Artificial Intelligence Review (2024) 57:110
https://doi.org/10.1007/s10462-024-10740-3

Applying the ethics of AI: a systematic review of tools for developing and assessing AI-based systems

Ricardo Ortega-Bolaños[1] · Joshua Bernal-Salcedo[1] · Mariana Germán Ortiz[3,7] · Julian Galeano Sarmiento[2] · Gonzalo A. Ruz[3,5,6] · Reinel Tabares-Soto[1,3,4,7]

Accepted: 24 February 2024 / Published online: 5 April 2024

Abstract
Artificial Intelligence (AI)-based systems and their increasingly common use have made it a ubiquitous technology; Machine Learning algorithms are present in streaming services, social networks, and in the health sector. However, implementing this emerging technology carries significant social and ethical risks and implications. Without ethical development of such systems, there is the potential for this technology to undermine people's autonomy, privacy, and equity, even affecting human rights. Considering the approaches necessary for ethical development and effective governance of AI, such as ethical principles, guidelines, and technical tools, the question arises regarding the limitations of implementing these measures by the highly technical personnel involved in the process. In this context, we propose the creation of a typology that distinguishes the different stages of the AI life-cycle, the high-level ethical principles that should govern their implementation, and the tools with the potential to foster compliance with these principles, encompassing both technical and conceptual resources. In addition, this typology will include relevant information such as developmental level, related tasks, sectors, and language. Our research is based on a systematic review in which we identified 352 resources and tools. We expect this contribution to be valuable in promoting ethical AI development for developers and leaders who manage these initiatives. The complete typology and the comprehensive list of resources are available for consultation at https://ricardo-ob.github.io/tools4responsibleai.

Keywords Artificial Intelligence · Responsible development · Governance · Ethics of AI · Machine learning · AI life cycle

1 Introduction

Artificial Intelligence (AI) can strengthen different economic and social sectors, thus improving our quality of life (Cath 2018). The health sector (Morley et al. 2020; Yu et al. 2018), financial or business sector (Buchanan 2019; Loureiro et al. 2021), educational sector (Chen et al. 2020), security and justice (Hoadley and Lucas 2018; Rigano 2018), among others, are already incorporating AI-based systems, specifically employing Machine Learning (ML) or Deep Learning (DL) algorithms. Government agencies and private companies

Extended author information available on the last page of the article

see significant advantages and benefits in these systems for decision-making or support due to their great precision, automation capacity, and data analytic (Wirtz et al. 2018). Furthermore, these systems can improve compliance with human rights and social welfare and contribute to policy formulation, public service provision, and internal management within the public sector (Cath 2018; Henman 2020; van Noordt and Misuraca 2022).

In the context of our research, AI is defined as computational systems capable of executing tasks that would usually require human intervention (Wang 2019). These systems have become integral to our society due to their remarkable ability to identify patterns in large datasets, appearing in streaming services, machine translation, voice assistants, and chatbots. However, implementing this emerging technology in this fields brings social, economic, and primarily ethical implications (Benefo et al. 2022; Floridi 2019; Jia and Zhang 2021; Martin 2019). Some implications of AI include possible discrimination against vulnerable populations, privacy violations, loss of autonomy or sense of agency, and even impacts on sustainability (Devillers et al. 2021; Galaz et al. 2021). The misuse and incorrect development of AI have been the focus of debate among academics, policymakers, private companies, and civil society (Cath et al. 2018; Jia and Zhang 2021; Mittelstadt et al. 2016). Consequently, numerous documents have been produced suggesting specific principles or guidelines for the ethical advancement of AI.[1] Notably, Jobin et al. (2019), Fjeld et al. (2020), Floridi and Cowls (2019), and Hagendorff (2020) scrutinize the most relevant documents proposed by organizations and governments. Table 1 shows some examples of the main ethical principles and other low-level principles identified and organized by the community. Their respective associated principles and descriptions are included.

The concern among the mentioned actors is well-founded.[2]Raso et al. (2018) noted that a single AI application -among the thousands already in use- can affect numerous civil, political, economic, social, and cultural rights. A poor implementation or development of AI can undermine citizens' trust, generate damage (tangible and intangible), and, ultimately, endanger human rights (Latonero 2018). Consequently, in recent years, 'AI ethics' has been taking shape to answer the question: How is an AI developed that benefits society? With clear guidelines and actions. The vast majority of methods to achieve ethical development of AI (Floridi 2015; Morley et al. 2020) are based on following ethical frameworks with widely accepted ethical principles (see Table 1), which provide minimum requirements to build and use an ethically sound AI system (Ashok et al. 2022). Other approaches focus on government laws and technical or practical tools (Cath 2018; Stahl 2021a).

In order to clarify the concept of a tool in the context of ethical AI development, from now on, it will be understood as a versatile resource capable of supporting this process. This term encompasses many resources, such as open-source libraries, procedures, good practices, and government frameworks. These tools have different ways of contributing and play specific roles in developing responsible or ethical AI. For example, technical tools like libraries and web applications generally focus on specific aspects, such as bias assessment in AI systems. On the other hand, non-technical tools, such as guidelines and abstract definitions, provide ethical guidance more conceptually. Although these tools may differ in nature, they share the fundamental purpose of guiding, informing, teaching, or facilitating the development and evaluation of ethical AI. In this sense, the term 'tool' is used

[1] See https://inventory.algorithmwatch.org and https://www.aiethicist.org/ai-principles.

[2] See https://github.com/daviddao/awful-ai, https://incidentdatabase.ai and https://www.aiaaic.org/aiaaic-repository.

Table 1 Ethical principles and their interpretation

Principles	Associated principles	Description
Transparency	Explicability, transparency, open access, understandability, right to information, interpretability	Transparency is described as the characteristic of a project to be explained, understood, replicated, and justified. It allows understanding the process to obtain specific results, being a bridge of connection with the other principles (Morley et al. 2020)
Justice and Fairness	Justice, fairness, inclusion, equality, redress, non-discrimination, prevention of bias, impartiality	It is an ethical perspective whereby ethical decisions are made based on universal principles and standards and in an impartial and verifiable manner to ensure all persons' fair and equitable treatment (Botes 2000)
Beneficence	Benefit of society, human values, wellbeing, peace, social good	"The principle of beneficence underscores the moral obligation to act for the benefit of others, including protecting the rights of others, preventing harm to others, and helping those in danger" (Cummings and Mercurio 2010, p. 2)
Non-maleficence	Safety, non-maleficence, security, precaution, prevention, integrity, non-subversion	Refers to avoiding excessive or improper use of AI technologies that may lead to various negative consequences, in other words, avoiding harm to people or minimizing risks so that new digital solutions do not harm users (Becker et al. 2022; Floridi and Cowls 2019)
Accountability	Responsibility, accountability, impact assessment, auditing, verifiability and replicability, legal obligation and liability, environmental responsibility, acting with integrity	It can be understood as an integration of the other principles and a way to answer the question of "who is responsible for the way it works?" (Floridi et al. 2018, p. 12) and the effects it may cause. It guides the actions or conduct and explains why the decisions and measures were taken, considering that the correct functioning must always be guaranteed (OECD 2019a)
Privacy	Privacy, consent, control over use of data, privacy by design, data protection laws, restrict processing, right of rectification	It is characterized by being closely related to the access of individuals and control over how personal data is used. Generally, this principle is the individual's right to control their information (storage and use) (Floridi et al. 2018; Khan et al. 2021)

Prepared by the authors based on Fjeld et al. (2020) and Jobin et al. (2019)

inclusively, reflecting the diversity of resources contributing to the process, just as a carpenter recognizes a hammer, chisel, and saw as essential instruments for creating a product. This holistic approach seeks to underline the importance of any resource capable of promoting ethics in AI development.

Existing technical tools are linked to certain ethical principles, e.g., technical solutions aimed at creating explainable systems that protect the privacy of sensitive or personal data and involve data collection aware of biases and discrimination (Lepri et al. 2017). However, regarding transparency, non-maleficence, and beneficence, the ethical criteria and technical mechanisms to safeguard them are left to the judgment of the person or organization that implements the AI. As a result, an ethical standard is often not met (Floridi 2019; Floridi et al. 2020). The debate on whether ethical principles are sufficient to develop ethical AI leans towards the fact that *they are not*. The documents' definitions remain abstract and unhelpful for developers (Mittelstadt 2019; Morley et al. 2021). In addition, this abstraction "tends to encourage a public opinion that there are good and bad algorithms, rather than well designed or poorly designed algorithms" (Morley et al. 2021, p. 2). As high-level principles continue to emerge, there is also a rise in resources addressing AI ethics. However, it remains problematic to determine which tool could aid adherence to these principles (Corrêa 2021; Morley et al. 2020).

Understanding the Collingridge dilemma is crucial in AI ethics, especially when exploring how technology ethics and values interact. This concept highlights the intricate link between technologies and the frameworks we use to judge them. The ethical side of the Collingridge dilemma poses an ongoing challenge: navigating a constantly changing landscape where technologies reshape societal values and ethical standards (Kudina and Verbeek 2019). This dilemma creates a paradoxical situation where ethical considerations swing between being "too early" -assessing technologies without knowing how they might alter ethical frameworks- and "too late" -understanding ethical implications only after the technology is firmly established, making change difficult. Fundamentally, evaluating technology from an ethical standpoint requires a nuanced equilibrium between foreseeing societal impacts with limited knowledge and comprehending those effects only after the technology has already influenced the frameworks for assessment (Kudina and Verbeek 2019; Strümke et al. 2022).

With the rise of tools such as ChatGPT,[3] Diagflow,[4] Dall-e,[5] and others, it is even more relevant to consider an ethical framework for AI systems. These tools are having great impact in how we work and perform tasks in everyday life. Because they directly impact people's lives, the ethical context in which they operate becomes more critical (Christoforaki and Beyan 2022). Topics such as the intellectual property of the data used to train algorithms, the false identities that can be created, and how these systems affect work and industry have been discussed (Frank et al. 2019). As these technologies become more ubiquitous, addressing the ethical concerns associated with their use is important. Regulation and establishing solid ethical standards are essential to ensure that AI is used responsibly and with respect for people's rights and well-being (Hickok 2021). Taking steps to ensure ethical AI makes it possible to maximize its potential while protecting human rights and values.

[3] See https://openai.com/blog/chatgpt.
[4] See https://cloud.google.com/dialogflow.
[5] See https://openai.com/research/dall-e.

