021). The other topic that several interviewees raised was that of governance. Ethical governance can be defined as the set of processes, procedures, cultures, and values designed to ensure the highest standards of ethical behaviour in both individual developers and the organisations for which they work (Winfield and Jirotka 2018). The main finding regarding these topics may be found in Table 8.

F11. Two critical issues for accountability are the quality of data and the monitoring of operations.

The concept of accountability is linked to business performance and is dependent on the type of application. Data quality is mentioned as one of the most critical aspects. Operations monitoring in production needs more attention and development.

Quote #22
If I have to be very "ethical", accuracy will also be affected. Then I think there is a dilemma there, in the end, of how ethical I am and how much business I am losing.
Quote #2
[…] the customer wants the highest degree of success. We are talking about the analysis of legal documentation […]. This information is very relevant and has a significant impact on the business.
Quote #8
The biggest catch we have at the production level is now mainly the quality of the data. Everybody thinks they have some good data. AI systems without good data do not work, and we are at a point where many companies think they have good data, and then when you scratch, they do not have them.
Quote #4
But historical circumstances do not have a sufficient quality of data.
Quote #6

Table 9 Summary of the recommendations Source: developed by the authors

R1. There is a need to promote and develop a culture of "data" in the organisation
R2. The purpose is to build trust and confidence in the systems
R3. Companies should consider the complete life cycle of the algorithm
R4. Organisations should develop, adopt, and use a specific code of ethics for AI systems
R5. Organisations should provide specific training on ethical issues of data and AI systems
R6. There is a clear trend to control data at the source

The whole monitoring part in operations is not mature, to put it elegantly.
Quote #11
It is very focused on fast development and fast deployment.

F12. Regarding decision support systems, it is very relevant to establish close collaboration with company experts.

Quote #3
We are providing a tool so that people who are water experts can be the ones who determine the price.
Quote #16
Then, in the end, it is a system of support to the decision, but that is very filtered, significantly intervened by an expert who also considers other factors that your model does not see.
Quote #20
I think it depends a lot on the type of service and the level of criticality. We use the human in the loop model that different types of human supervision may be based on different criticality levels.
Quote #12
We work with CDOs and data teams […]. It is not like they leave us in a room; we handle the data, and then we give them a report. [...] We usually work with customer teams closely.

F13. There is a lack of formalisation of procedures and policies.

In general, there is an absence of formalisation of procedures and policies. Companies do not have a guide of their own with indications for the ethical design, development, and control of these systems.

Quote #15
Well, it is not formalised at all. It depends a lot on people, not on the process.
Quote #12
There is very little control, very little, and that is like having the wheel continually reinvented. There is a lot of reliance on teams. When the team that's developing leaves you, you often have to start back again.
Quote #10
We do not have a guide of our own; we do not.
Quote 2
Any code or protocol that includes ethical principles? I asked for it, but we don't have it.
Quote #22
We are starting to develop it now. But we just started before the summer. So, we do not have it advanced yet.

Regarding the previous research, accountability is highlighted together with responsibility and transparency (Orr and Davis 2020; Stahl and Wright 2018; Taylor and Dencik 2020; Vakkuri et al. 2019), usually considering accountability tied to liability and responsibility being more vague. Our study delves into some of the factors contributing to accountability and highlights the lack of formalisation, which might be a critical factor on the road to data governance.

5.6 Recommendations

Based on the qualitative analysis and findings related to the research questions, there might be some recommendations worth highlighting (see Table 9):

R1. There is a need to promote and develop a culture of "data" in organisations.

The culture of the organisation is a fundamental element for the adoption and implementation of ethical principles. The organisation's culture can be an obstacle or a catalyst for adopting and implementing the different principles. A clear effort towards developing a sound culture of "data" in organisations should be taken.

Quote #6
[...] another obvious problem is the lack of a culture of data, lack of digital knowledge, lack of technological knowledge. I mean, businesses say "I need this "requirement, and you are the one who sometimes has to go one step back and let them know about data governance, data quality, data privacy. For them, it is automatic. It is a magic box where they want something good, nice and cheap and [where they] do not have to worry.
Quote #20
It has to do with different organisation stakeholders being part of that definition of business objectives that have to contemplate ethical principles, right at the beginning of any artificial intelligence project.
Quote #17

Who defines guiding principles depends on each organisation and dependencies and culture. That has a lot to do with an agile culture […]. There are still many people with more of a waterfall mentality, or more of the old school, whip and budget and deadlines. Then it is tough to consider ethics, privacy, and data governance.

R2. The purpose is to build trust and confidence in the systems.

Explainability and accountability are not to be pursued as an end in themselves, but the goal should be to build trust and confidence in the systems.

Quote #14
[…] this is a personal perception, and I think more than explainability, what I think is that companies should build trust in systems. […] at least in the industrial sector, it may be more relevant to build confidence in the algorithm than to make, in a way, micromanagement of the algorithm.
Quote #20
But what we want to convey, beyond whether sensitive data are processed or not, is trust in systems and trust in systems does create a business impact. Low-trust AI implementations are known not to be as productive as systems where all participants have been taken into account.

R3. Companies should consider the complete life cycle of the algorithm.

Whether the topics are the ethical impact of the product or service, privacy issues, explainability, or accountability, these concepts must cover the algorithm's entire lifecycle. It is necessary to establish indicators in the different phases of design and control during the algorithm's lifetime.

Quote #11
Ethics has to be contained throughout the process; it has to be the whole process.
Quote #6
You have to monitor in operations and make that monitoring simple for alerts and notifications of "your model has become obsolete". […] It is crucial that explainability touches every one of these points in the lifecycle, development, system and makes it easy for the audience to see it.
Quote #12
In a study that we have done, 70% of the algorithm that companies apply, they have no control over their impact. They are not connected to the business. There are no metrics that determine how they impact, such as seeing the effect before, during and after the application. There is also no identification of those responsible for the algorithm creation phases. It is not that it is a monolithic issue. There are several decisions in its life cycle because there is still much to be done at the governance level. […] the monitoring of the algorithm. How do I guarantee that the algorithms are working or not.
Quote #6
In my opinion, at least in Spain, there is one thing that is difficult to understand: that engineering systems, by definition, degrade. There is no system, whether physical or cyber-physical, that does not degrade.

R4. Organisations should develop, adopt, and use a specific code of ethics for AI systems.

Many interviewed companies pointed out the need for an ethical code that collects considerations for the correct design and use of systems.

Quote #2
It is necessary for an ethical code that is accepted and that the company requires compliance for every person working in product development. I do not have any manuals, and I don't have anything to tell me this way or this other way.

R5. Organisations should provide specific training on the ethical issues of data and AI systems.

Specific training is needed in data ethics and artificial intelligence ethics.

Note: The following quotations are in response to the question of whether any specific ethical training is provided.

Quote #19
The training is more oriented right now to the more technical and functional aspects than to ethical issues.
Quote #18
We have the classic training that we have for employees, but it is already more internal to avoid harassment, discrimination, etc.
Quote #11
I'm not doing it with the team, and maybe just like we give technical training, it's important to give this kind of training.

R6. There is a clear trend to control data at the source.

Source data control is recognised as good practice by many of the interviewed companies.

