ISSN: 1130-3743 - e-ISSN: 2386-5660
DOI: https://doi.org/10.14201/teri.31821

# TOWARDS AN ETHICAL FRAMEWORK FOR ARTIFICIAL INTELLIGENCE IN EDUCATION

## *Hacia un marco ético de la inteligencia artificial en la educación*

Ana María ALONSO-RODRÍGUEZ
*Centre for Research in the Philosophy of Science and Technology (CIFYT). University of A Coruña. Spain.*
*a.alonsor@udc.es*
*https://orcid.org/0000-0002-2379-657X*

Date received: 29/12/2023
Date accepted: 22/02/2024
Online publication date: 04/06/2024

**How to cite this article / Cómo citar este artículo**: Alonso-Rodríguez, A. Mª. (2024). Towards an Ethical Framework for Artificial Intelligence in Education [Hacia un marco ético de la inteligencia artificial en la educación]. *Teoría de la Educación. Revista Interuniversitaria*, *36*(2), 79-98. https://doi.org/10.14201/teri.31821

ABSTRACT

This article reflects on the use of Artificial Intelligence in education from an ethical perspective. It does so from an external perspective, considering its impact on educational contexts as a breeding ground for the ethical and political challenges that society faces. This allows us to measure their scope and depth and propose actions to address them.

The *objectives* of this article focus on the ethical bases of Artificial Intelligence related to educational activity, seeking to identify: a) the opportunities, associated risks and ethical impact on education; and b) the ethical principles that should guide the development, deployment and use of these intelligent systems.

To achieve these objectives, a qualitative study was carried out, supported by a literature review based on the following method: (i) academic works on the current uses and potential risks of Artificial Intelligence and (ii) a comparative analysis of different ethical codes, exploring the convergence of principles applicable to Artificial Intelligence Systems in educational contexts.

The results obtained, in the first instance, place the identified problems in the ethical tradition and question the proliferation of subdomains within the discipline. The possibility of a unified ethical framework that avoids the overlap of principles for each specific domain is then investigated. The findings confirm the usefulness of a widely recognized and influential framework with principles that adapt well to the ethical challenges of education.

It concludes by indicating potential lines of future research: (I) ethical foundation and normative regulation for the development and use of Artificial Intelligence in education in accordance with the selected principles; and (II) definition of a new professional teaching profile and its implications for initial teacher training.

*Keywords:* education; artificial intelligence; ethics, ethical codes; ethical framework.

RESUMEN

Este artículo reflexiona sobre el uso de la Inteligencia Artificial en educación desde una perspectiva ética. Lo hace desde un punto de vista externo, considerando su incidencia en los contextos educativos como vivero de los desafíos éticos y políticos que encara la sociedad. Esto permite dimensionar su alcance y profundidad y proponer medidas para afrontarlos.

Sus *objetivos* se enfocan hacia las bases éticas de la Inteligencia Artificial relacionada con la actividad educativa, buscando identificar: las oportunidades, riesgos asociados y su impacto ético en educación; y b) los principios éticos que puedan guiar el desarrollo, despliegue y uso de estos sistemas inteligentes.

Para ello se realizó un estudio cualitativo, apoyado en una *metodología* de revisión bibliográfica de: (i) trabajos académicos sobre sus usos actuales y riesgos potenciales de la Inteligencia Artificial; y (ii) un análisis comparativo de distintos códigos éticos, explorando la convergencia de principios aplicables a Sistemas de Inteligencia Artificial en los contextos educativos.

Los resultados obtenidos, en primer lugar, sitúan los problemas identificados en la tradición ética, cuestionando la proliferación de subdominios de la disciplina. Se indaga, después, la posibilidad de un marco ético unificado que evite la superposición de principios para cada dominio específico. Se constata la utilidad de un marco ampliamente reconocido e influyente, cuyos principios se adaptan bien a los desafíos de la educación.

Se concluye señalando las líneas en las que se debe avanzar en la investigación: (I) fundamentación ética y regulación normativa para el desarrollo y uso de la Inteligencia Artificial en educación conforme a los principios seleccionados; y (II) definición del nuevo perfil profesional docente y sus implicaciones para la formación inicial del profesorado.

*Palabras clave:* educación; inteligencia artificial; ética; códigos éticos; marco ético.