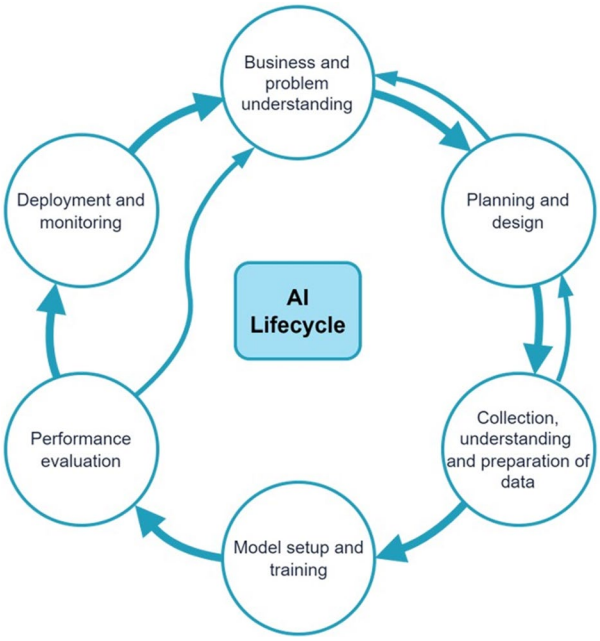

Fig. 1 Stages in the life cycle of an AI-based system. *Note* Created based on Haakman et al. (2021) and Hermosilla et al. (2021)

AI ethics needs to improve to comply with the stipulated ethical principles; it lacks control and reinforcement mechanisms, and infractions do not have significant consequences (Hagendorff 2020). Additionally, the technical personnel in charge of the development and deployment of the system (data scientists, data engineers, and developers, among others) still do not receive adequate education on the ethical implications and how to make a judgment when developing ML or DL algorithms (Burton et al. 2017). These personnel generally follow the life cycle of the AI-based system (see Fig. 1) or the recent MLOps cycle[6] without explicitly addressing ethical aspects. The latter could arise from inadequate awareness or perception of ethical considerations, rendering it non-essential, and some firms may consider it a hindrance (Morley et al. 2021). The AI life cycle stages consider technical aspects of both the model and data sets. However, it is imperative to incorporate responsible practices in every stage, not only definitions or suggestions, to comply with the principles outlined in Table 1, which serve as the core framework. It is crucial to prevent undesired situations like discrimination resulting from biases in either the model or the dataset (Bogina et al. 2021; Hermosilla et al. 2021) or a lack of transparency caused by system opaqueness and processes (Kroll 2018). All the involved parties, including AI developers and other stakeholders, must acknowledge the importance of ethical principles, their implications, and the risks that emerge when they neglect them. Furthermore, they should possess adequate practical resources to implement AI ethics.

The review will examine the application of AI ethics, current shortcomings and challenges, and theoretical and practical tools to aid in developing AI-based systems. Specifically, the section 'Towards a more ethical design of AI' provides an overview of the current scenario for responsible and ethical AI development and the limitations

[6] See https://ml-ops.org/content/mlops-principles.

of existing approaches, and, finally, a conceptual framework is proposed to compensate for these limitations (focused on operationalization). The 'Methodology' section outlines the systematic review conducted to identify the practical resources and tools that multidisciplinary teams can use when developing systems. Likewise, a detailed explanation of the creation of a graph incorporating multiple typologies to classify the tools discovered is provided. The 'Results' section exposes the main findings of incorporating the tools in the typologies. The 'Limitations and Future Work' section highlights ideas that may help close the gap for implementing AI ethics and the limitations of this review. Lastly, the study concludes by presenting the challenges that need to be addressed to create AI that benefits society.

2 Theoretical background

Most frameworks, guides, and guidelines (even less abstract resources) proposed in recent years by the community to achieve ethical (also called responsible) AI development are based on ethical principles (Hagendorff 2022b). There is a wide variety of these documents, with a plethora of principles (Jobin et al. 2019), that can be "overwhelming and confusing" for AI developers (a small number of the principles can be seen in Table 1). However, some researchers have identified and unified them into more actionable fundamental principles (Becker et al. 2022; Fjeld et al. 2020; Floridi and Cowls 2019). These attempts to clarify the picture, while valuable, remain limited. The level of abstraction of the principles, and even more worryingly, the fact that more technical personnel are unaware of them, means that this approach is still immature (Morley et al. 2021). We are not discouraging or discrediting these valuable efforts; on the contrary, we believe that they are the basis for developing AI governance and should be addressed with methodologies that encourage multidisciplinary development, participation of all stakeholders, and initiatives that support a more practical approach such as the typologies proposed by Morley et al. (2020) or Ayling and Chapman Ayling and Chapman (2021). Their research presents several tools classified into two main categories: fundamental ethical principles and AI lifecycle stages. Morley et al. (2020) identify 106 tools, while Ayling and Chapman (2021) list 39. Similar to our research, these tools vary in nature; some are theoretical, and others abstract. However, they share the goal of providing developers and stakeholders with resources to apply ethics in AI at specific points in development.

Governments are still reserved or cautious about exercising mandatory regulation on these emerging technologies (Marchant and Gutierrez 2022; Maslej et al. 2023; The Law Library of Congress, 2023), probably for two reasons. The first is that since "laws and norms cannot keep pace with code", this may also explain the great variety of existing soft laws (Fjeld et al. 2020, p. 57; Gutierrez and Marchant 2021). The second possible reason is that "policymakers and legislators [sometimes fall, but] need to push against the false logic of the Collingridge dilemma"; therefore, they don't intervene until technologies are fully developed and the use is widespread (Morley et al. 2021, p. 8). In any case, the forms of governance and initiatives to achieve the ethical development of an AI system follow some basic principles. These serve as "normative constraints on the do's and don'ts of [developing and using AI-based systems] in society", noting that normative constraints should apply to all parties involved, rather than just technical personnel (Morley et al. 2020, p. 4).

2.1 Operationalization of AI ethics

As mentioned, the fundamental or high-level principles proposed by governmental institutions (China's 2019; European 2019; Government 2019; Espaÿnol et al. 2021), different private companies (de Laat (2021)) give a broader view of the principles of private companies), international organizations (OECD 2019b; UNESCO 2021), academia (Brundage et al. 2018; Diakopoulos et al. xxx; Floridi et al. 2018; Future 2017), among others, have dominated the field of AI ethics and it is hoped that these principles can prevent AI practitioners from "crossing societal red lines" (Morley et al. 2021, 2020). Ideally, abstract principles would suffice to achieve ethical AI development, but this is not the case. There are several reasons or causes why high-level principles are insufficient. We clarify that there may be more reasons why principles per se are insufficient. The first reason is that most of these documents (guides, guidelines, directives, principles) are limited to defining the "what of ethics", i.e., what must be followed for an AI system to develop ethically, and focus little on "how" to achieve it (Morley et al. 2020). A second possible cause may be because these principles are seen as a replacement for regulation but are not enforceable, so compliance cannot be guaranteed (Rességuier and Rodrigues 2020) or, put another way, ethics "lacks mechanisms to reinforce its normative claims" (Hagendorff 2020). The last cause refers to the fact that in real-world environments, the interests of companies may be above compliance or adoption of the principles. Therefore, they are considered a constraint when developing and implementing AI systems, even creating conflicts with their employees (Hagendorff 2020; Ryan et al. 2022).

With the above in mind, our study gives importance to the first two causes, and we attempt to close further the gap in the operationalization of AI ethics, i.e., how to move from principles to practice (Morley et al. 2020). To address the described gap and achieve our goal, we propose to answer the following questions:

- What tools exist to apply the ethical principles in AI systems?
- How can the above tools be classified according to their real-world application?
- At what level of technological development are these tools?

Achieving ethically aligned results with AI development requires using technical and abstract (non-technical) principle-based tools. We also argue that differentiating between stages of system development (see Fig. 1) will help more technical personnel (Morley et al. 2021). So far, few resources on toolkits can be consulted by the personnel involved to comply with AI ethics, regardless of whether the project is starting or the system is in production. For example, a developer or any person engaged in the system's pre-development or post-development stage may find a helpful tool to measure the impact of the proposed solution (according to the principle of Non-Maleficence). Even so, we know this principle-centered approach to deontological ethics (normative rules to which we must adhere, sometimes seen as obligations) has shortcomings. We believe that it should be accompanied by a virtue-based approach, i.e., "ideals that AI practitioners can aspire to" and at the same time, decrease the active responsibility gap (Hagendorff 2022b, p. 4; Santoni de Sio and Mecacci 2021).

The conceptual framework we suggest as a fundamental basis for the responsible development of AI-based systems can be seen in Fig. 2, and its main features are shown in Table 2. To address AI ethics, we propose a holistic approach of two approaches or

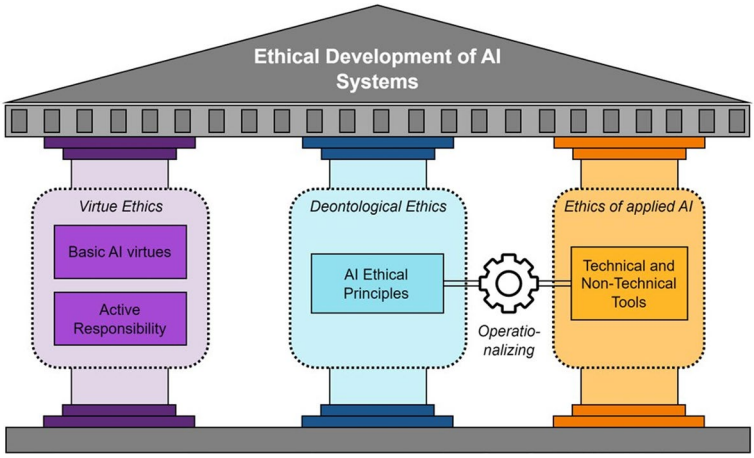

Fig. 2 Pillars for the ethical development of AI

theories of ethics: i) virtue ethics, which "focuses on an individual's character development" (Hagendorff 2022b, p. 19), these virtues create character dispositions that serve as the basis for ethical decision-making and further relate to the concept of active responsibility, which encourages AI professionals to promote and comply with legal standards, goals, and values, as well as to foresee, prevent and avoid unintended consequences (Santoni de Sio and Mecacci 2021); and ii) deontological ethics, based on normative rules to which people must adhere, this approach focuses more on the action and not on the actor (Hagendorff 2022b). These two theories and tools pave the way for applying ethics to AI. Integrating virtue ethics augments ethical growth by relying on the inherent moral character of the individual, which deontology and its external principles fail to tackle entirely (Hagendorff 2022b; Mittelstadt 2019). While the principles of deontology provide external guidelines, virtue ethics goes further, cultivating the internal dispositions that drive the correct behavior. Deontology may present a set of rules to follow, but without ingrained virtue, there is a risk that adherence to those principles will lack depth and commitment (Stahl 2021b; Vallor 2016). For example, a developer might adhere to a list of principles, but without personal solid virtues, he or she might apply them in a lax or superficial manner. The balanced combination of both perspectives, virtues, and principles provides a more holistic framework that guides external actions and drives authenticity and accountability from within. While we focus primarily on the third pillar in our research, we recognize that the other two ethical approaches will also benefit all those involved in AI development and the target population (Morley et al. 2020).

This study aims to provide a resource for people involved in AI development (not just technical staff) so that when operationalizing AI ethics, they are not overwhelmed by unfamiliar notions. To this end, and taking as a reference two studies conducted[7] on AI operationalization (Ayling and Chapman 2021; Morley et al. 2020), we propose to close further

[7] This research focuses on tools aligned with ethical principles and distinguishing between stages of the AI lifecycle.