Quote #10
We from the design only use controlled data, and we can explain what sources have been used, in case there were biases, or there were behaviours that I could not avoid or detect […] So from the design, we assure you what the data sources are.
Quote #22
The first thing is what variables will put some bias in my model, all that are sex, race, politics, etc. Eve-

rything, that is; personal variables do not enter the models. For example, they neither enter business intelligence nor the risk part; in general, they are eliminated from the models.

6 Conclusions

This paper aimed to better understand the apparent gap between principles and practices in AI ethics. The qualitative approach allowed us to gather diverse information and offer helpful insight from the perspective of 22 companies. The research questions were focused on how companies try to apply existing guidelines or principles and about the issues of privacy, explainability and fairness.

Most of these organisations agree on the fact that there is indeed a gap between practice and principles. However, we could distinguish two main perspectives: a bottom-up approach, whereby companies first develop AI systems and then try to understand how to generate ethical principles and standardise best practices, and a top–bottom approach, whereby companies receive high-level guidelines and principles and try to find a way to land them into their day-to-day operations. Fact-checking points out that most interviewed companies follow the first approach. Therefore, this needs to be accounted for when addressing the gap between practices and principles since bridging the gap is different if we depart from practices to principles (as most companies seem to be doing) or from principles to practices (as most guidelines seem to suggest).

Regarding the concepts explored: privacy, explainability and fairness, one of the research's main findings has been the relevance of considering the ethical issues related to these concepts throughout the life cycle of the project, system, or product. This has several implications such as considering the ethical dimension throughout the whole process rather than merely compiling a "checklist". Since we have found that most of these concepts are also application dependent, the type of application becomes very relevant because both the process and the issues may be different for different applications, and some ethical aspects might need to be more or less emphasised. This concept of "process" is also considered much more relevant than the specific use of a set of tools. This can also result from a relatively low knowledge and use of available tools by the companies.

Another conclusion is that specific regulations need to be promoted and developed. Although regulations usually tend to be quite generic, it could be complemented by normative sectorial initiatives. Certification by third parties could also contribute to the formalisation and standardisation of policies and procedures. These results clearly point to developing both regulations and tools adapted to specific sectors and applications.

Regarding privacy, there are two main conclusions to be drawn concerning the research goal. There are two different meanings for companies regarding privacy: one is related to data protection relative to the users (as stated in the GDPR and most principles). The other is the protection of the company's data, which appears to be more relevant for companies. This is a straightforward issue that needs to be addressed because principles and practice are clearly divergent in this case. Therefore, since both meanings are important, they should be presented as different terms. As with the other topics, there is a need to promote and develop more tools in the case of privacy beyond anonymization techniques. New approaches, such as federated learning or differential privacy, are still unknown or unused by most companies.

The same conclusion concerning the divergence between principles and practice can be applied to explainability. The desirability shown by the principles and the genuine concerns of companies might not be aligned, and companies are naturally more interested in the results than in the "why". However, there is an agreement pointing towards the development of trust and confidence in systems. Explainability is also considered a complex issue, and, again, the knowledge and use of available tools are limited.

As a set of values, processes, and procedures, organisational culture significantly influences the ethical dimension of AI projects. Therefore, companies must foster an adequate culture despite the higher relevance placed on regulation over ethics. There are specific actions that might contribute to it, such as the following:

- To clearly identify and distinguish the different parties and stakeholders involved in the project and manage their expectations appropriately. For example, it should be taken into consideration that the customer's perspective is different from that of the development team or business team.
- To create multidisciplinary teams (engineers, philosophers, sociologists, data scientists) that contribute different perspectives in the development of AI projects.
- To develop and implement a code of ethics or guidance for the design, development, use and exploitation of AI systems.
- To assess and formalise the processes, procedures and policies related to AI projects, embedding the ethical dimension.
- To provide specific training on AI ethics that must be adapted to the organisation and to the depth of knowledge required in each application and the available procedures and tools.

7 Limitations and further research

Given the nature of qualitative research, some apparent limitations must be considered such as the need to limit the number of companies interviewed. In future research, the study could be extended to other regions and more companies or more deeply into specific sectors or applications. It would also be interesting to build a maturity level survey to assess how companies address the identified issues.

Appendix 1: Interview script

Topic	Question	Time
Opening	Project objectives	2′
	Presentation—role in the organisation—approach to AI	5′
	Describe a project you are working on or have worked on recently (last 3–6 months) Objective Users/segment AI's influence on the project Tools used	5′
Guidelines and principles	Principles/guidelines used for development—in which phases (design, test, control, implementation?) Product design specific code (AI) Does it incorporate ethics criteria into the decision-making model?	10′
Bias and fairness	What do you mean by bias and equity? Incorporate criteria for these concepts Examples? how an AI system can perpetuate existing biases)	10′
Explicability	What do you mean by explainability and interpretability? Are they different? Incorporate criteria for this concepts Examples? Kind of algorithms, kind of tools	10′
Privacy	What do you mean by privacy? Incorporate criteria for this concept Examples? Data source, data type, collection, use, storage	10′
Issues	Problems/dilemmas encountered	5′
Closing	Any final reflections? Thank you and next steps	5′

Source: developed by the authors

Appendix 2: Focus groups script discussion group

Welcome		3′
A brief description of the objective of the investigation		2′
Introductory question	General impression on the draft report	10′
Key issues	Main question 1: (show the findings table). From this list of findings, which ones do you consider most relevant and why? Main question 2: What has caught your eye?	15′ 15′
Closing	Closing question 1: What next step would you consider attractive to deepen the investigation?	15′
Thank you and next steps		5′

Source: developed by the authors

Appendix 3: Codes and groups of codes

Codes		
Accuracy	Against profit margin	Algorithm supervision
Algorithm training	Algorithms carry bias	Anonymization not enough
Anonymization solutions	Anonymization Used	Applications
Asilomar principles	Autonomy	Balanced data
Beneficence	Benefit vs. the risk of being wrong	BIAS
Bias by correlation	Bias due to the analyst work	Bias due to the origin of data
Bias in the data	Bias is a key topic	Bias is no concern
Bias is not critical	Bias not found	Bias people vs. machines
Bias vs. diversity	Bias vs. explainability	Bias vs. objective criteria
Bias vs. unbalanced data	Broad approach to explainability	Business KPIs and Model KPIs
Business team's expectations	Business teams vs. technical teams	Certification
C-level's interest in reputation	Client does not ask ethical questions	Client more interested in the result than in the why
Client wants to know why	Clients	Client's criteria on privacy
Client's ethical criteria	Client's specific training	Colinearity
Company data and external data	Company size	Complexity of explainability