## 1. Current context

The growing digitalisation of everyday tasks in different sectors of human activity has created informational changes that not only modulate our interaction with information and communication technology (ICT) but also how relationships and social processes are articulated. This drives a reontologization of the world rooted in the infosphere and places us in a new era known as hyperhistory (Floridi, 2014). In this era, information is the fundamental resource, meaning that we are vitally dependent on ICT.

Education is what makes this disruptive innovation possible and as such it has now become "lifelong". In guaranteeing the transition to information as a resource, the 2030 Agenda establishes that one of the Sustainable Development Goals (SDG 4) is "to ensure inclusive and equitable quality education and promote lifelong learning opportunities for all" (United Nations, 2019). Artificial Intelligence (AI) can make an effective contribution to these goals as it already offers numerous opportunities in multiple areas of social activity (industry, communication, health, etc.).

Its incorporation in education, driven by *e-learning*, has been discreet, but its use is constantly growing even in the most traditional educational environments. Its potential to improve education is immense, but it is essential that we learn to manage the numerous social, ethical and deontological implications. Above all, because the normalisation of AI in different areas of social life entails an inevitable change to which education is obligated to respond.

According to this "discourse of imperative change" (Bearman *et al.*, 2023), it is unrealistic to simply dispense with these tools because they entail certain risks. Furthermore, it would be contrary to the ultimate aims of education, which include preparing students to enter the world in which they will have to live – a world that is already shaped by digitalisation and AI. It is essential, therefore, to think about *Artificial Intelligence for Education* in the context of *Education for Artificial Intelligence.*

Generally speaking, the social impact of AI raises legitimate ethical concerns about its use in educational settings. Being an especially high-risk area—and an excellent testing ground for addressing these types of suitably focused problems—, it requires serious ethical reflection. This can be approached from an "internal" perspective, as done by the scientists who create AI, the technologists who implement it and the educators who use it. However, the aim in this study is to focus on an "external" approach, allowing us to consider the scientific/technological/social three-part dimension of the issues and to suggest criteria for assessing the potential solutions, expected results and potential consequences.

As such, the *objectives* of this study focus on the ethical bases of Artificial Intelligence linked to educational activity, seeking to: a) identify the opportunities, risks and ethical impact associated with its incorporation, which will help understand the need for ethical principles; and b) to specify these principles in order to create an ethical framework to regulate the development, deployment and use of intelligent systems. Other key points to consider include its use in making pedagogical decisions and to determine the skills needed by teachers and students when interacting with AI in order to prevent the associated risks.

To outline a proposal, a qualitative study was performed supported by a literature review based on two types of texts: (i) academic works on uses, risks and ethical guidelines in Artificial Intelligence—selected based on the specific topic, authority of the authors and impact of the work—that confirmed the need to link AI programs to ethical codes; and (ii) ethical codes for intelligent systems, identifying the convergence of certain principles. Given the amount and diversity of the published codes, an intentional sampling method was selected with the aim of comparing the perspectives of different stakeholders: international organisations, countries involved in the AI race and industries representative of the sector.

In this context, the research follows various steps that set the structure of this article. First, the problem is placed within the philosophical-methodological framework of design science. Second, the uses, opportunities and risks of AI tools in education are described. Third, serious ethical implications are confirmed. In this respect, after confirming continuity with problems in the tradition, the proliferation of subdomains within the discipline is questioned. Fourth, as a result of the analysis of different ethical codes, the convergence of principles that allow for the proposal of a unified framework for AI in education is confirmed. Fifth, the article concludes that future research is required along two different lines: (i) ethical foundation and normative regulation for the development and use of Artificial Intelligence Systems (AIS) in education; and (ii) outlining a new professional teaching profile to identify teacher training needs.

## 2. Design science: goals, processes and results

Artificial Intelligence in Education (AIED) is a scientific undertaking in the realm of the science of design (González, 2017a). Its aim is to create intelligent tools that allow for broader possibilities for different users in specific contexts of use in education. It arises from the confluence of Artificial Intelligence and Education, which are also disciplines of design science as they seek new goals to broaden human possibilities and are articulated in terms of objectives, processes and results.

In contrast to natural and social sciences, design science expressly focuses on predicting how things might be and on prescribing guidelines that help change how things are now to "how they should be" (Simon, 1996; González, 2007). As such, its main objective is to modulate the future according to a set of predefined aims.