Table 2 Ethical theories and their characteristics

Virtue ethics	Deontological ethics	
Basic virtues of AI[a]	Corresponding principles[a]	Basic principles of AI[b]
Justice	Algorithmic fairness, non-discrimination, bias mitigation, inclusion, equality, diversity	- Justice: promoting prosperity, preserving solidarity, avoiding unfairness
Honesty	Organizational transparency, openness, explainability, interpretability, technological Conflict of interest, open source, acknowledge errors and mistakes	- Explicability: enabling the other principles through intelligibility and accountability
Responsibility	Responsibility, liability, accountability, replicability, legality, accuracy, considering (long-term) technological consequences	- Beneficence: promoting well-being, preserving dignity, and sustaining the planet
Care	Non-maleficence, harm, security, safety, privacy, protection, precaution, hidden costs, beneficence, well-being, sustainability, peace, common good, solidarity, social cohesion, freedom, autonomy, liberty, consent	- Non-Maleficence: privacy, security, and "capability caution'- Autonomy: the power to decide (to decide)

The principles and virtues were mapped based on each framework's shared concepts and associated principles. For further details, please refer to Table 1 of Hagendorff (2022b) (virtue-based ethical theory) and Sect. 3 of Floridi and Cowls (2019) (deontological ethical theory)

[a]There are four "basic AI virtues," each corresponding to principles. Second-order virtues (prudence and fortitude) that enable practitioners to deal with "bounded ethics" should also be taken into account (see Hagendorff (2022b) for details)

[b]The five basic AI principles are based on classical bioethics principles (see Floridi et al. (2018) for details)

Table 3 Databases and search strings for Phases 1 and 2

Phase	Database	Search strings
1	Scopus IEEE Xplore Google Scholar PubMed Nature SpringerLink	("technical tool" OR "ethical toolkit" OR "technology ethics" OR "ethics tools" OR "tool categories" OR "ethical design tool" OR "assessment risk" OR bias OR audit OR ethics OR transparency OR privacy OR "privacy and data") AND ("artificial intelligence" OR "machine learning" OR "deep learning") AND (review OR "systematic review" OR overview OR "state of the art" OR "systematic mapping")
2	Scopus IEEE Xplore Google Scholar ACM Digital Library Springer Link arXiv	("ethical toolkit" OR "technology ethics" OR "ethics tool" OR "technical tool" OR "ethical design tool" OR "responsible tool" OR ethics OR "ethical framework" OR audit OR evaluation OR "impact assessment" OR instrument) AND ("artificial intelligence" OR "machine learning" OR "deep learning")

Phase 1 was carried out from August 2022 to November 2022, and Phase 2 was conducted between February 2023 and May 2023

the gap between ethical principles and the actions or methods to enforce them. On the one hand, we will update the list of tools since, to the best of our knowledge, there have been no further studies on tools available to enforce AI ethics; on the other hand, we will extend the scope with two new ideas discussed in the next section.

3 Methodology

This research was divided into four phases: phases one and two follow a systematic review methodology, and the third and fourth phases focus on creating a typology and categorizing the tools. Five researchers participated in all stages of the process, and a random exchange of reviewer roles was implemented among them. This practice aimed to receive diverse and equitable feedback, thus promoting a comprehensive and objective evaluation. The level of consensus was achieved by identifying common ground and clarifying disagreements. Each step is described in detail below.

The **first phase** was to review the growing literature on AI ethics to provide a theoretical foundation for the study. This phase also aimed to identify research on tools that emerge from the research that promote pro-ethical[8] AI development, taking into account ethical principles and the various stages of the AI lifecycle.[9] We created a search string and

[8] Floridi (2015) explains that a pro-ethical development or design does not fully constrain agent decision-making, contrary to paternalistic design ethics. In pro-ethical design, the agent receives a nudge to act ethically, but the actual choice rests with the agent.

[9] Before conducting the systematic search, we were aware of the study by Morley et al. (2020) and Ayling and Chapman (2021), but we intended to expand our search.

Table 4 Eligibility criteria for literature and tool selection

Literature eligibility criteria	Tools eligibility criteria
Inclusion criteria were as follows: Documents in English or Spanish language. Articles published between 2017 and 2022. Articles included in the databases in Table 3 (Phase 1)	Inclusion criteria were as follows: Documents or websites in English or Spanish. Articles published between 2015 and 2022. Articles included in the databases in 3 (Phase 2). Applications, systems, or methods that will help to develop ML or DL aligned with the principles described
Exclusion criteria were as follows: Documents unrelated to AI ethics or liability issues (ML and DL algorithms). Articles prior to 2017. Conference proceedings, newspapers, non-academic journals and dissertations. Articles with purely technical tools.[a]	Exclusion criteria were as follows: Applications, systems, or methods focused on Reinforcement Learning (RL) or Artificial General Intelligence (AGI). Articles prior to 2015. Conference proceedings, newspapers, non-academic journals and dissertations. Articles that present tools but lack effective practical implementation (in the case of code libraries).[b]

[a]This item refers to articles with an extensive mathematical description of a software tool, usually with little guidance on its application in practice

[b]The content of this item focuses on articles that describe the operation and construction of algorithmic methods or techniques for solving specific problems. These articles often include experimental code, which may be challenging to implement or comprehend due to minimal maintenance and support, distinguishing them from libraries with ongoing support

explored six databases (see Phase 1 in Table 3) with necessary modifications to accomplish this. The **second phase** entailed searching for tools using an additional search string through six databases (see Phase 2 in Table 3). Additionally, a non-exhaustive exploration was conducted using the Google search engine without the search string previously mentioned, due to its capacity to offer a broader range of information than traditional academic search engines.

The PRISMA methodology (Page et al. 2021) was used for the systematic search described in phases one and two. For their respective selection (on literature and tools), the eligibility criteria shown in Table 4 were applied. The literature and tools selection flow is shown in Figs. 3 and 4, respectively.

Keeping in mind that the tools identified in the second phase needed to be aligned with the ethical principles mentioned in Table 2, the stages of the AI life cycle (refer to Fig. 1), the type of algorithm task, and two additional categories described below, the **third phase** aimed to generate a diagram with various typologies that summarizes all the gathered tools and their relevant information. This diagram can aid AI developers in identifying these resources more efficiently and determining their applicability in specific cases. Inspired by the research of Morley et al. (2020), Fjeld et al. (2020), and Ayling and Chapman (2021), we created a diagram found at Tools4ResponsibleAI, specifically in the 'Typology' section of the website.

Each level of the circle is related to an ethical principle (5 levels corresponding to 5 principles, except for the center of the diagram, which is explained in the next paragraph), and each section of the circle represents a stage in the life cycle of an AI (6 sections corresponding to 6 stages). Within each section, there are dots, and inside is a tool identifier

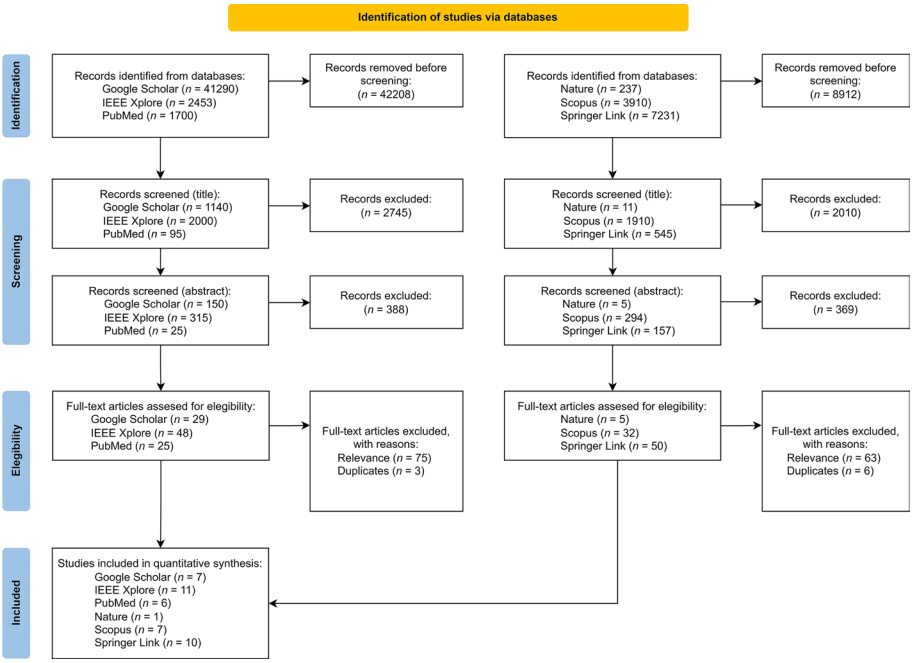

Fig. 3 Flowchart illustrating literature selection.
Note. Flowchart to illustrate our literature selection process, which was based on the PRISMA structure developed by Moher et al. (2010) for systematic reviews. The articles incorporated at this stage were used as theoretical foundations for our study, although they were not entirely cited

number. Pressing a dot displays relevant information, such as the type of task, the sector that created the tool, and the level of development, described below.

The typology includes a sixth principle called Governance, which relates to "establishing and implementing policies, procedures and standards for the proper development, use and management of the infosphere" (Floridi 2018, p. 3). This concept ranges from regulations to social moral values (Ashok et al. 2022). Governance encompasses guidelines and recommendations (Floridi 2018), such as the General Data Protection Regulation (GDPR)[10] or the proposal to implement an AI Law in the European Union (European 2021). This principle is at the center because it unifies the resources that adhere to the other five ethical principles while addressing a range of social, ethical, and normative considerations.

Finally, the **fourth phase** involved synthesizing the information from the tools identified in Phase 2 and framing this collated data into the diagram that was created earlier. To ensure the alignment of the tools with ethical principles and lifecycle stages, we used the description provided by the respective tool. Many of these tools clearly indicate which ethical principle they address and in which phases they are best suited. In addition to the ethical principle definitions by Floridi et al. (2018), Becker et al. (2022), Fjeld et al. (2020), and The Ethics Guidelines for Trustworthy Artificial Intelligence (European 2019), we

[10] Available at https://gdpr-info.eu/.

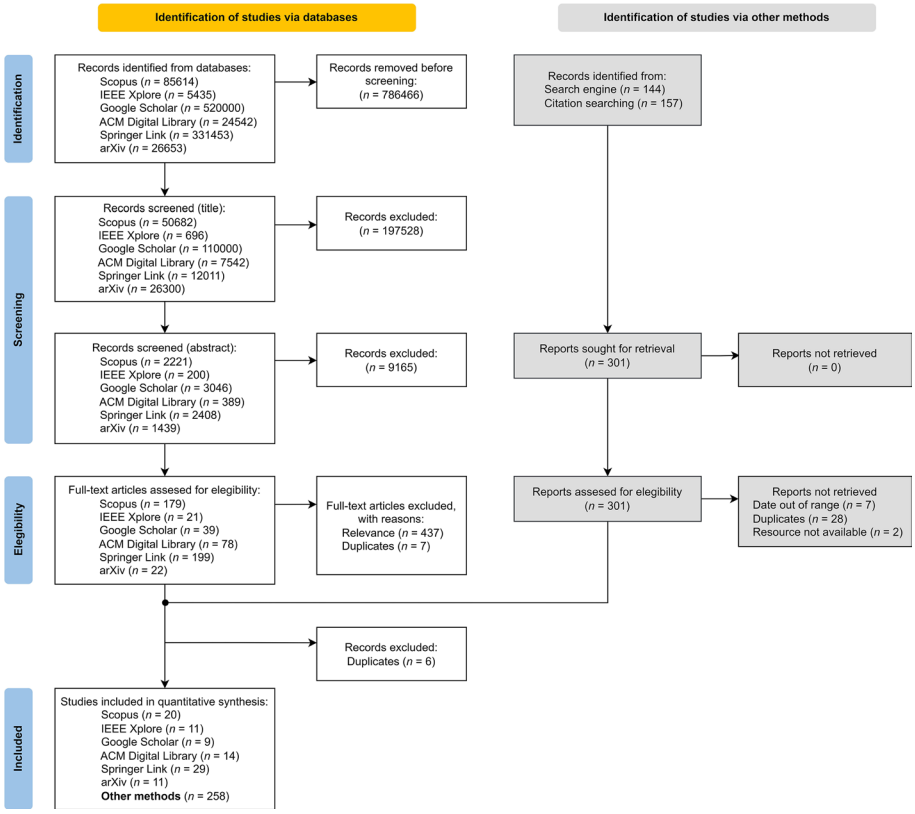

Fig. 4 Flowchart illustrating the tool and resource selection.
Note. Flowchart to illustrate our tool and resource selection process, which was based on the PRISMA structure developed by Moher et al. (2010) for systematic reviews

include our knowledge of AI system development stages and relevant literature. In addition, we outline other relevant typologies such as:

- Task type, i.e., the task(s) the tool or resource addresses (classification, regression, clustering, NLP, among others). In some cases, as in theoretical resources, it is mentioned that its use is directed toward AI-based applications, but the type of task is not specified.
- Sector that developed the tool (private, public, NGO, or academic).
- Level of development of the tool. It was determined according to its functionality, documentation and examples, updates, and requirements. Some of these variables do not apply to theoretical tools. The level of development is classified into ranks: level 1 Insufficient, level 2 Basic, level 3 Intermediate, level 4 High, and level 5 Advanced. Passing level three (intermediate threshold) is graded every 0.5 points, although the scores are continuous.