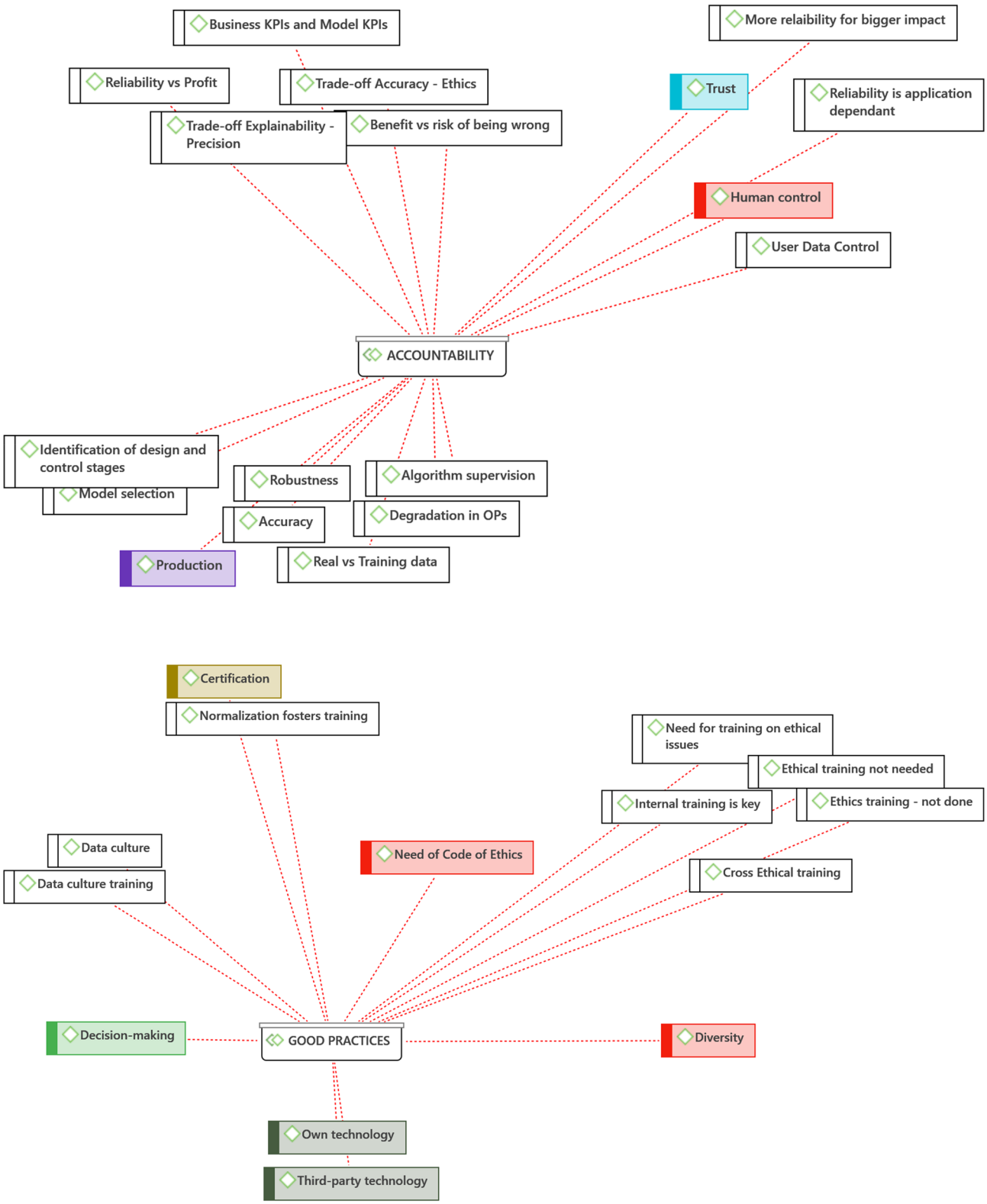
Codes		
Compliance mechanisms	Control of data in origin	Controlled access
Cross data analysis	Cross ethical training	Cross-integration
Data availability is key	Data culture	Data culture training
Data gathering	Data governance	Data misuse risk
Data ownership	Data quality	Dataset selection
Data source problem due to decalibration	Decision-making	Degradation in OPs
DevOps principles	Differential privacy is not well known	Diversity
Effect on employment	Ethical checkpoints	Ethical concern is not that important
Ethical dimension	Ethical training not needed	Ethics
Ethics as a process	Ethics by design not used	Ethics by design used
Ethics of data vs. system	Ethics of system vs. data	Ethics training—not done
EU vs. USA differences	Explainability	Explainability—life cycle
Explainability—not feasible	Explainability fostered by regulation	Fast deployment
Federated learning	Federated learning and bias	Future
Future of regulation	Gap between principles and practice	GDPR
Gender bias	General knowledge about principles	Global explanations
Google principles	Hierarchical analysis to control for bias	Human being
Human control	Hype	Identification of design and control stages
Importance of regulation vs. ethics	Individual knowledge vs. procedures	Inequality
Integrative approach	Internal client vs. external client	Internal training is key
Interpretability	ISO 27001	Knowledge about LIME
Knowledge about SHAP	Lack of auditing	Lack of formal checkpoints
Lack of metrics and indicators	Local explanations	Low interest on the why but high interest in results
Magic expectations	Misuse	MLOps principles
Model selection	Montreal principles	More reliability for bigger impact
Necessity vs. correlation	Need for training on ethical issues	Need for code of ethics
Nilsson principles	No personal data from users	Nondisclosure agreement
Normalisation fosters training	Open data	Opportunities
Organisational culture	Organisation's role is key	Origin of data
Own guidelines	Own technology	Penalties and fines
Personal data usage responsibility	Personal interest in principles	Principles
Principles are too general	Principles vs. ethical culture	Privacy
Privacy check	Privacy during model training	Privacy is the top concern
Privacy vs. profit	Procedures	Production
Reactive behaviour	Real vs. training data	Recommendation to customers
Regressions are explainable	Regulation fosters privacy	Regulation is not enough
Reliability is application dependent	Reliability vs. profit	Reputation
Respect to people	Results-oriented	Risk of anonymization
Robustness	Safety vs. freedom	Sectors
Sensitive variables	Shared responsibility	SLA—service level agreement
Social responsibility	Standardisation	Stored data vs. manual scanning
Sustainability	Symbolic IA—explainable	Synthetic data used
Terms of use	Textual explanations	Third-party technology
Third-party tools	To prevent bias during model training	Too many principles
Tools	Tools are not the main issue	Towards explainability
Traceability	Trade-off accuracy—ethics	Trade-off explainability—precision
Transparency	Transparent algorithms	Trust
Trust over explainability	UE principles	Use of anonymised data
User data control	User does not know or does not care	Utility of bias
Visualisation	Work together with client's team	