On this scientific basis, interventions can be designed in different fields of human activity and, to put them into practice, they may include technological designs. This is the case with the use of AI to solve problems and improve possibilities in education in terms of achieving objectives. However, this all has serious ethical implications that need to be considered in relation to various aspects:

(i) The objectives, which can be debatable or lead to conflict between different stakeholders. This also includes the selection of problems that AI is used to solve. (ii) The processes: which processes we want to automate, accelerate, etc. and which AI technologies are most suitable for the intended purpose. (iii) The results, which can call into question the intervention itself (that is, the whole sequence of pedagogical activity improved with AI). (iv) The consequences that follow the implemented actions, which may not be as intended.

## 3. Uses and opportunities of AI in education

AI has a broad range of uses in education (Flores Vivar y García Peñalvo, 2023; European Union, 2022; Jara y Ochoa, 2020; Ocaña Fernández *et al.*, 2019; Moreno Padilla, 2019; Sánchez Vila y Lema Penín, 2007), including the following:

(i) To facilitate management tasks, including timetabling, resource allocation, etc.

(ii) To automate daily tasks for teachers, including student tracking, sharing information about students with their families or guardians, marking exercises or tests, etc.

(iii) To support teaching through intelligent tutoring systems that offer students support depending on the difficulties that they may encounter. These systems can provide students with effective support, even outside the classroom, in the preparation of tasks. In contexts and cases where a human teacher is not available, they can be replaced by *virtual facilitators* (realistic virtual characters based on a combination of AI technology, 3D games and computer animation).

(iv) To personalise the learning experience, which is arguably AI's main contribution. The so-called *adaptive teaching systems* are tools related to this objective. They genuinely place the student at the centre of the experience, adjusting the learning path to the individual profile, characteristics and behaviour of each student. By analysing information about each student's progress, they can predict their future performance, allowing for the optimisation of resources and content and even anticipating corrective measures to improve performance.

Various AI-based solutions for special educational needs also contribute to a more personalised form of education; for example, live subtitling in the case of hearing impairment or audio description for blind or visually impaired students.

In addition, there are many other algorithmic applications from the world of informal education that are part of the learning ecosystem: through social media, online gaming platforms or mobile apps, among others.

However, although the use of AI in education is becoming the norm, there is still very little empirical research into its real impact. Some studies show that these tools, with a step-based approach to well-defined problems, are successful in supporting teaching and learning in the field of STEM (Humble & Mozelius, 2022). And it is precisely these disciplines that seek to develop the right type of skills, from a technological and scientific perspective, to face the challenges of the 21st century (Moreno Padilla, 2019). These AI tools also work well in the learning of foreign languages. In any case, it is important to maintain a critical and supervised attitude (European Union, 2022, p. 14).

## 4. Potential risks and ethical implications

Limited teacher training in the use of AI is a key concern in terms of its proper implementation in education as this lack of training could lead to the incorrect use or even abuse of the technology. However, to understand the potential risks, it is important to look, above all, at the design of these tools and ask what is behind their introduction in the education "market", who is driving it and why.

The development of AI in education can interfere with people's autonomy and responsibility and obstruct universal rights (UN General Assembly, 1948) such as *privacy* (art. 12), *equality* (art. 1) and *non-discrimination* (art. 2). This has broad social and ethical implications (Crawford, 2023) that also arise in other areas (Linares Salgado, 2022) but which require special consideration in educational contexts.

*Privacy*: Machine learning systems are trained using large amounts of data. In education, this means: (i) personal information about students and their families; (ii) records of academic performance; and (iii) tracking data generated from online use and learning activities.

Data is key to the success of AI and forms the basis of personalised education. However, the logic of the technology sector, where everything is seen as data that can be taken and used (Crawford, 2023), poses a series of questions in the education sector about how this data is obtained (consent and privacy), how it is analysed (transparency and trust) and the risk of it being used beyond its approved purpose, where students and families end up as victims of consumer manipulation or other practices (Unesco, 2019).

There is a high risk of cyber-attacks when there are no security protocols. This is especially concerning when minors are involved. Education centres are required to have suitable procedures in place to ensure the protection and ethical use of personal data (EU, 2022, p. 11), but this is no easy task when the skills and knowledge to develop AIED systems are in the hands of for-profit organisations and not within the education sector (Humble & Mozelius, 2022).

*Equality and non-discrimination*: The data used to train machine learning algorithms is affected by biases from certain contexts and people. These systems then internalise partial or discriminatory criteria (about "race", gender, age, etc.) from these sources and can even end up amplifying them.