The tools identified were classified into three distinct categories, each playing a specific role in promoting ethical AI. Firstly, hybrid tools blend technical and non-technical

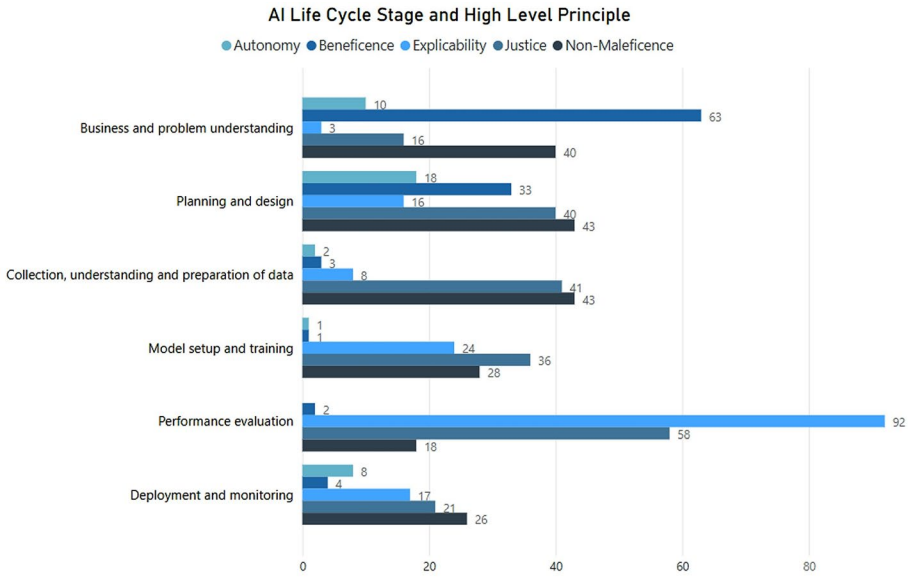

Fig. 5 Tools by AI lifecycle stage and high-level principle.
Note. The governance principle is not included due to its holistic scope described above. The graph was made according to the occurrence of each principle. An example of a correct graph interpretation is that 92 tools can facilitate compliance with the Explicability principle in the Performance Evaluation stage

approaches, including books, checklists, courses, examples, guides, resource compendiums, tools compendiums, tutorials, and websites. These tools provide resources that integrate theoretical comprehension with practical applicability, furnishing developers and stakeholders with a comprehensive outlook.

Secondly, the non-technical tools cover a broad spectrum, including articles, case studies, certifications, codes of ethics, codes of practice, contractual terms, guidelines, laws, licenses, playbooks, practical frameworks, principles (documents with their respective definitions, and sometimes with theoretical guidance), reports, standards, theoretical frameworks, white papers and theory documents (documents that did not fully fit into the other types but were chosen not to be mixed with too many different types). These tools focus on providing conceptual, ethical, and legal guidance, establishing a solid framework for the ethical development of AI.

Thirdly, technical tools encompass codes, apps, datasets, programs (desktop and web-based programs), and voice assistants. This group is oriented towards development's practical and technical aspects, providing concrete and applicable solutions. Codes in this group, for example, may include open-source libraries that facilitate the implementation of ethical practices in AI systems. At the same time, datasets improve models' quality and fairness. The following section presents graphs that illustrate the three macro categories and each type of tool.

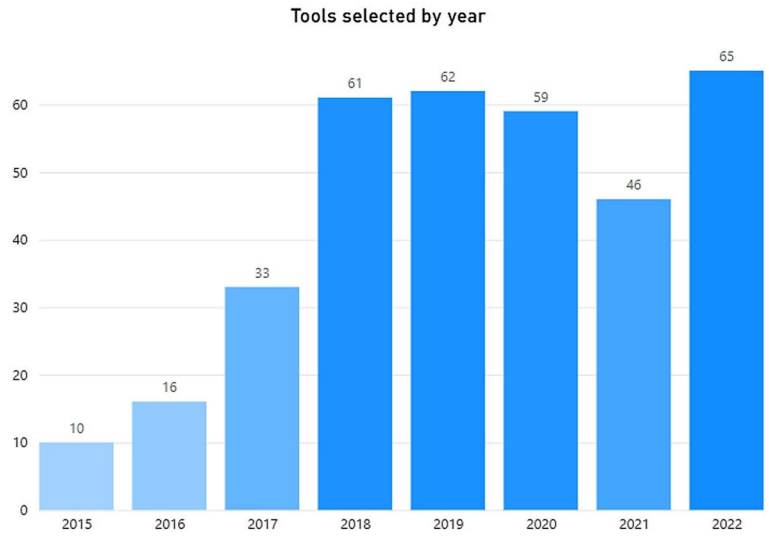

Fig. 6 Distribution by year of the resources found in the study

Table 5 AI lifecycle stages and high-level principles identified in the tools

AI life cycle stage	Number of tools	High level principle	Number of tools
Business and problem understanding	101/352 (28.7%)	Beneficence	95/352
Planning and design	113/352	Non-maleficence	159/352
Collection, understanding and preparation of data	93/352	Justice	167/352
Model setup and training	89/352	Autonomy	36/352
Performance evaluation	159/352	Explicability	124/352
Deployment and monitoring	72/352	Governance	26/352

4 Results

Our research identified 352 tools (from now on referred to interchangeably as resources) that have made it possible to complete the typology described in the previous section. Each resource was examined and categorized based on the life cycle stage, high-level principle, resource type, task type, and level of development. Other data were also collected, such as the sector that created the tool, year of creation, links of interest, and programming language (applicable to technical tools). The complete data set is shown in Appendix A, and an interactive version of the diagram with the different typologies -still under development- can be seen at https://ricardo-ob.github.io/tools4responsibleai. The main points of the search and framing of the resources found are highlighted below.

Figure 5 illustrates the six stages of the AI life cycle and the five high-level principles (hereafter, stage and principle, respectively). The figure reflects more resources because each stage may present several principles. The above is a distinguishing feature of the resources; they are not monothematic, at least not for the most part, but span numerous principles and stages. In the first stage, many resources focus on beneficence

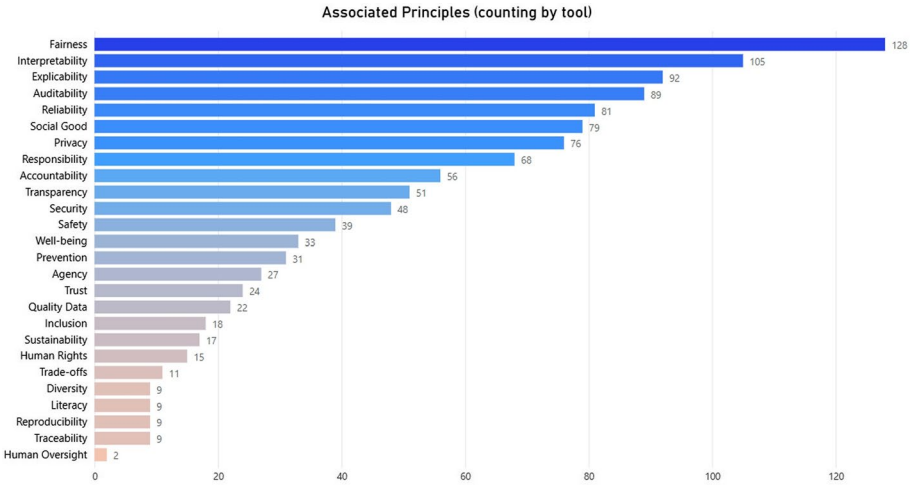

Fig. 7 Count of low-level or associated principles.
Note. The associated principles were counted by their identifier. A tool may have more than one associated principle depending on the stage at which it is used

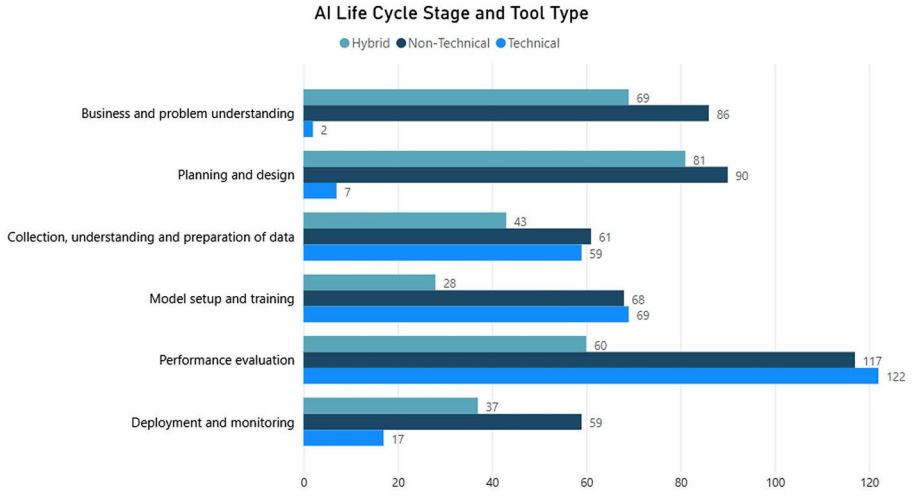

Fig. 8 Type of tool according to AI life cycle stage

and Non-maleficence, where the AI-based solution's objectives, scope, and approach are defined. The principles of Non-maleficence and justice maintain a notorious relevance in the other five stages. About 57% of the tools focus on helping to meet these two principles throughout the lifecycle. Table 5 shows the number of principles and stages identified in each tool, showing the most frequent principles (Justice, Non-maleficence, Explicability) and the most frequent stages per tool (Performance Assessment, Planning and Design, Understanding the Business and the Problem).

Figure 6 shows a significant increase in production from 2018 onwards, doubling the output compared to the previous year. This upward trend has persisted, despite a slight

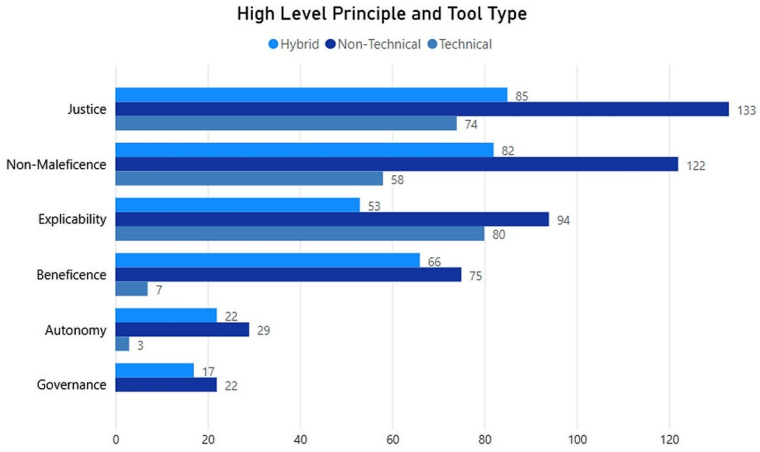

Fig. 9 Tool type according to high-level principle

decrease in 2021. In contrast, the creation of such tools was considerably more restricted during 2015 and 2016 compared to 2018.