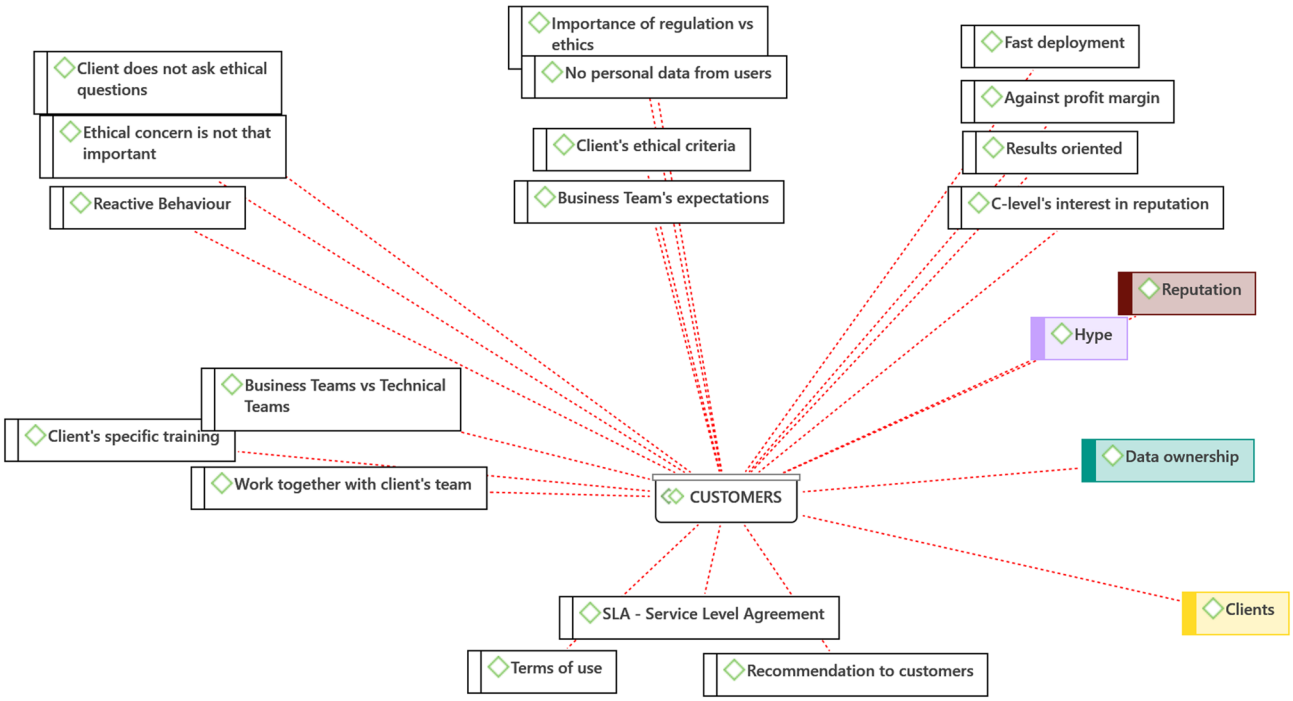
Source: developed by the authors

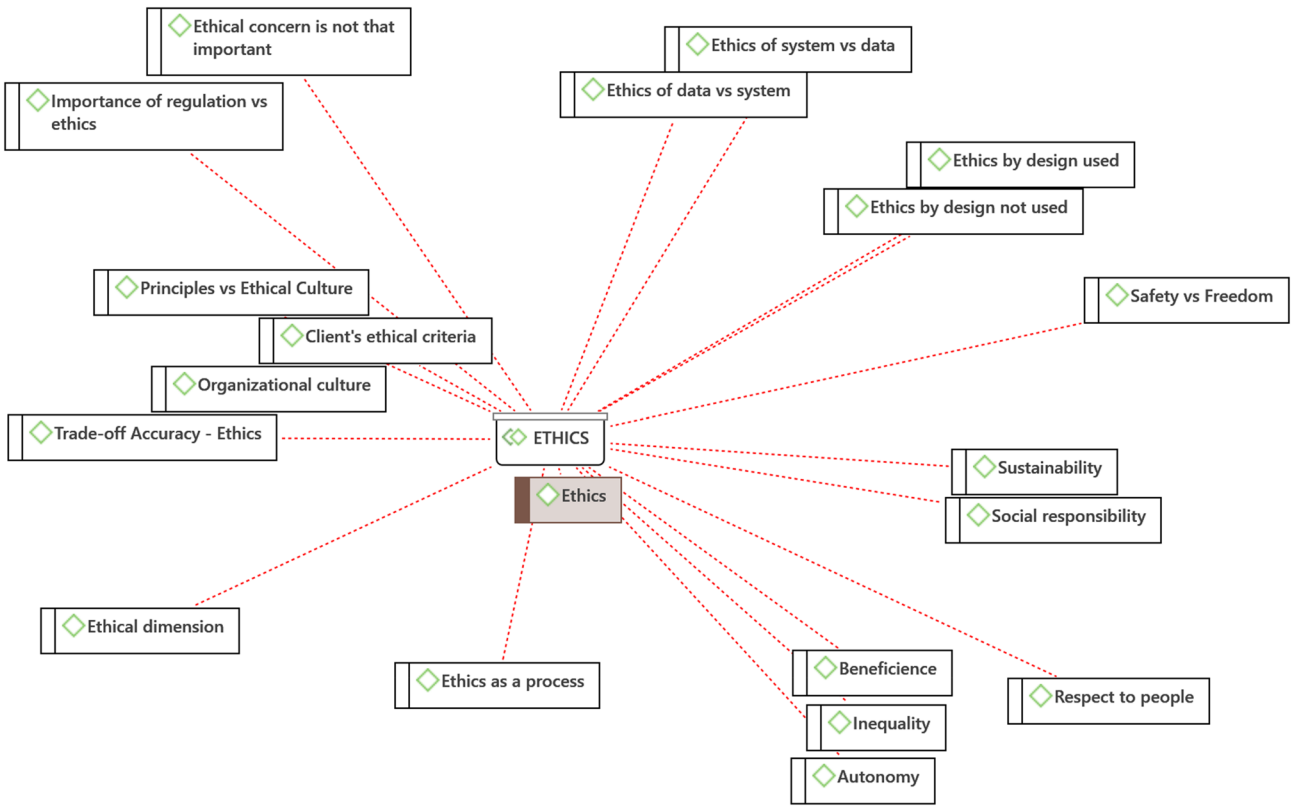
Groups of codes
Accountability
Customers
Ethics
Explainability
Fairness
Good practices
Origin of data
Other issues
Principles
Privacy
Standardization

Source: developed by the authors

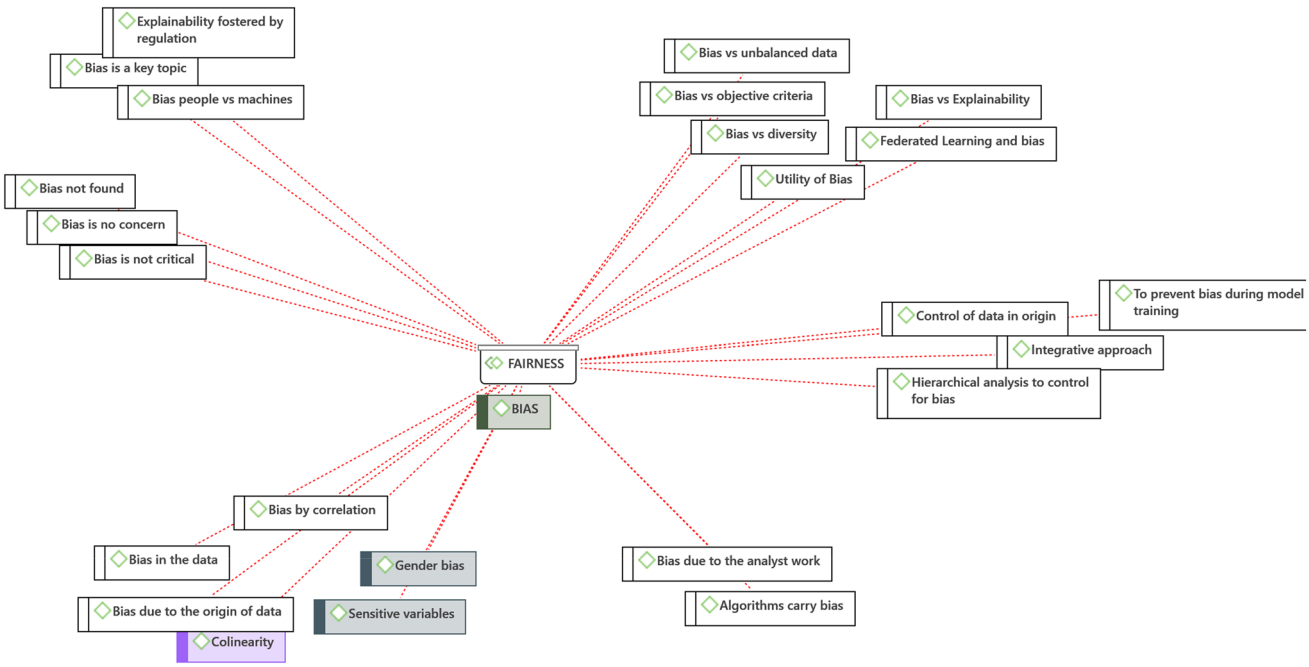
Appendix 4: Network maps (groups of codes)