Historically, different forms of inequality conditioned access to resources and opportunities, and this is now affecting the data used to train algorithms to classify and recognise patterns. As such, "under the guise of technical neutrality" (Crawford, 2023, p. 231), where groups of users about whom there is less available data often get left behind, AI systems end up perpetuating social inequality. The use of AI is also expanding the digital gap, socially, as the countries and groups of people with the most resources are those who benefit from the opportunities it provides. All this compromises educational fairness.

AI classification practices are also affected by inherent bias. It is the designers who have the final say in deciding what variables are used in the training data sets and which differences should be considered to correctly *classify* new observations (Crawford, 2023). This contributes to maintaining and amplifying stereotypes (with clear epistemological implications that go beyond the aims of this article).

*Autonomy:* Interactions with AIS can hinder students' development of autonomy and affect their reasoning and decision-making abilities. This adds an additional complication in achieving the goal of creating independent learners (Humble & Mozelius, 2022). In addition, relinquishing freedom of choice and delegating decision-making undermines autonomy defined as the capacity for self-legislation and self-determination, linked to the recognition of human dignity.

*Responsibility*: AIS are agents –and might be replacing human agency–, but they are not moral agents. This creates ethical problems related to responsibility (and accountability) for their actions, including cases of legal liability that would need to be addressed in the event of negative consequences for other people. These problems are exacerbated because machine learning algorithms currently present unpredictable elements due to the opacity with which they function.

### 4.1. *Ethical challenges specific to the field of education*

Together with these general ethical questions, each of the domains in which AI is used has specific ethical challenges that need to be analysed in context. In education, some of the key considerations are: 1) The potential harm that could arise from educational diagnosis and the prediction of student results that could shape their future development. 2) Problems created by the decisions that an AI system might make due to their impact on the educational decisions of teachers, families and other stakeholders (including legislators). 3) The effect on people's development and maturity, especially in the early years of education, due to a shift in roles that changes the relationship between teachers and students.

It is also important to consider: 4) The pedagogical methods that guide the design of AI systems. Generally speaking, they seek solutions that are most readily monetised, and they are not very innovative. They contribute, therefore, to maintaining the *status quo* of the education system (Holmes *et al*, 2022), losing their immense potential to change education and genuinely broaden people's horizons. 5) The image of the world and the tacit social and political ideas to which AI responds that are transferred to the field of education. This technology is not innocuous and "the classifications that are casually chosen to shape a technical system can play a dynamic role in shaping the social and material world" (Crawford, 2023, p. 128). 6) It is also important to mention the ethical dimension of some bad practices engaged in by students when using AI tools, which involve fraud and scams, threaten or damage property and compromise intellectual honesty.

## 5. The importance of ethics

The positive aspects and risks associated with AI are very closely linked, and it is impossible to enjoy the huge benefits that it has to offer without facing its negative consequences. As such, these ambiguities need to be analysed and understood in order to (i) anticipate and calibrate the impacts of certain developments and (ii) select the values and principles that must be protected. This is why the ethics of AI have emerged as a multidisciplinary field of research in applied ethical issues related to the normative problems posed by the development, deployment and use of intelligent systems. These ethics are not for the intelligent systems *per se* but for the people who design, develop or use these systems as they are the ones who could potentially cause, albeit involuntarily, numerous moral problems.

A normative analysis of the impacts of AI reveals parallels with other specific subdomains in the ethics of technology (regarding the internet, data, robots, etc.) formed around different innovations. They often address the same issues (privacy, bias, etc.) with similar approaches, but they are developed in isolation and discussions about them are never connected nor positioned in relation to historical approaches (Sætra & Danaher, 2022). This leads to a duplication of work with no increase in actual knowledge and even to the omission of consolidated ideas in the ethics tradition. As such, the difficulties of addressing the ethical challenges of AI often have more to do with forgetting the fundamental questions of ethics –and acquired historical knowledge– than with our actual understanding of AI.

The ethics of AI have received a lot of attention from researchers (Boddington, 2017; Coeckelberg, 2021; Floridi, 2022) and from different bodies, institutions and corporations. The current focus is on drawing up codes of ethics to identify the values and principles that should be put into practice. Multiple initiatives and proposals have been put forward in recent years and different countries have also tried to establish suggestions and regulations. Often, the proliferation of codes with their guidelines and principles is seen as a problem (Sætra & Danaher, 2022) as it can be confusing (Floridi & Cowls, 2021). However, many of these codes have shared principles with certain emphases and nuances based on different cultural and social values.