The high-level principles also have associated or low-level principles (see Fig. 7); 26 associated principles were compiled. The main one is the principle of Equity, with 128 tools addressing issues related to the bias in models and data sets. The principles of Interpretability (present in 105 tools) and Explicability (present in 92 tools) are the next in number. The latter two focus on the Performance Evaluation and Model Configuration and Training stages (see Explicability principle in Fig. 5). It can be seen that many low-level principles are summarized in the high-level principles of Justice, Explicability, and Non-maleficence.

Figure 8 shows a count of tools by type (three macro categories as described in the methodology section) for each stage in the life cycle addressed by the tool. It is evident that for the first two stages, there is a significant representation of hybrid and non-technical tools (frameworks, guidelines, reports, articles, among others), while for the following three stages, there is a notable presence of technical tools (mostly code libraries). In the complete cycle, it is evident that non-technical tools predominate, mostly resources that introduce theoretical concepts, principles, requirements, or methodologies for developing responsible AI.

The identification of the types of tools and their presence in the high-level principles can be seen in Fig. 9. The Justice principle contains the highest number of non-technical tools, about 79% concerning the total number of tools in this principle (see principles in Table 5), followed by Non-maleficence with ~76% and Explicability with ~75%. It is essential to mention that many technical tools, such as code libraries, are accompanied by research articles (classified as non-technical tools). Therefore, the percentages of technological tools in the three predominant principles (~44% in Justice, ~64% in Explicability, and ~36% in Non-maleficence) may be at the level of non-technical tools and not correspond to the proportion shown in Fig. 9. Figure 10 shows that the top two resources are scientific articles and code libraries, where 217 resources contain research articles, and 162 are or include code libraries, although these are not mutually exclusive.

Finally, Fig. 11 shows the number of tools by the sector that created them and their level of development, where it can be seen that the academic sector is the primary producer of

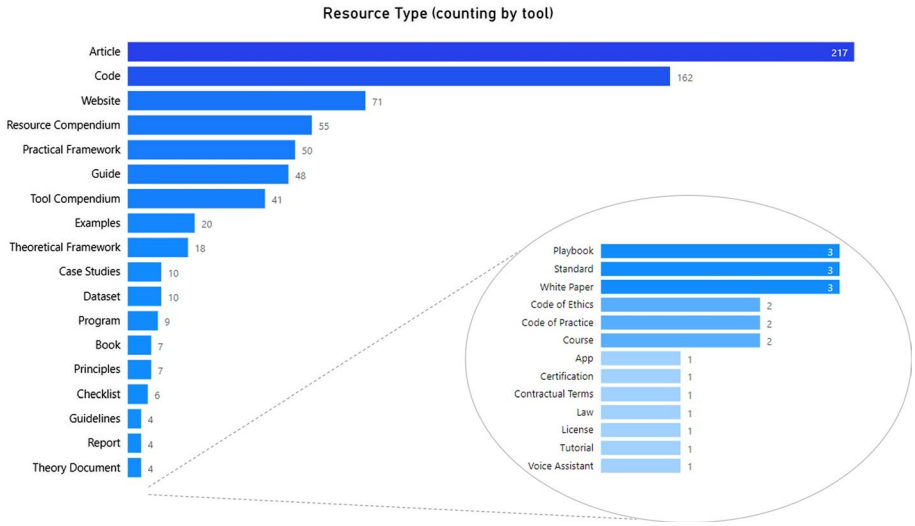

Fig. 10 Tool type count by quantity found

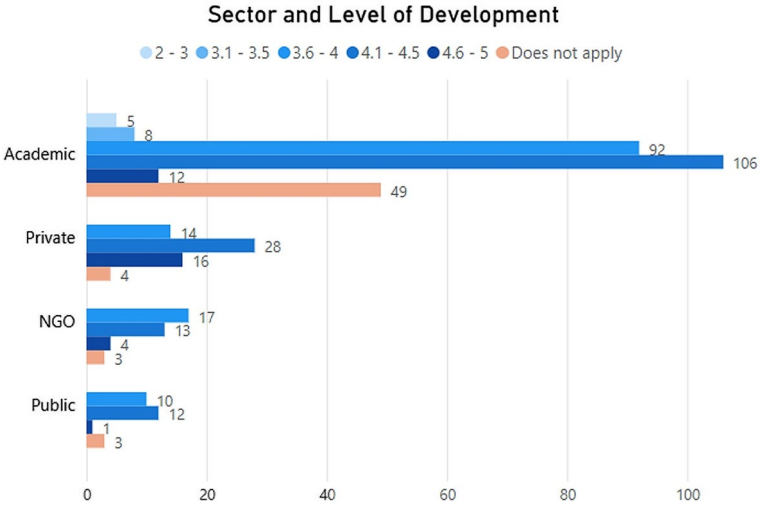

Fig. 11 Number of tools according to the level of development and the sector that created them.
Note. Some tools do not have a development level because the criteria defined for their classification were not applied. The content is too abstract or theoretical, so it would be pertinent to have stricter measures to classify these tools, which is beyond the scope of our research

tools. This sector has 106 tools with a high level of development. The private sector has the most significant number of tools at a high-advanced level ($n = 16$) compared to the public sector ($n = 12$), non-profit organizations ($n = 14$), and the public sector ($n = 1$). The graph also shows that, although the academic sector is the leading producer of tools, about 39% are at insufficient-high levels. On the other hand, the private sector has ~22% of its tools

in the insufficient-high levels. The Public and Non-Profit sectors are among the lowest producers of tools, with most of their tools at the intermediate-high levels.

It is necessary to clarify that the sectors are presented separately for graphical representation purposes, even though about 12% of the tools were created in alliance with the different sectors. This collaboration between academic, private, and public research is mainly due to companies funding research in universities, private companies with interests in the point of view of non-profit organizations, or academic researchers hired by government institutions.

The research articles taken as a basis for the theoretical development, the tools obtained from databases, citation search, and search engine (Google Search) can be found in Appendix B, Appendix C, and Appendix D, respectively.

5 Discussion

During our research and analysis of results, we observed a rising trend in producing ethical tools. Each resource attempts to solve one or more social and ethical issues at different stages of the cycle, depending on the context of each project. This section discusses the main findings, which are addressed in detail in the following four sections. It concludes with a short assessment of how the compendium of resources should be interpreted and used.

5.1 The academic sector as main tool contributor

Many tools of a 'high-advanced' development level are produced by academia, corresponding primarily to research papers, frameworks, and open-source libraries. The sector makes the most tools, followed by the private sector, with about 66% of the resources found (not including collaborations). Among some tools developed exclusively by the academic sector, we highlight the assessment model centered on human rights (Human Rights, Ethical and Social Impact Assessment-HRESIA) proposed by Mantelero (2018), the Medical Algorithmic Audit scheme (Liu et al. 2022), the Dataset Nutritional Label (Holland et al. 2018), the SHAP open-source library (Lundberg and Lee 2017), the DaRe4TAI framework (Thiebes et al. 2021), the Aequitas bias audit toolkit (Saleiro et al. 2018), and the FairSight visual analytics for equity system (Ahn and Lin 2020). These tools have a multidisciplinary and cross-sectoral target audience, so stakeholders are considered. Private sector tools have a higher level of development, albeit in lesser quantities. The last may respond to a concern about strict regulation and ethical washing (de Laat 2021; Floridi 2019) or, on the other hand, to generate value for their companies by obtaining commercial benefits (Mills et al. 2021).

Although the public sector is the primary beneficiary of these resources, it has a low production of ethical tools, and most of them have a below-average level of development. Despite this low production of tools, it is notable that governments in the last six years have increased their initiatives and results in AI regulation (Maslej et al. 2023). Governments must ensure social good by focusing on equity, accountability, sustainability, privacy, and security. Therefore, they see the need to use and produce tools, including responsibly developed AI-based systems, to improve the quality and reliability of bureaucratic processes.

Finally, sectors see a significant advantage in collaborations. We found that the tools created through partnership are mostly at 'high-advanced' levels, in contrast to the 'intermediate-high' levels developed by a single sector. The most notable resources resulting from cross-sector work are the DEDA dialogic framework (Franzke et al. 2021), the open source library InterpretML (Nori et al. 2019), the AI-RFX contracting framework (Institute yyy), the interdisciplinary framework proposed by Fetic et al. (2020), the HAI dashboard for addressing agile ethics in AI created by Butnaru et al. (2018), the PySyft open source library (OpenMined 2018), and the Evaluation Checklist for Trusted Artificial Intelligence (Ala-Pietilä et al. 2020).

5.2 Gaps in tool types and stages

There is a large gap between the type of tools in the first two stages of the life cycle (see Fig. 8), where non-technical resources stand out, and in the following four stages, technical tools become more relevant. On the other hand, hybrid tools show a constant permanence in the different stages of the life cycle. The absence of technological resources in the first two stages poses a need to develop tools that can be used to understand the problem to be solved and the scope of the project without neglecting the deployment and monitoring stage. Although the resources in the described stages are usually theoretical, implementing automated processes -correctly planned- to assist and evaluate these stages would allow streamlining processes that can be laborious for development teams unfamiliar with social and ethical domains.

The number of tools is also not equally distributed in stages. According to Table 5, the third, fourth, and sixth stages have approximately a quarter of the total tools, meaning that data processing, model training, and model implementation have fewer resources and methodologies available to developers and decision-makers. However, the issues associated with data processing were the first approaches in the development of ethical AI (Butterworth 2018), and its maturity is expected to complement the imbalance in the number of resources.

5.3 Justice, non-maleficence, and explicability as generalized principles

To the principle of explicability, described as the only "all-encompassing principle" in the review by Morley et al. (2020), are added the principles of Justice and Non-maleficence (see Figs. 5 and 9). Therefore, the distribution of high-level principles is also uneven. An extensive literature on AI project requirements (Fjeld et al. 2020; Hagendorff 2020; Jobin et al. 2019) gives evidence of the need for such principles. Figure 7 exposes the associated principles of fairness, interpretability, explicability, reliability, security, privacy, and accountability, among others, as the most recurrent in the tools found.

Various technical resources have recently been developed to audit biases in models and datasets (Hagendorff 2020; Lee and Singh 2021; Mehrabi et al. 2021). These tools allow for identifying biases at individual and group levels through statistical metrics and even allow mitigation during the training stage. In the Justice principle, the most well-represented stages in resource allocation are performance evaluation, data processing, and planning and design, with most of these resources consisting of post-hoc analysis tools. The tools that stand out the most are Amazon's SageMaker Clarify (Hardt et al. 2021), IBM's AI Fairness 360 (Bellamy et al. 2018), community-driven Fairlearn and Microsoft (Agarwal et al. 2018), Fairness Indicators created by the TensorFlow community and Google

(TensorFlow 2019), the AEKit toolkit (Krafft et al. 2021), the UnBias toolkit (Lane et al. 2018), and FairML (Adebayo 2016).

Fairness, Explainability, and Interpretability remain principles with many technical resources (Guidotti et al. 2018; Namatēvs et al. 2022; Ras et al. 2022). Hagendorff (2020) suggests one possible reason for this is the ease with which the principles can be mathematically operationalized. On the other hand, there is a shortage of tools that aid in upholding principles like data quality, prevention, agency, or inclusion. These principles are highly significant in fostering confidence while dealing with sensitive data and responding to public sector initiatives. Autonomy, a principle in about 10% of all tools, is often neglected. Ensuring individuals are well-informed and free to make their own decisions is crucial. A system based on AI that "limits people's autonomy will discriminate against them, and if these discriminations are not addressed, they will not be detected nor made visible" (Subías-Beltrán et al. 2022, p. 15).