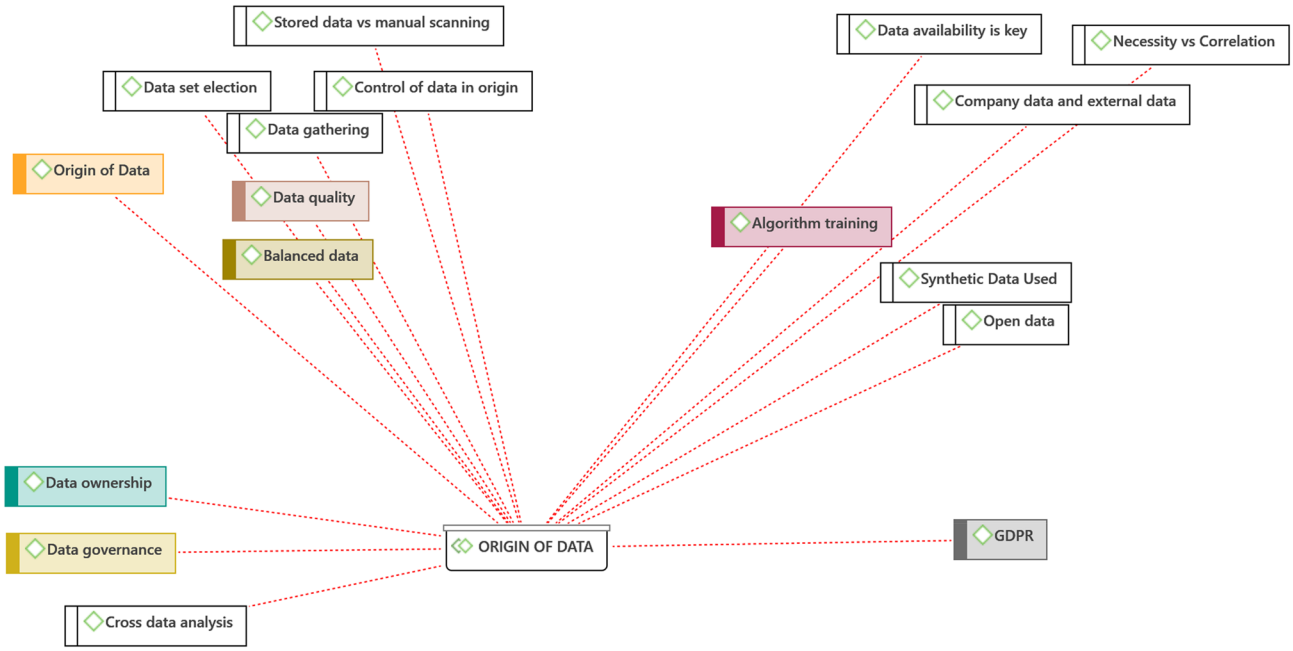


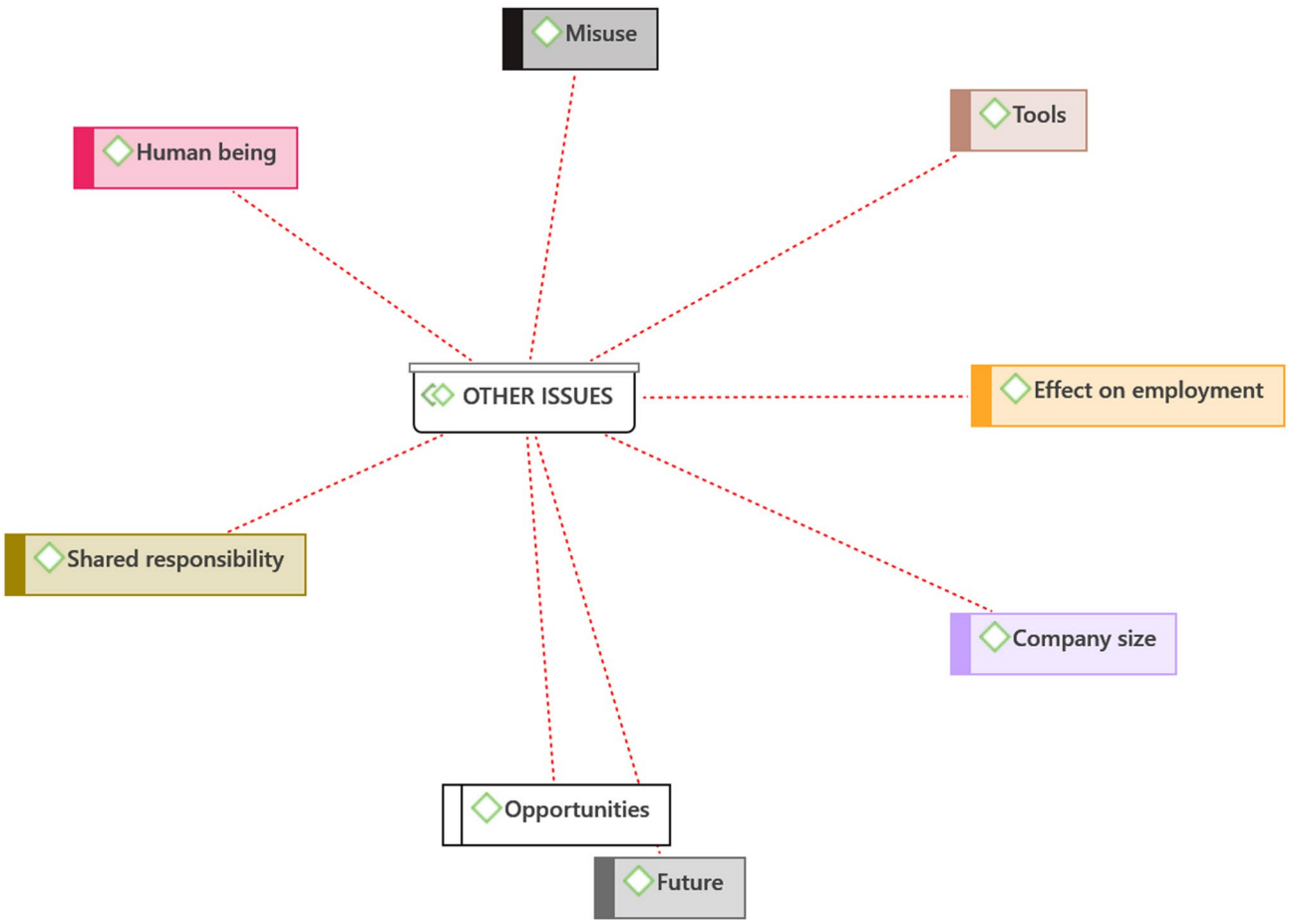

Importance of regulation vs ethics
No personal data from users
Client's ethical criteria
Business Team's expectations
Client does not ask ethical questions
Ethical concern is not that important
Reactive Behaviour
Fast deployment
Against profit margin
Results oriented
C-level's interest in reputation
Reputation
Hype
Business Teams vs Technical Teams
Client's specific training
Work together with client's team
CUSTOMERS
Data ownership
Clients