A comparative analysis of six high-profile ethical codes for AI –published between 2017 and 2018[1]– allowed Floridi & Cowls (2021) to establish a general framework with five core principles, four of which are also used in Bioethics: *beneficence, non-maleficence, autonomy* and *justice* (Beauchamp & Childress 2019). However, their transfer to the challenges of AI is not perfect and, in some cases, requires a degree of translation:

*Beneficence* refers to promoting wellbeing, preserving dignity and sustaining the planet. *Non-maleficence* refers to privacy, security and caution. *Autonomy* is understood as the power to decide. *Justice* entails promoting prosperity, preserving solidarity and avoiding unfairness. Issues arising from the development of AI itself call for the inclusion of a new principle: *explicability*, which enables other principles through intelligibility and accountability (Floridi & Cowls, 2021).

This framework could serve as the architecture within which regulations, technical standards and best practices for specific sectors can be developed. It can play both an enabling role (e.g., the use of AI to achieve the UN Sustainable Development Goals) and a restrictive role (Floridi & Cowls, 2021, p. 14). It is a general and realistic ethical framework for monitoring and assessing the design, development and use of AIS because it contains a manageable number of principles that are compatible with universal and irrefutable fundamental values. These principles offer a basis for specific regulations that could arise in a broad range of contexts, for international laws and agreements, the monitoring and assessment mechanisms set up by different countries and even citizen observatories, thereby providing moral, legal and political legitimacy.

## 6. A UNIFIED ETHICAL FRAMEWORK FOR ARTIFICIAL INTELLIGENCE IN THE CONTEXT OF EDUCATION

Based on the foregoing ideas, we can now move from the general to the specific. This requires adopting ethical frameworks for the use and development of AI in education. The importance of doing so has been highlighted in previous research (Holmes *et al.*, 2022; Holmes & Porayska-Pomsta, 2023). However, since the *Ethical guidelines for AI in education* published by Aiken & Epstein over two decades ago (Aiken & Epstein 2000), there has been a surprising lack of published work that focuses specifically on ethics (Holmes, *et al.*, 2022, pp. 505-506). Recently, it has once again become a topic of interest in the education research

[1] Asilomar AI Principles (2017); Montreal Declaration for Responsible AI (2017); Version 2 of Ethically Aligned Design (2017); Statement on artificial intelligence, robotics and 'autonomous' systems, from the European Group on Ethics in Science and New Technologies (European Commission, 2018); UK House of Lords Select Committee on Artificial Intelligence Report (2018); Multi-Stakeholder AI Association (2018).

community (Chu *et al*, 2022; Holmes y Porayska-Pomsta, 2023; Boddington, 2023; Nguyen *et al.*, 2023) and, in the past few years, a number of significant initiatives have been launched by international organisations (Unesco, 2019; European Union, 2022).

The European Union, in its Digital Education Action Plan (2021-2027) —which aims to support a sustainable and effective adaptation of the education and training systems of Member States to the digital age— established a series of *ethical guidelines on the use of AI and data in teaching and learning for educators* (European Union, 2022). These guidelines, based on the principles and requirements for trustworthy AI (European Commission, 2019), present four key considerations that should guide educators in their decisions about the use of AI: human agency, fairness, humanity and justified choice.

In the academic community, ethical reports and guidelines by international organisations have been systematically analysed and applicable principles selected to ensure that the AI systems developed for education are essentially ethical by design: Principle of governance and stewardship, principle of transparency and accountability, principle of sustainability and proportionality, principle of privacy, principle of security and safety, principle of inclusiveness and principle of human-centered AIED (Nguyen *et al.* 2023).

Considering these proposals, a unified framework is needed to ensure responsible, ethical and trustworthy AI. With this, the aim is to avoid the overlap of principles and recommendations for each specific domain. There is a prestigious framework that has played a key role in the discussions of the European Commission high-level expert group on AI and influenced various recommendations (Floridi & Cowls, 2021). Its principles are surprisingly adaptable to a broad range of contexts (Linares Salgado, 2022). The question is whether they can also be applied to issues regarding when, how and why an intelligent system is used in education in order to ensure that AI is used properly in specific teaching decisions, actions and practices.