Although not a principle, governance is found in about 7% of the tools. It is holistic in unifying resources that adhere to the five core principles. It is essential to ensure that AI is developed and used positively for society. It becomes the foundation that underpins integrity and trust in the development and use of AI (Taeihagh 2021). Its inclusion highlights the importance of establishing sound structures to guide the ethical evolution of AI for social welfare and respect for human rights (Mäntymäki et al. 2022).

Ethical aspects such as transparency, accountability, and traceability are present in most non-maleficence stages (see Figs. 5 and 7), which shows that it is necessary and desirable to provide information about the decisions taken (at least the most important ones), who is responsible, and what measures are taken to counteract or mitigate unwanted effects.

5.4 Increase in the process of collecting and classifying resources

In our initial search, no research distinguished between tools, stages, and principles, but we found significant progress on the part of the community. The development of tools to facilitate the search for and selection of ethical tools for projects is becoming increasingly common. We highlight the following compendiums that distinguish between tools, stages, and principles: AI Ethics Tool Landscape (Wenink 2021), the comprehensive Catalogue of Tools & Metrics for Trustworthy AI (OECD 2018), the PLOT4AI threat modeling library and methodology (Barberá 2022), the interactive guide to ethical mitigation strategies proposed by Boyd (2022), and the review of methods and tools by Kaur et al. (2022). In addition, resources that recommend tools at a particular stage (without mentioning principles) or, on the contrary, focus on one principle only are becoming more common (Baxter 2019; BSA 2021; CNIL 2019; Corrêa 2021; NIST 2021; Thoughtworks 2021).

5.5 Final considerations

The proposed conceptual framework could serve as a helpful foundation for generating ethical AI development, as virtue ethics could complement deontology and address its limitations. As previously stated, deontological ethics prioritizes adherence to rules or principles, but these principles lack mechanisms to support their normative claims. Furthermore, their content is highly abstract and does not guide achieving these principles (Hagendorff 2020, 2022b). The principles of deontology serve as external guidelines, while virtue ethics develops internal dispositions that encourage appropriate behavior (Shafer-Landau 2012; Besser-Jones and Slote 2015). This integration would be

unproblematic when acting rightly, as a person's moral virtues would drive authenticity and responsibility from within.

The connection between the two theories has been previously addressed. According to Camps (2015), virtues are necessary to effectively ensure principled ethics or great values function. Although this research pertains to deontological conduct in healthcare practitioners, it has applicability to AI ethics as the five elementary principles of AI are generated from bioethical principles. Furthermore, Camps (Camps 2015, p. 6) affirms that virtues, as they are more related to personal development than abstract principles, values, or norms, serve two crucial objectives. Firstly, fundamental principles are broadened and extended; however, merely recognizing them is insufficient. Proper implementation of these principles is crucial, and this is where virtues prove essential in ensuring these principles are reflected in daily conduct, particularly in challenging situations. Being a competent professional requires acknowledging these ethical principles and having the disposition to behave correctly (Camps 2015; Vallor 2016).

Sganzerla et al. (2022) assert that virtues do not hold inherent value but possess instrumental value, as individuals who exhibit greater virtue are more likely to adhere to rules. The researchers also discuss the recognition by Beauchamp and Childress (who proposed the four bioethics principles) of the necessity of virtue ethics in laying the groundwork for principlism. Without considering virtue ethics, attaining objectives aimed at the betterment of society is rendered more challenging. Furthermore, they not only strengthen principled practice but also often constitute the condition for its correct application, given the variety of circumstances that may arise because principles cannot provide a clear guideline to follow, and it is up to the agent to judge what should be done (Camps 2015, p. 7). In ethical AI development, integrating virtue ethics with deontological ethics can aid stakeholders in creating ethical systems in their design and application (MacIntyre 2007).

On the other hand, our main contribution focuses on the diagram, the resources found, and their correspondence with various ethical principles and stages of development. The diagram is intended as a compendium to facilitate the application of abstract ethical principles. The diagram is hoped to be under constant evaluation, to which the community can contribute. The tool's usefulness extends to technical professionals and high-level decision-makers, providing an overview of resources that lead to informed decision-making. However, it is crucial to note that simply using one or more tools does not guarantee responsible AI development. It is essential to avoid technical or methodological solutionism, the mistaken belief that complex socio-technical and political problems can be solved or avoided entirely by applying new techniques (Hagendorff 2022a; Santoni de Sio and Mecacci 2021). Although the gap between ethical tools and ethical AI development persists (Morley et al. 2021), our research aims to bridge it by collecting and categorizing resources, some of which are still unknown to the community. However, we acknowledge that a comprehensive solution to this challenge requires a broader and more holistic approach. Our conceptual framework (Fig. 2) aims to address this challenge from a holistic perspective. This approach aims to align the selection of ethical tools with a sound conceptual framework to drive more effective and sustainable ethical development.

5.6 Limitations and future work

The limitations of our research are centered on i) an extensive search for resources and their corresponding search string, ii) the qualitative relationship of ethical principles with

the resources found, and iii) the exclusion criteria defined and the search method used. Although we intended to compile a considerable amount of resources, our work was based on too broad search strings, making the filtering and selection process too extensive, with results appearing for the most part from the areas of medicine, economics, or social sciences (outside ethics and AI). Therefore, narrowing and improving the search strings may be a starting point for future work in systematic reviews.

The number of resources found is diverse, ranging from theoretical papers to repositories with code. In some cases, framing the resource/tool with the principles and stage was straightforward (primarily with technical tools). However, because our background is in engineering and computer science, biases could have been introduced during the qualitative analysis to determine which ethical principle applied to the resource. Nevertheless, we believe that the framing and analysis of each resource in the typology was done best and shows valuable resources for developers and decision-makers. Finally, the omission of valuable resources is not excluded from our research. Exclusion criteria such as language, databases, and Google search engine made us overlook valuable resources. Other factors with significant impact are the year, as we did not include 2023 (a year marked with substantial advances in generative AI) and outcomes other than ML, DL, and data.

In future work, we propose more rigorous filtering of the resources found to reduce the resources that do not contribute to a particular principle, stage, or even to the general purpose of responsible AI development. We also propose to create methodologies that complement the typology, such as the tool based on filters, objectives, and stages presented by Boyd (2022), to facilitate the selection of tools for the interested person and not being overwhelmed by the plethora of resources. Another possible future helpful work is that analyzing each resource is an essential step in facilitating the choice of tools. A possible outcome would be a new classification of resources by difficulty of use or implementation since the functioning of each resource needed to be thoroughly inspected.

6 Conclusion

This review has focused on gathering information on the current state of AI ethics, continuing with the creation of a base framework for the responsible development of AI-based solutions, and then focusing on the search for tools, where we were able to find 352 ethical tools that address different areas and have different levels of technological development that facilitate the implementation or putting into practice the myriad of abstract ethical principles. The resources found were classified into different typologies, such as the AI lifecycle stage, ethical principles, applicable tasks, and development sector, among others, to facilitate the search and selection of resources for development groups.

The tools are not evenly distributed in terms of both stages and principles. Many tools focus on the first, second, and fifth stages of the life cycle, and the most frequent principles are Justice, Non-maleficence, and Explicability, with Equity and Explicability being the low-level principles with the most tools (technical and non-technical). The principles of governance and autonomy present a deficit of resources, which are paramount in creating legal frameworks and generating trust among the target audience (Kaur et al. 2022). The academic and private sectors are the leading creators of tools and academia being the sector with the most technologically advanced tools, with the resources of private companies being the most developed.

Finally, the compendium and the classification of resources resulting from the review do not seek to create or provide a single solution for the ethical development of AI-based systems, as this is a technology of a socio-technical nature. Each stage needs to be reflected upon and determine which ethical principles are most relevant (depending on the context). We hope that the ML community, less familiar with ethical issues, will find the tools useful, and emphasise the need for comprehensive academic training to shape the basic virtues of AI and broader dissemination of such resources (Morley et al. 2021).

Appendix A Classified resources and tools

https://bit.ly/3ZcADOQ.

Appendix B Theoretical basis articles

https://bit.ly/3PshJA3.

Appendix C Tools from databases

https://bit.ly/3R64BSz.

Appendix D Tools from google search engine and citation searching

https://bit.ly/3PpwR1e.

Acknowledgements This research was supported by: National Agency for Research and Development, ANID + Applied Research Subdirection/ IDeA I+D 2023 Grant [folio ID23110357]. IDB Lab, [Project ATN/ME-18240-CH] Algoritmos Éticos, Responsables y Transparentes. ANID PIA/BASAL FB0002. ANID/PIA/ANILLOS ACT210096. Clasificación de los estadios del Alzheimer utilizando Imágenes de Resonancia Magnética Nuclear y datos clínicos a partir de técnicas de Deep Learning [873-139] Universidad Autónoma de Manizales, Manizales, Colombia. ORIGEN 0011323. Additionally, we express our gratitude for the invaluable contribution of an anonymous reviewer whose meticulous review significantly improved the quality of our article.

Author Contributions Tabares-Soto guided and structured the proposal for the systematic review based on his experience in previous work and research. Ortega-Bolaños and Bernal-Salcedo were in charge of structuring and constructing the six sections of the review with the support of Germán Ortiz and Galeano Sarmiento. The figures were developed by Ortega-Bolaños and verified by Bernal-Salcedo. The complementary material was done by Ortega-Bolaños, Bernal-Salcedo, Germán Ortiz and Galeano Sarmiento. Tabares-Soto and Ruz served as reviewers of the manuscript and supplementary material.

Declarations

Conflict of interest The authors declare no competing interests.

References

Adebayo JA (2016) FairML: ToolBox for diagnosing bias in predictive modeling (Doctoral dissertation). Massachusetts Institute of Technology. https://github.com/adebayoj/fairml

Agarwal A, Beygelzimer A, Dudik M, Langford J, Wallach H (2018) A reductions approach to fair classification. In Dy J, Krause A (Eds.) Proceedings of the 35th international conference on machine learning, PMLR, pp 60–69 https://proceedings.mlr.press/v80/agarwal18a.html

Ahn Y, Lin Y-R (2020) FairSight: visual analytics for fairness in decision making. IEEE Trans Vis Comput Gr 26(1):1086–1095. https://doi.org/10.1109/TVCG.2019.2934262

Ala-Pietilä P, Bauer W, Bergmann U, Bieliková M, Boujemaa N, Bonefeld-Dahl C, Bonnet Y, Bouarfa L, Brunessaux S, Chatila R, Coeckelbergh M, Dignum V, Floridi L, Gagné J-F, Giovannini C, Goodey J, Haddadin S, Hasselbalch G, Heintz F, Yeung K (2020) The Assessment List for Trustworthy Artificial Intelligence (ALTAI). European Commission. https://doi.org/10.2759/002360

Ashok M, Madan R, Joha A, Sivarajah U (2022) Ethical framework for artificial intelligence and digital technologies. Int J Inf Manag 62:102433. https://doi.org/10.1016/J.IJINFOMGT.2021.102433

Ayling J, Chapman A (2021) Putting AI ethics to work: Are the tools fit for purpose? AI Ethics 2(3):405–429. https://doi.org/10.1007/S43681-021-00084-X

Barberá I (2022) Privacy Library of Threats 4 Artificial Intelligence. https://plot4.ai/