SLA - Service Level Agreement
Terms of use
Recommendation to customers

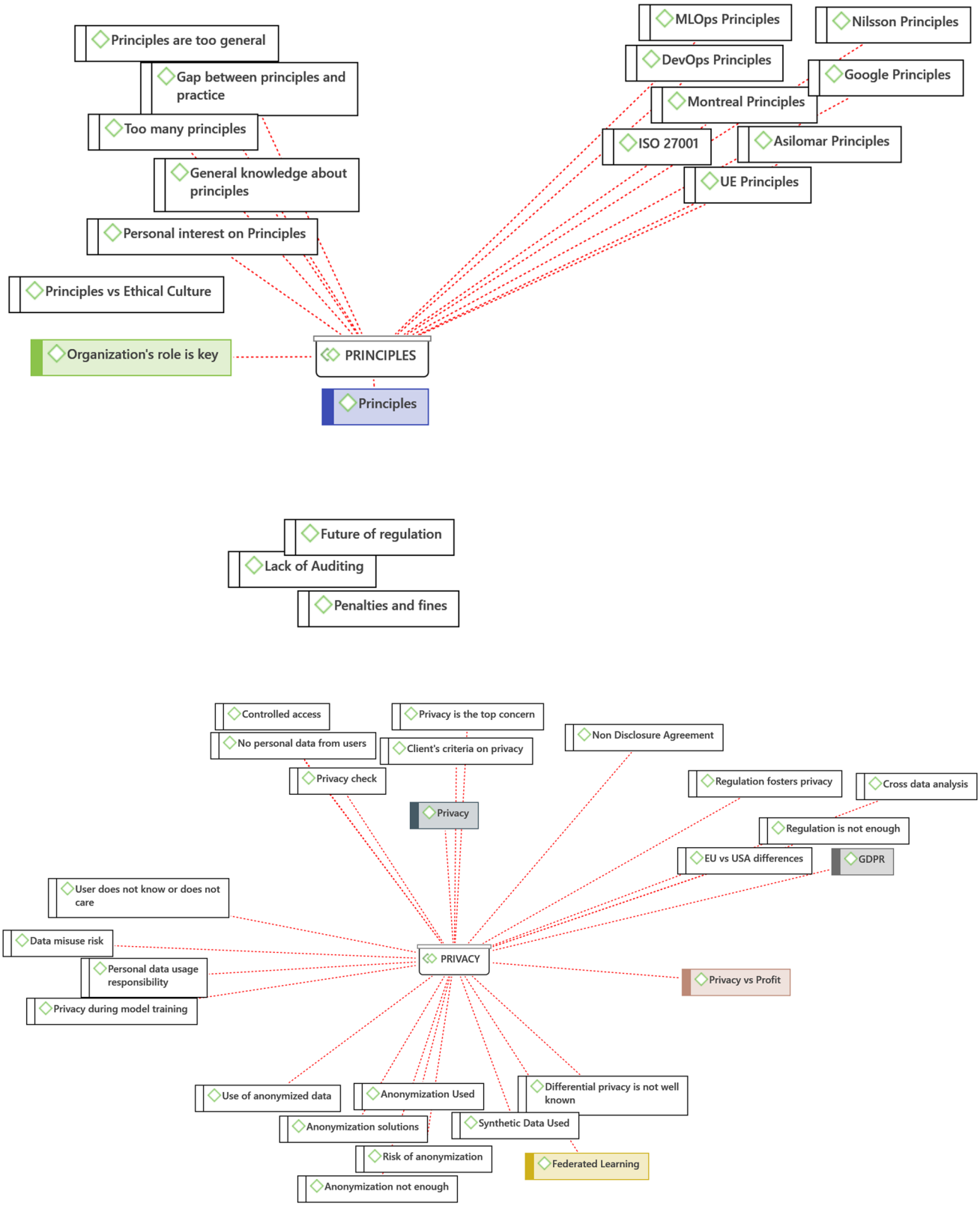

Ethical concern is not that important
Importance of regulation vs ethics
Ethics of system vs data
Ethics of data vs system
Ethics by design used
Ethics by design not used
Principles vs Ethical Culture
Client's ethical criteria
Organizational culture
Trade-off Accuracy - Ethics
Safety vs Freedom
ETHICS
Ethics
Sustainability
Social responsibility
Ethical dimension
Ethics as a process
Beneficience
Inequality
Autonomy
Respect to people