This question led to a new three-phase analysis:

1) Review of guides and recommendations that contain ethical codes deemed relevant due to their impact and published after those analysed by Floridi & Cowls (2021). The aim is to verify the validity and prevalence of the proposed ethical principles. In the sample group, the aim was to ensure that the sector's different influential perspectives were represented: a) recommendations from international organisations: EU (2019), UNESCO (2022) and OECD (2019); b) documents from countries involved in the AI race: USA (Executive Office of the President, 2020), China (China's Ministry of Science and Technology, 2017), EU (EC, 2019) and Spain (Gobierno de España, 2020); and c) initiatives from various key companies in the sector: Microsoft (2022), Google (2022) and Meta (Meta Platforms, 2021).

2) Confirmation of the convergence, in the previous ethical framework, of the following aspects: (i) the principles selected by Nguyen *et al*. (2023) and (ii) the European Commission's ethical guidelines for the use of AI in education.
3) Verification that the values behind these principles are aligned with the values of education (Alonso, 2022 y 2023).

### 6.1. *Results*

The comparative analysis, as seen in Table 1, shows a high degree of overlap among the principles of the nine codes, which clearly converge in the general framework proposed by Floridi & Cowls (2021). Although only *explicability* appears in identical form –or as *transparency*– in all the codes, the "translation" performed to transfer the principles of bioethics to AI allows for their identification.

I) The principle of *non-maleficence* (which refers back to the principle of *non-laedere* in Roman law)[2] appears expressly as *prevention of harm* in the EU guidelines, as *security and safety* in six codes and is centred around the protection of *privacy* in two others. In Spain's *National AI Strategy*, prevention is associated with the principle of data and systems governance: "data cannot be used to harm people or to violate their fundamental rights" (Gobierno de España, 2020, p. 67).

II) *Justice* (*suum cuique tribuere*) is referenced explicitly in one of the codes (China), while *fairness* appears more frequently as a criterion of justice (in seven codes). It can be seen in the principle of *inclusion* (Spain) and in the prevention of threats to justice like *avoid creating or reinforcing unfair bias* (Google) and *non-discrimination*.

III) The principle of autonomy is associated with human-led decision-making, which could be threatened as the agency of machines increases. Interestingly, this principle is not included in the US *Guidance for the regulation of AI applications* and in two of the three private sector incentives (Microsoft and Meta). It appears in the other codes in the form of *autonomy,* as *human oversight, control or direction* or as the *right to choose.*

IV) The same codes that do not include the principle of autonomy also do not include the principle of *beneficence*, which limits the creation of AIS to systems that are of benefit to humankind. Nor does it appear explicitly in the ethical guidelines from the EU. However, *social and environmental wellbeing* does appear among the requirements derived from the principle of *fairness*. In the other codes, it appears in various forms linked to *wellbeing* and *sustainability.*

[2] "Iuris praecepta sunt haec: honeste vivere, alterum non laedere, suum cuique tribuere" (Ulpiano. Digesto 1, 1, 10, 1).

TABLE 1
SYNOPTIC OVERVIEW 1

| | **Beneficence** | **Non-maleficence** | **Autonomy** | **Justice** | **Explicability** |
|---|---|---|---|---|---|
| | Promoting wellbeing, preserving dignity and sustaining the planet | Privacy, security and 'capability caution' | Power to decide (whether to decide) | Promoting prosperity, preserving solidarity and avoiding unfairness | Enabling the other principles through intelligibility and accountability |
| **UNESCO** | Sustainability | Proportionality and do no harm<br>Safety and Security | Human oversight and determination | Fairness and non-discrimination | Transparency and explainability |
| **OECD** | Inclusive growth, sustainable development and wellbeing | Robustness, security and safety | Respecting human values and fairness (including autonomy) | Respecting human values and fairness | Transparency and explicability<br>Accountability |
| **EU** | | Prevention of harm | Respect for human autonomy | Fairness | Explicability |
| **USA** | | Safety and Security | | Fairness and non-discrimination | Disclosure and transparency |
| **CHINA** | Harmony and friendship: promoting the common good | Privacy: respecting and protecting privacy | Privacy: respecting and protecting people's right to know and choose | Fairness and justice | Security: transparency, explainability, traceability, trustworthiness, auditability and safety |

(*continued*)