Baxter K (2019) Ethical AI frameworks, tool kits, principles, and certifications-Oh my! https://blog.salesforceairesearch.com/frameworks-tool-kits-principlesand-oaths-oh-my

Becker SJ, Nemat AT, Lucas S, Heinitz RM, Klevesath M, Charton JE (2022) A Code of Digital Ethics: laying the foundation for digital ethics in a science and technology company. AI Soc 1:1–11. https://doi.org/10.1007/S00146-021-01376-W

Bellamy RKE, Dey K, Hind M, Hoffman SC, Houde S, Kannan K, Lohia P, Martino J, Mehta S, Mojsilovic A, Nagar S, Ramamurthy KN, Richards J, Saha D, Sattigeri P, Singh M, Varshney KR, Zhang Y (2018) AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. arXiv. https://doi.org/1048550/arxiv:1810.01943

Benefo EO, Tingler A, White M, Cover J, Torres L, Broussard C, Shirmohammadi A, Pradhan AK, Patra D, Tingler A, White M, Broussard C (2022) Ethical, legal, social, and economic (ELSE) implications of artificial intelligence at a global level: a scientometrics approach. AI Ethics 2(4):667–682. https://doi.org/10.1007/S43681-021-00124-6

Besser-Jones L, Slote M (2015) The routledge companion to virtue ethics. Routledge, Routledge

Bogina V, Hartman A, Kuflik T, Shulner-Tal A (2021) Educating software and AI stakeholders about algorithmic fairness, accountability, transparency and ethics. Int J Art Intell Educ 32(3):808–833. https://doi.org/10.1007/S40593-021-00248-0

Botes A (2000) A comparison between the ethics of justice and the ethics of care. J Adv Nurs 32(5):1071–1075. https://doi.org/10.1046/J.1365-2648,2000.01576.X

Boyd K (2022) Designing up with value-sensitive design: building a field guide for ethical ML development. In: Proceedings of the 2022 ACM conference on fairness, accountability, and transparency, pp 2069–2082 https://doi.org/10.1145/3531146.3534626

Brundage M, Avin S, Clark J, Toner H, Eckersley P, Garfinkel B, Dafoe A, Scharre P, Zeitzoff T, Filar B, Anderson H, Roff H, Allen GC, Steinhardt J, Flynn C, HÉigeartaigh SÓ, Beard S, Belfield H, Farquhar S, Amodei D (2018) The malicious use of artificial intelligence: forecasting, prevention, and mitigation. https://doi.org/10.48550/arxiv.1802.07228

BSA (2021) Confronting Bias: BSA's Framework to Build Trust in AI (tech. rep.). https://ai.bsa.org/confronting-bias-bsas-framework-to-build-trust-in-ai

Buchanan B (2019) Artificial intelligence in finance. The Alan Turing Institute. https://doi.org/10.5281/zenodo.2626454

Burton E, Goldsmith J, Koenig S, Kuipers B, Mattei N, Walsh T (2017) Ethical considerations in artificial intelligence courses. http://arxiv.org/abs/1701.07769

Butnaru C, Theodorou A, Benrimoh D (2018) Agile Ethics for AI. Humans in AI (tech. rep.). https://trello.com/b/SarLFYOd/agile-ethics-for-ai-hai

Butterworth M (2018) The ICO and artificial intelligence: the role of fairness in the GDPR framework. Comput Law Secur Rev 34(2):257–268. https://doi.org/10.1016/j.clsr.2018.01.004
Camps V (2015) Los valores éticos de la profesión sanitaria. Educ Méd 16(1):3–8. https://doi.org/10.1016/j.edumed.2015.04.001
Cath C (2018) Governing artificial intelligence: ethical, legal and technical opportunities and challenges. Philos Trans R Soc A Math Phys Eng Sci. https://doi.org/10.1098/RSTA.2018.0080
Cath C, Wachter S, Mittelstadt B, Taddeo M, Floridi L (2018) Artificial Intelligence and the 'Good Society': the US, EU, and UK approach. Sci Eng Ethics 24(2):505–528. https://doi.org/10.1007/s11948-017-9901-7
Chen L, Chen P, Lin Z (2020) Artificial intelligence in education: a review. IEEE Access 8:75264–75278. https://doi.org/10.1109/ACCESS.2020.2988510
China's Ministry of Science and Technology (2019) China: AI Governance Principles Released. https://www.loc.gov/item/global-legal-monitor/2019-09-09/chinaai-governance-principles-released/
Christoforaki M, Beyan O (2022) AI Ethics-A bird's eye view. Appl Sci. https://doi.org/10.3390/app12094130
CNIL (2019) GDPR Toolkit. https://www.cnil.fr/en/gdpr-toolkit
Corrêa NK (2021) Artificial Intelligence Ethics and Safety: practical tools for creating ôgoodö models. arXiv. https://arxiv.org/vc/arxiv/papers/2112/2112.11208v1.pdf
Cummings CL, Mercurio MR (2010) Ethics for the pediatricianautonomy, beneficence, and rights. Pediatr Rev 31(6):252–255. https://doi.org/10.1542/PIR.31-6-252
de Laat PB (2021) Companies committed to responsible AI: From principles towards implementation and regulation? Philos Technol 34(4):1135–1193. https://doi.org/10.1007/s13347-021-00474-3
Devillers L, Fogelman-Soulié F, Baeza-Yates R (2021) AI & human values: inequalities, biases, fairness, nudge, and feedback loops. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 12600 LNCS, 76–89. https://doi.org/10.1007/978-3-030-69128-8_6
Diakopoulos N, Friedler S, Arenas M, Barocas S, Hay M, Howe B, Jagadish HV, Unsworth K, Sahuguet A, Venkatasubramanian S, Wilson C, Yu C, Zevenbergen B (n.d.) Principles for accountable algorithms and a social impact statement for algorithms. https://www.fatml.org/resources/principles-for-accountable-algorithms
European Comission (2019) Ethics guidelines for trustworthy AI (tech. rep.). https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai
European Commission (2021) Laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain union legislative acts
Fetic L, Fleischer T, Grünke P, Hagendorff T, Hauer M, Hauschke A, Heesen J, Herrmann M, Hillerbrand R, Hubig EC, Kaminski A, Krafft T, Loh W, Otto P, Puntschuh M, Hustedt C, Hallensleben S (2020) From principles to practice. An interdisciplinary framework to operationalise AI ethics. https://doi.org/10.11586/2020013
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 (2015) Tolerant paternalism: pro-ethical design as a resolution of the dilemma of toleration. Sci Eng Ethics 22(6):1669–1688. https://doi.org/10.1007/S11948-015-9733-2
Floridi L (2018) Soft ethics, the governance of the digital and the General Data Protection Regulation. Philos Trans R Soc A Math Phys Eng Sci. https://doi.org/10.1098/RSTA.2018.0081
Floridi L (2019) Translating principles into practices of digital ethics: five risks of being unethical. Philos Technol 32(2):185–193. https://doi.org/10.1007/s13347-019-00354-x
Floridi L, Cowls J (2019) A unified framework of five principles for AI in society. Harvard Data Sci Rev. https://doi.org/10.1162/99608F92.8CD550D1
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. Minds Mach 28(4):689–707. https://doi.org/10.1007/S11023-018-9482-5/
Floridi L, Cowls J, King TC, Taddeo M (2020) How to design AI for social good: seven essential factors. Sci Eng Ethics 26(3):1771–1796. https://doi.org/10.1007/S11948-020-00213-5
Frank MR, Autor D, Bessen JE, Brynjolfsson E, Cebrian M, Deming DJ, Feldman M, Groh M, Lobo J, Moro E, Wang D, Youn H, Rahwan I (2019) Toward understanding the impact of artificial intelligence on labor. Proc Natl Acad Sci 116(14):6531–6539. https://doi.org/10.1073/pnas.1900949116

Franzke AS, Muis I, Schäfer MT (2021) Data Ethics Decision Aid (DEDA): a dialogical framework for ethical inquiry of AI and data projects in the Netherlands. Ethics Inf Technol 23(3):551–567. https://doi.org/10.1007/s10676-020-09577-5

Future of Life Institute (2017) Asilomar AI Principles. https://futureoflife.org/openletter/ai-principles/

Galaz V, Centeno MA, Callahan PW, Causevic A, Patterson T, Brass I, Baum S, Farber D, Fischer J, Garcia D, McPhearson T, Jimenez D, King B, Larcey P, Levy K (2021) Artificial intelligence, systemic risks, and sustainability. Technol Soc 67:101741. https://doi.org/10.1016/J.TECHSOC.2021.101741

Government of Canada (2019) Directive on automated decision-making. https://www.tbs-sct.canada.ca/pol/doc-eng.aspx?id=32592

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2018) A survey of methods for explaining black box models. ACM Comput Surv. https://doi.org/10.1145/3236009

Guio Espaÿnol A, Tamayo Uribe E, Gómez Ayerbe P, Mujica MP (2021) Marco Ético para la Inteligencia Artificial en Colombia. https://dapre.presidencia.gov.co/TD/MARCO-ETICO-PARA-LA-INTELIGENCIA-ARTIFICIALEN-COLOMBIA-2021.pdf

Gutierrez CI, Marchant GE (2021) A global perspective of soft law programs for the governance of artificial intelligence. SSRN Electron J. https://doi.org/10.2139/ssrn.3855171

Haakman M, Cruz L, Huijgens H, van Deursen A (2021) AI lifecycle models need to be revised: an exploratory study in Fintech. Empir Softw Eng 26(5):1–29. https://doi.org/10.1007/S10664-021-09993-1

Hagendorff T (2020) The ethics of AI ethics: an evaluation of guidelines. Minds Mach 30(1):99–120. https://doi.org/10.1007/S11023-020-09517-8

Hagendorff T (2022) Blind spots in AI ethics. AI Ethics 2(4):851–867. https://doi.org/10.1007/s43681-021-00122-8

Hagendorff T (2022) A virtue-based framework to support putting AI ethics into practice. Philos Technol 35(3):1–24. https://doi.org/10.1007/S13347-022-00553-Z

Hardt M, Chen X, Cheng X, Donini M, Gelman J, Gollaprolu S, He J, Larroy P, Liu X, McCarthy N, Rathi A, Rees S, Siva A, Tsai E, Vasist K, Yilmaz P, Zafar MB, Das S, Haas K, Kenthapadi K (2021) Amazon SageMaker clarify: machine learning bias detection and explainability in the cloud. In: Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining, pp 2974–2983. https://doi.org/10.1145/3447548.3467177

Henman P (2020) Improving public services using artificial intelligence: possibilities, pitfalls, governance. Asia Pac J Public Admin 42(4):209–221. https://doi.org/10.1080/23276665.2020.1816188

Hermosilla M, González Alarcón N, Pombo C, Sánchez Ávalos R, Denis G, Aracena C (2021). Uso responsable de IA para política pública: manual de formulación de proyectos. https://doi.org/10.18235/0003631

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

Hoadley DS, Lucas NJ (2018) Artificial Intelligence and National Security (tech. rep.). Congressional Research Service Washington, DC. https://a51.nl/sites/default/files/pdf/R45178.pdf

Holland S, Hosny A, Newman S, Joseph J, Chmielinski K (2018) The Dataset nutrition label: a framework to drive higher data quality standards. https://arxiv.org/abs/1805.03677v1

Jia K, Zhang N (2021) Categorization and eccentricity of AI risks: a comparative study of the global AI guidelines. Electron Markets 32(1):59–71. https://doi.org/10.1007/S12525-021-00480-5

Jobin A, Ienca M, Vayena E (2019) The global landscape of AI ethics guidelines. Nat Mach Intell 1(9):389–399. https://doi.org/10.1038/s42256-019-0088-2