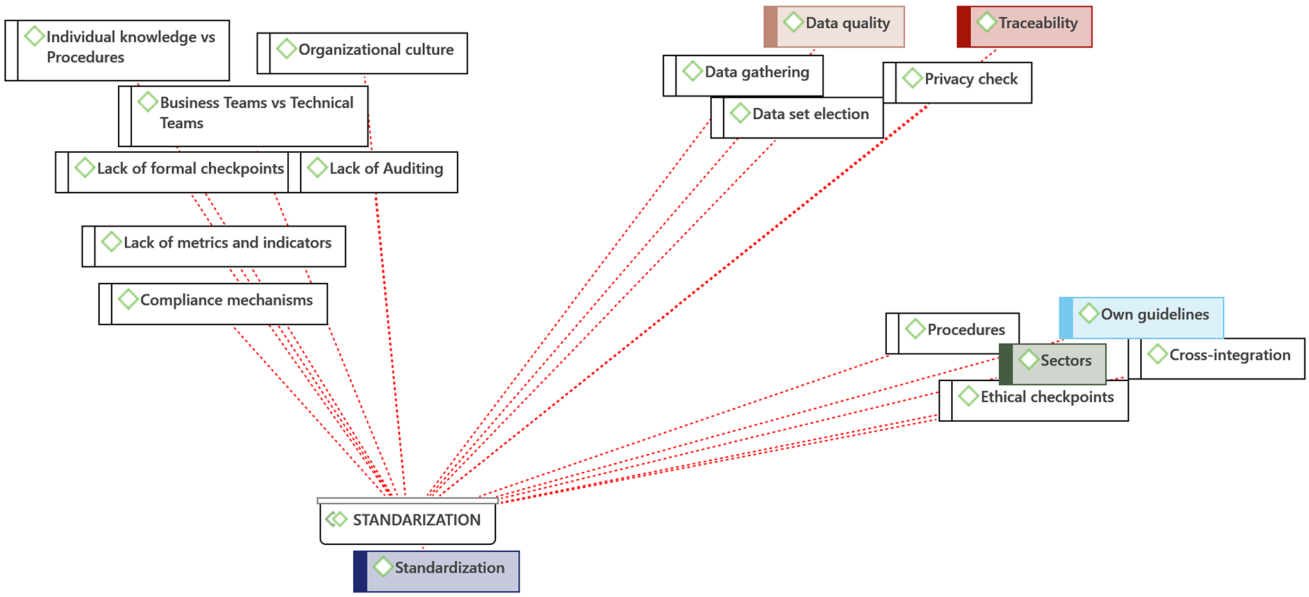

Explainability - not feasible
Complexity of explainability
Trade-off Explainability - Precision
Tools are not the main issue
Knowledge about SHAP
Client more interested in the result than in the why
Local explanations
Symbolic IA - explainable
Low interest on the why but high interes on results
Third-party tools
Textual explanations
Client wants to know why
Global explanations
Knowledge about LIM
Transparent algorithms
Internal client vs External client
Regressions are explainable
Visualization
Lack of metrics and indicators
EXPLAINABILITY
Transparency
Explainability
Traceability
Degradation in OPs
Applications
Explainability - Life cycle
Interpretability
Trust over explainability
Broad apporach to explainability
Towards explainability


Explainability fostered by regulation
Bias vs unbalanced data
Bias is a key topic
Bias vs objective criteria
Bias vs Explainability
Bias people vs machines
Bias vs diversity
Federated Learning and bias
Utility of Bias
Bias not found
Bias is no concern
Bias is not critical
To prevent bias during model training
Control of data in origin
Integrative approach
FAIRNESS
Hierarchical analysis to control for bias
BIAS
Bias by correlation
Bias in the data
Gender bias
Bias due to the analyst work
Bias due to the origin of data
Sensitive variables
Algorithms carry bias
Colinearity


Stored data vs manual scanning
Data availability is key
Necessity vs Correlation
Data set election
Control of data in origin
Company data and external data
Data gathering
Origin of Data
Data quality
Algorithm training
Balanced data
Synthetic Data Used
Open data
Data ownership
Data governance
ORIGIN OF DATA
GDPR
Cross data analysis


Misuse
Tools
Human being
OTHER ISSUES
Effect on employment
Shared responsibility
Company size
Opportunities
Future


Principles are too general
Gap between principles and practice
Too many principles
General knowledge about principles
Personal interest on Principles
Principles vs Ethical Culture
Organization's role is key
PRINCIPLES
Principles
MLOps Principles
Nilsson Principles
DevOps Principles
Google Principles
Montreal Principles
ISO 27001
Asilomar Principles
UE Principles
Future of regulation
Lack of Auditing
Penalties and fines
Controlled access
Privacy is the top concern
No personal data from users
Client's criteria on privacy
Non Disclosure Agreement
Privacy check
Privacy
Regulation fosters privacy
Cross data analysis
Regulation is not enough
EU vs USA differences
GDPR
User does not know or does not care
Data misuse risk
Personal data usage responsibility
Privacy during model training
PRIVACY
Privacy vs Profit
Use of anonymized data
Anonymization Used
Differential privacy is not well known
Anonymization solutions
Synthetic Data Used
Risk of anonymization
Federated Learning
Anonymization not enough



References

Abrams M, Abrams J, Cullen P, Goldstein L (2019) Artificial intelligence, ethics, and enhanced data stewardship. IEEE Secur Priv 17(2):17–30. https://doi.org/10.1109/MSEC.2018.2888778

Bietti E (2020) From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In: FAT* 2020—proceedings of the 2020 conference on fairness, accountability, and transparency, pp 210–219. https://doi.org/10.1145/3351095.3372860

Braun V, Clarke V (2006) Using thematic analysis in psychology. Qual Res Psychol 3(2):77–101. https://doi.org/10.1191/1478088706QP063OA

Brundage M, Avin S, Wang J, Belfield H, Krueger G, Hadfield G, Khlaaf H, Yang J, Toner H, Fong R, Maharaj T, Koh PW, Hooker S, Leung J, Trask A, Bluemke E, Lebensold J, O'Keefe C, Koren M, et al. (2020) Toward trustworthy AI development: mechanisms for supporting verifiable claims. http://arxiv.org/abs/2004.07213

Bryant A, Charmaz K (eds) (2007) The Sage handbook of grounded theory. Sage

Canca C (2020) Operationalising AI ethics principles. In: Communications of the ACM, vol. 63, issue 12. https://doi.org/10.1145/3430368

Charisi V, Dennis L, Fisher M, Lieck R, Matthias A, Slavkovik M, Sombetzki J, Winfield AFT, Yampolskiy R (2017) Towards moral autonomous systems. http://arxiv.org/abs/1703.04741

Confessore N (2018) Cambridge analytica and facebook: the scandal and the fallout thus far. NY Times.https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html. 4 Apr 2018

Creswell JW (2012) Qualitative inquiry and research design: choosing among five approaches. SAGE, Los Angeles

Daly A, Hagendorff T, Li H, Mann M, Marda V, Wagner B, Wang WW (2020) AI, Governance and ethics: global perspectives. SSRN Electron J. https://doi.org/10.2139/ssrn.3684406

Dastin J (2018) Amazon scraps secret AI recruiting tool that showed bias against women. Reuters, London, UK. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G. Accessed 1 Mar 2021

Data Ethics (2021) Data ethics readiness test. https://dataethics.eu/home/dataethics-readiness-test-2021/. Accessed 1 Mar 2021

de Bruin B, Floridi L (2017) The ethics of cloud computing. Sci Eng Ethics 23:21–39. https://doi.org/10.1007/s11948-016-9759-0

Edwards L, Veale M, Welbl J, van KM, Binns R, Lane G, Henderson T (2018) Slave to the algorithm? Why a "right to an explanation" is probably not the remedy you are looking for (Vol. 16). https://perma.cc/PJX2-XT7X

Eitel-Porter R (2020) Beyond the promise: implementing ethical AI. AI Ethics 1(1):73–80. https://doi.org/10.1007/S43681-020-00011-6

European Commission (2019) Ethics guidelines for trustworthy AI. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai. Accessed 1 Mar 2021

Fernow J, de Miguel Beriain I, Brey P, Stahl B (2019) Setting future ethical standards for ICT, Big Data, AI and robotics. ORBIT J 2019(1):1–8. https://doi.org/10.29297/ORBIT.V2019I1.115

Fjeld J, Achten N, Hilligoss H, Nagy A, Srikumar M (2020) Principled artificial intelligence: mapping consensus in ethical and rights-based approaches to principles for AI. SSRN Electron J. https://doi.org/10.2139/SSRN.3518482

Floridi L (2019) Translating principles into practices of digital ethics: five risks of being unethical. Philosophy and technology, vol 32. Springer Netherlands, pp 185–193. https://doi.org/10.1007/s13347-019-00354-x