TABLE 1
SYNOPTIC OVERVIEW 1

| | **Beneficence** | **Non-maleficence** | **Autonomy** | **Justice** | **Explicability** |
|---|---|---|---|---|---|
| **SPAIN** | Social wellbeing<br>Sustainability | Data and systems governance: data cannot be used to harm people or to violate their fundamental rights | Social wellbeing: not reducing, limiting or diverting people's autonomy | Inclusion | Transparency |
| **GOOGLE** | Be socially beneficial | Be built and tested for safety<br>Incorporate privacy design principles | Be accountable to people:<br>the systems must be subject to appropriate human control and direction | Avoid creating or reinforcing unfair bias | Be accountable to people: providing appropriate opportunities for feedback, relevant explanations and appeal |
| **MICROSOFT** | | Privacy and security | | Fairness | Transparency<br>Accountability |
| **META** | | Robustness and safety | | Fairness and inclusion | Transparency and control |

Source: prepared by authors

In conclusion, the framework presented by Floridi & Cowls (2021) offers an ample and complete list of the key ethical principles for AI. –As suggested by the authors–, it could be a useful guide for regulations and practices in specific domains. It should also include the principles proposed in the field of education, focusing on two different areas: (ii) directing the development, deployment and use of AI tools and (ii) guiding educators in their professional practices.

There are clear parallels between the principles of *non-maleficence* and *beneficence* and the meta-principles proposed by Aiken and Epstein as a core philosophical basis for any discussion about AIED systems. According to the *negative meta-principle*, "AIED technology should not diminish the student along any of the fundamental dimensions of human being". The *positive meta-principle* establishes that "AIED technology should augment the student along at least one of the fundamental dimensions of human being" (Aiken & Epstein, 2000, p. 170).

Also, as seen in Table 2, there is a high degree of convergence between the proposal from Nguyen *et al.* (2023) and the unified framework: (i) The principle of *non-maleficence* implies *privacy, safety* and *capability caution***.** Therefore, it also incorporates the *principles of privacy and of security and safety* (ii) *Beneficence* refers to *promoting wellbeing, preserving dignity and sustaining the planet*, which relates to the *principle of sustainability and proportionality*. (iii) *Autonomy* means *the power to decide*. Linked to the recognition of people's dignity, it leads to the basic principle that should govern our relationship with AI: *the irreplaceable human element*, as required by the *human-centered AIED principle*. (iv) *Justice* is *promoting prosperity, preserving solidarity and avoiding unfairness*, which is the key aim of the *principle of inclusiveness*. (v) *Explicability* means enabling the other principles through *intelligibility and accountability*. Intelligibility requires *transparency,* and *accountability* implies the *attribution of responsibilities*, as required by the *principle of transparency and accountability*.

These principles are also aligned with the ethical considerations identified by the European Union to guide educators in their decisions about AI (European Union, 2022, p. 18). The following are recognised in these considerations: (i) the values that form the core aims of education (autonomy, self-determination, responsibility, social cohesion, wellbeing, etc.); (ii) the indispensable values for the development of education (equal opportunities, non-discrimination, transparency, explicability, etc.); and (iii) the values used to inform decision-making in education, including inclusion, fair distribution of rights and responsibilities or meaningful human connection (Alonso, 2022 y 2023).

The ethical principles of *beneficence, non-maleficence, autonomy, justice* and *explicability* applied to AIS can prevent the risks associated with these systems for them to be used to benefit the Right to Education (SDG 4) as an end in themselves and as a means to achieve the other rights. Any tools that —in their design, deployment and use— might violate any of these principles would not be legitimate as they would jeopardise the respect for human dignity that human rights seek to preserve.

TABLE 2
SYNOPTIC OVERVIEW 2

| **Floridi & Cowls, 2021** | | **AIED** | | |
|---|---|---|---|---|
| | | **Nguyen *et al.*, 2023** | **Education values (EU, 2022)** | |
| **Principles** | | **Principles** | **Ethical considerations** | **Values** |
| **Non-maleficence:** | Privacy | Privacy | | |
| | Safety | Security and Safety | | |
| | 'Capacity caution' | | | |
| **Beneficence:** | Promoting wellbeing | | Humanity (consideration for people, identity and dignity) | Wellbeing<br>Safety |
| | Preserving dignity | | | Social cohesion |
| | Sustaining the planet | Sustainability and proportionality | | Meaningful human connection |
| **Autonomy:** | Power to decide (to decide) (leading to the irreplaceable human element) | Human-centered AIED | Human agency | Autonomy<br>Self-determination<br>Responsibility |
| **Justice:** | Promoting prosperity | Principle of inclusion | Fairness | Equal opportunities<br>Inclusion |
| | Preserving solidarity | | | Non-discrimination |
| | Avoiding unfairness | | | Fair distribution of rights and responsibilities |
| **Explicability:** | Intelligibility | Principle of transparency and accountability | Justified choice<br>(use of knowledge, facts and data to justify the choices of the various stakeholders in a school setting) | Transparency |
| | Accountability<br>(Intelligibility requires transparency, and accountability implies the attribution of responsibilities) | | | Explicability |