Kaur D, Uslu S, Rittichier KJ, Durresi A (2022) Trustworthy artificial intelligence: a review. ACM Comput Surv. https://doi.org/10.1145/3491209

Khan AA, Badshah S, Liang P, Khan B, Waseem M, Niazi M, Akbar MA (2021) Ethics of AI: a systematic literature review of principles and challenges. https://doi.org/10.48550/arxiv.2109.07906

Krafft PM, Young M, Katell M, Lee JE, Narayan S, Epstein M, Dailey D, Herman B, Tam A, Guetler V, Bintz C, Raz D, Jobe PO, Putz F, Robick B, Barghouti B (2021) An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp 772–781 https://doi.org/10.1145/3442188.3445938

Kroll JA (2018) The fallacy of inscrutability. Philos Trans R Soc A Math Phys Eng Sci. https://doi.org/10.1098/RSTA.2018.0084

Kudina O, Verbeek P-P (2019) Ethics from within: Google glass, the collingridge dilemma, and the mediated value of privacy. Sci Technol Human Values 44(2):291–314. https://doi.org/10.1177/0162243918793711

Lane G, Angus A, Murdoch A (2018) UnBias fairness Toolkit. https://doi.org/10.5281/zenodo.2667808

Latonero M (2018) Governing Artificial Intelligence: upholding human rights & dignity
Lee MSA, Singh J (2021) The landscape and gaps in open source fairness Toolkits. In: Proceedings of the 2021 CHI conference on human factors in computing systems. https://doi.org/10.1145/3411764.3445261
Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P (2017) Fair, transparent, and accountable algorithmic decision-making processes. Philos Technol 31(4):611–627. https://doi.org/10.1007/S13347-017-0279-X
Liu X, Glocker B, McCradden MM, Ghassemi M, Denniston AK, Oakden- Rayner L (2022) The medical algorithmic audit. Lancet Digit Health 4(5):e384–e397. https://doi.org/10.1016/S2589-7500(22)00003-6
Loureiro SMC, Guerreiro J, Tussyadiah I (2021) Artificial intelligence in business: state of the art and future research agenda. J Bus Res 129:911–926. https://doi.org/10.1016/J.JBUSRES.2020.11.001
Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30:4765–4774
MacIntyre A (2007) After virtue: a study in moral theory, 3rd edn. University of Notre Dame Press
Mantelero A (2018) AI and Big Data: a blueprint for a human rights, social and ethical impact assessment. Comput Law Secur Rev 34(4):754–772. https://doi.org/10.1016/J.CLSR.2018.05.017
Mäntymäki M, Minkkinen M, Birkstedt T, Viljanen M (2022) Defining organizational AI governance. AI Ethics 2(4):603–609. https://doi.org/10.1007/s43681-022-00143-x
Marchant GE, Gutierrez CI (2022) Soft law 2.0: an agile and effective governance approach for artificial intelligence. Minnesota J Law Sci Technol 24(2):52
Martin K (2019) Ethical implications and accountability of algorithms. J Bus Ethics 160(4):835–850. https://doi.org/10.1007/S10551-018-3921-3
Maslej N, Fattorini L, Brynjolfsson E, Etchemendy J, Ligett K, Lyons T, Manyika J, Ngo H, Niebles JC, Parli V, Shoham Y, Wald R, Clark J, Perrault R (2023) The AI Index 2023 Annual Report (tech. rep.). AI Index Steering Committee. Stanford, CA
Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. ACM Comput Surv. https://doi.org/10.1145/3457607
Mills S, Duranton S, Santinelli M, Hua G, Baltassis E, Thiel S, Muehlstein O (2021) Are you overestimating your responsible AI maturity? (Tech. rep.). BCG. https://www.bcg.com/publications/2021/the-four-stages-ofresponsible-ai-maturity
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 BD, Allo P, Taddeo M, Wachter S, Floridi L (2016) The ethics of algorithms: mapping the debate. Big Data Soc. https://doi.org/10.1177/2053951716679679
Moher D, Liberati A, Tetzlaff J, Altman DG (2010) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. Int J Surg 8(5):336–341. https://doi.org/10.1016/J.IJSU.2010.02.007
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
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, Kinsey L, Elhalal A, Garcia F, Ziosi M, Floridi L (2021) Operationalising AI ethics: barriers, enablers and next steps. AI Soc 1:1–13. https://doi.org/10.1007/S00146-021-01308-8
Morley J, Machado CC, Burr C, Cowls J, Joshi I, Taddeo M, Floridi L (2020) The ethics of AI in health care: a mapping review. Soc Sci Med 260:113172. https://doi.org/10.1016/J.SOCSCIMED.2020.113172
Namatevs I, Sudars K, Dobrajs A (2022) Interpretability versus explainability: classification for understanding deep learning systems and models. Comput Assist Methods Eng Sci 29(4):297–356. https://doi.org/10.24423/cames.518
NIST (2021) AI risk management framework. https://airc.nist.gov/AI_RMF_ Knowledge_Base/AI RMF
Nori H, Jenkins S, Koch P, Caruana R (2019) InterpretML: a unified framework for machine learning interpretability. arXiv. https://arxiv.org/abs/1909.09223
OECD (2018) Catalogue of tools & metrics for trustworthy AI.https://oecd.ai/en/catalogue/overview
OECD (2019a) Accountability (OECD AI Principle). https://oecd.ai/en/dashboards/ai-principles/P9
OECD (2019b) Recommendation of the council on artificial intelligence (tech. rep.). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449
The Law Library of Congress (2023) Regulation of artificial intelligence around the world. https://tile.loc.gov/storage-services/service/ll/llglrd/2023555920/2023555920.pdf
OpenMined (2018) PySyft. https://github.com/OpenMined/PySyft
Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW,

Mayo-Wilson E, McDonald S, Moher D (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ. https://doi.org/10.1136/BMJ.N71
Ras G, Xie N, Van Gerven M, Doran D (2022) Explainable deep learning: a field guide for the uninitiated. J Art Intell Res 73:329–396
Raso F, Hilligoss H, Krishnamurthy V, Bavitz C, Kim LY (2018) Artificial intelligence & human rights: opportunities & risks. SSRN Electron J. https://doi.org/10.2139/SSRN.3259344
Rességuier A, Rodrigues R (2020) AI ethics should not remain toothless! A call to bring back the teeth of ethics. Big Data Soc 7(2):2053951720942541. https://doi.org/10.1177/2053951720942541
Rigano C (2018) Using artificial intelligence to address criminal justice Needs. https://www.ojp.gov/pdffiles1/nij/252038.pdf
Ryan M, Christodoulou E, Antoniou J, Iordanou K (2022) An AI ethics 'David and Goliath': value conflicts between large tech companies and their employees. AI Soc 1:1–16. https://doi.org/10.1007/S00146-022-01430-1
Saleiro P, Kuester B, Hinkson L, London J, Stevens A, Anisfeld A, Rodolfa KT, Ghani R (2018) Aequitas: a bias and fairness audit toolkit. arXiv. https://arxiv.org/abs/1811.05577
Santoni de Sio F, Mecacci G (2021) Four responsibility gaps with artificial intelligence: why they matter and how to address them. Philos Technol 34(4):1057–1084. https://doi.org/10.1007/S13347-021-00450-X
Sganzerla A, Siqueira J, Guérios T (2022) Ética de las virtudes aplicada a la deontología médica. Rev Bioét 30:482–491. https://doi.org/10.1590/1983-80422022303541es
Shafer-Landau R (2012) Ethical theory: an anthology, 2nd edn. Wiley-Blackwell, Hoboken
Stahl BC (2021a) Addressing Ethical Issues in AI. pp 55–79 https://doi.org/10.1007/978-3-030-69978-9_5
Stahl BC (2021) Concepts of ethics and their application to AI. Artificial intelligence for a better future: an ecosystem perspective on the ethics of AI and emerging digital technologies. Springer International Publishing, New York, pp 19–33
Strümke I, Slavkovik M, Madai VI (2022) The social dilemma in artificial intelligence development and why we have to solve it. AI Ethics 2(4):655–665. https://doi.org/10.1007/s43681-021-00120-w
Subías-Beltrán P, Pujol O, de Lecuona I (2022) The forgotten human autonomy in Machine Learning. CEUR Worksh Proc 3221:45–64
Taeihagh A (2021) Governance of artificial intelligence. Policy Soc 40(2):137–157. https://doi.org/10.1080/14494035.2021.1928377
TensorFlow (2019) Fairness indicators - GitHub. https://github.com/tensorflow/fairness-indicators
The Institute for Ethical AI & Machine Learning. (2018). The AI-RFX Procurement Framework. https://ethical.institute/rfx.html
Thiebes S, Lins S, Sunyaev A (2021) Trustworthy artificial intelligence. Electron Markets 31(2):447–464. https://doi.org/10.1007/S12525-020-00441-4
Thoughtworks (2021) Responsible Tech Playbook. https://www.thoughtworks.com/en-us/about-us/social-change/responsible-tech-playbook
UNESCO (2021) Recommendation on the Ethics of Artificial Intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137
Vallor S (2016) Technology and the virtues: a philosophical guide to a future worth wanting. Oxford University Press, Oxford
van Noordt C, Misuraca G (2022) Artificial intelligence for the public sector: results of landscaping the use of AI in government across the European Union. Govern Inf Q 39(3):101714. https://doi.org/10.1016/j.giq.2022.101714
Wang P (2019) On defining artificial intelligence. J Art Gener Intell 10(2):1–37. https://doi.org/10.2478/jagi-2019-0002
Wenink E (2021) AI Ethics Tool Landscape. https://edwinwenink.github.io/aiethics- tool-landscape/
Wirtz BW, Weyerer JC, Geyer C (2018) Artificial intelligence and the public sector-applications and challenges. Int J Public Admin 42(7):596–615. https://doi.org/10.1080/01900692.2018.1498103
Yu K-H, Beam AL, Kohane IS (2018) Artificial intelligence in healthcare. Nat Biomed Eng 2(10):719–731. https://doi.org/10.1038/s41551-018-0305-z

Authors and Affiliations

Ricardo Ortega-Bolaños[1] · Joshua Bernal-Salcedo[1] · Mariana Germán Ortiz[3,7] · Julian Galeano Sarmiento[2] · Gonzalo A. Ruz[3,5,6] · Reinel Tabares-Soto[1,3,4,7]

✉ Ricardo Ortega-Bolaños
ricardo.ortegab@autonoma.edu.co

Joshua Bernal-Salcedo
joshua.bernals@autonoma.edu.co

Mariana Germán Ortiz
mgerman@alumnos.uai.cl

Julian Galeano Sarmiento
juagaleanosa@unal.edu.co

Gonzalo A. Ruz
gonzalo.ruz@uai.cl

Reinel Tabares-Soto
reinel.tabares@ucaldas.edu.co

1 Electronics and Automation Department, Universidad Autónoma de Manizales, Manizales 170001, Caldas, Colombia

2 Department of Mathematics and Statistics, Universidad Nacional de Colombia Sede Manizales, Manizales 170001, Caldas, Colombia

3 Faculty of Engineering and Sciences, Universidad Adolfo Ibáñez, 7941169 Santiago, Chile

4 Department of Systems and Informatics, Universidad de Caldas, Manizales 170001, Caldas, Colombia

5 Center of Applied Ecology and Sustainability (CAPES), 8331150 Santiago, Chile

6 Data Observatory Foundation, 7510277 Santiago, Chile

7 GobLab School of Government, Universidad Adolfo Ibáñez, Santiago, Chile