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F, Schafer B, Valcke P, Vayena E (2018) AI4People—an ethical framework for a good AI Society: opportunities, risks, principles, and recommendations. Mind Mach 28(4):689–707. https://doi.org/10.1007/s11023-018-9482-5

Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé III H, Crawford K (2018) Datasheets for datasets. https://arxiv.org/abs/1803.09010v7

Gonzalez Fabre R, Camacho Ibáñez J, Tejedor Escobar P (2021) Moral control and ownership in AI systems. AI & Soc 36:289–303 https://doi.org/10.1007/s00146-020-01020-z

Govia L (2020) Coproduction, ethics and artificial intelligence. J Dig Soc Res. https://doi.org/10.33621/jdsr.v2i3.53

Greene D, Hoffmann AL, Stark L (2019) Better, nicer, clearer, fairer: a critical assessment of the movement for ethical artificial intelligence and machine learning. In: Proceedings of the 52nd Hawaii international conference on system sciences

Hagendorff T (2020) The ethics of AI ethics: an evaluation of guidelines. Mind Mach 30(1):99–120. https://doi.org/10.1007/s11023-020-09517-8

Hagerty A, Rubinov I (2019) Global AI ethics: a review of the social impacts and ethical implications of artificial intelligence. https://arxiv.org/abs/1907.07892v1

Hennink MM, Kaiser BN, Marconi VC (2016) Code saturation versus meaning saturation: how many interviews are enough? Qualitative Health Research 27(4):591–608. https://doi.org/10.1177/1049732316665344

Hickok M (2021) Lessons learned from AI ethics principles for future actions. AI Ethics 1(1):41–47. https://doi.org/10.1007/S43681-020-00008-1

Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1(9):389–399

Kevin M, Ana FI (2019) Case study—customer relation management, smart information systems and ethics. ORBIT J 2(2):1–24. https://doi.org/10.29297/ORBIT.V2I2.114

Kitchin R (2016) Thinking critically about and researching algorithms. Inf Commun Soc 20(1):14–29. https://doi.org/10.1080/1369118X.2016.1154087

Krefting L (1991) Rigor in qualitative research: the assessment of trustworthiness. Am J Occup Ther 45(3):214–222. https://doi.org/10.5014/AJOT.45.3.214

Krueger RA, Casey MA (2000) Focus groups—a practical guide for applied research, 3rd edn. SAGE Publications Inc., USA

Mark R, Anya G (2019) Ethics of using smart city AI and Big Data: the case of four large European Cities. ORBIT J 2(2):1–36. https://doi.org/10.29297/ORBIT.V2I2.110

Marshall C, Rossman GB (1995) Designing qualitative research, 2nd edn. Sage Publications, Thousand Oaks

Mason M (2010) Sample size and saturation in PhD studies using qualitative interviews. In: Forum qualitative Sozialforschung/Forum: qualitative social research, vol 11, no 3

McDuie-Ra D, Gulson K (2020) The backroads of AI: the uneven geographies of artificial intelligence and development. Area 52(3):626–633. https://doi.org/10.1111/AREA.12602

McNamara A, Smith J, Murphy-Hill E (2018) Does ACM's code of ethics change ethical decision making in software development? In: ESEC/FSE 2018: proceedings of the 2018 26th ACM joint meeting on European software engineering conference and symposium on the foundations of software engineering. https://doi.org/10.1145/3236024.3264833

Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. Nat Mach Intell 1(11):501–507. https://doi.org/10.1038/s42256-019-0114-4

Mittelstadt B, Russell C, Wachter S (2019) Explaining explanations in AI. In: FAT* 2019—proceedings of the 2019 conference on fairness, accountability, and transparency, pp 279–288. https://doi.org/10.1145/3287560.3287574

Morgan DL (1997) Focus groups as qualitative research. SAGE, New York

Morley J, Floridi L, Kinsey L, Elhalal A (2020) From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Sci Eng Ethics 26(4):2141–2168. https://doi.org/10.1007/s11948-019-00165-5

Morley J, Elhalal A, Garcia F, Kinsey L, Mökander J, Floridi L (2021) Ethics as a service: a pragmatic operationalisation of AI ethics. Minds Mach 31(2):239–256. https://doi.org/10.1007/S11023-021-09563-W

Orr W, Davis JL (2020) Attributions of ethical responsibility by artificial intelligence practitioners. Inf Commun Soc 23(5):719–735. https://doi.org/10.1080/1369118X.2020.1713842

Rothenberger L, Fabian B, Arunov E (2019) Relevance of ethical guidelines for artificial intelligence—a survey and evaluation. In: Progress papers. https://aisel.aisnet.org/ecis2019_rip/26

Stahl BC, Wright D (2018) Ethics and privacy in AI and Big Data: implementing responsible research and innovation. IEEE Secur Priv 16(3):26–33. https://doi.org/10.1109/MSP.2018.2701164

Stahl BC, Antoniou J, Ryan M, Macnish K, Jiya T (2021) Organisational responses to the ethical issues of artificial intelligence. AI Soc. https://doi.org/10.1007/s00146-021-01148-6

Taylor L, Dencik L (2020) Constructing commercial data ethics. Technol Regul 2020:1–10

Vakkuri K-KKV (2020) AI ethics in industry: a research framework. https://arxiv.org/ftp/arxiv/papers/1910/1910.12695.pdf

Vakkuri V, Kemell KK (2019) Implementing AI ethics in practice: an empirical evaluation of the RESOLVEDD strategy. In: Hyrynsalmi S, Suoranta M, Nguyen-Duc A, Tyrväinen P, Abrahamsson P (eds) Software business. ICSOB 2019. Lecture Notes in Business Information Processing, vol 370. Springer, Cham. https://doi.org/10.1007/978-3-030-33742-1_21

Vakkuri V, Kemell K-K, Kultanen J, Siponen M, Abrahamsson P (2019) Preprint notes. In: Ethically aligned design of autonomous systems: industry viewpoint and an empirical study. arXiv preprint arXiv:1906.07946

Vallés M (2009) Entrevistas cualitativas (cuadernos metodológicos). Segunda edición. Centro de Investigaciones Sociológicas, Madrid

Watson GJ, Desouza KC, Ribiere VM, Lindič J (2021) Will AI ever sit at the C-suite table? The future of senior leadership. Bus Horiz. https://doi.org/10.1016/j.bushor.2021.02.011

Whittlestone J, Nyrup R, Alexandrova A, Cave S (2019) The role and limits of principles in AI ethics: towards a focus on tensions. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pp 195-200

Winfield AF, Jirotka M (2018) Ethical governance is essential to building trust in robotics and artificial intelligence systems. Philosoph Trans Royal Soc A: Math Phys Eng Sci 376(2133):20180085

Winfield AF, Michael K, Pitt J, Evers V (2019) Machine ethics: the design and governance of ethical AI and autonomous systems. Proc IEEE 107(3):509–517. https://doi.org/10.1109/JPROC.2019.2900622

Yeung K, Howes A, Pogrebna G (2019) AI governance by human rights-centred design, deliberation and oversight: an end to ethics washing. The Oxford handbook of AI ethics. Oxford University Press