Source: prepared by authors

*Non-maleficence.* AI should be designed to neither aggravate existing harm nor create new harm. As such, it would not be ethical to design tools that hinder the development of any of the inherent facets of human intelligence in its three core domains: perception —knowledge—, volition and values (González, 2017b), limiting the use of our senses, imagination, thinking and reasoning in a "truly human" way (Nussbaum, 2012, p. 33).

*Beneficence.* In addition to doing no harm, educational AIS should be designed so that students can develop their potential and use all their skills and abilities to choose the life that they want to live in a full and creative manner, in line with human dignity.

*Autonomy.* In ethics, autonomy is understood as the ability for self-legislation and self-determination, which is only applicable to human beings. This requires and justifies human control over "machines". As such: (a) decisions that affect people's lives cannot be delegated to intelligent machines without human supervision. As a result, educational decisions cannot be solely based on automatic data processing. (b) A human being should always know if they are interacting with human or with a machine. Therefore, when AI tools are used, the special precautions required for vulnerable people, such as children, require informed consent from the parents or legal guardians.

Education is an especially high-risk area due to its role in the development of human beings and in shaping how we think and act throughout our lives. It is essential to stress caution, which now appears as a requirement of non-maleficence, and to incorporate the *Precautionary Principle* in any educational intervention where AI might present a threat to autonomy and human dignity (expanding on the Wingspread Statement).

## 7. Conclusion and discussion

The external perspective on AI and its use in education focuses on social repercussions and responsibility in practices that involved these systems (Bearman *et al.*, 2023) and highlights the importance of ethics in the contemporary debate about education (Holmes *et al.*, 2022). In this article, a unified framework has been sought for this current and future problem, where science, technology and society have all been taken into account.

This framework is a useful starting point as (i) it can guide and inform coherent and carefully planned regulation, (ii) it allows for the identification of good practices, (iii) it identifies the skills that educators need to develop and (iv) it offers students a set of criteria for interacting with AI and using it to its full potential in learning environments. But there is still a long way to go. There has been a notable increase in research in this field and, in future projects, it would be interesting to compare the parallels detected in this study with the results of a meta-analysis of a much larger group of documents. In turn:

1) The principles represent the values that should be achieved through the end goals (e.g., an end goal that, if achieved, could cause harm, would not be legitimate). However, in addition to ensuring ethical goals for AI, we must also determine how they can be achieved in an ethical way because it is one thing to do ethical things and another to do them ethically (Holmes *et al.*, 2022, p. 504). As such, even if the end goal is ultimately beneficial, it does not justify the use of all means to achieve it (think about the misuse of data).
2) The values to which the principles refer are "trusted" values, representing some of our core certainties such as the guarantee of human dignity and democratic coexistence. However, (a) they can be conceptualised in different ways, as seen in the different criteria for *justice* used in the different codes and (b) the values are rooted in the Western tradition and, to ensure a true global reach, perspectives from other regions and cultures must be included. This brings us back to the age-old debate of multiculturalism. As argued by Adela Cortina, "there can be no response to universal challenges other than to adopt a universal ethical approach, where the end goal of decision-making is the universal good" (Cortina, 1997, p. 261).
3) The effective mechanisms to ensure that all stakeholders adopt these principles and accept responsibility for their transgression are often not specified. In particular, this refers to the technology companies that develop AI products for education and that also control the current research (Ahmed *et al.*, 2023). As social subjects, they can and must take responsibility for their actions insofar as: a) they demonstrate intentionality in the deployed activities and b) decision-making is controlled by the organisation itself (González, 2020).

The concerns raised by AI are connected to traditional ethical issues. They are, in general, not new. The degree of significance is different as, to a large extent, AI shapes how we interact with the world, modulates social relationships and, even, affects the configuration of our own identity. The consequences of this affect how we view traditional ideas of morality and individual responsibility, which perhaps need to be reviewed in light of the new conditions.

The debate should develop in two main directions: Firstly, the ethical foundations that give the principles their moral legitimacy. The oversimplification of especially complex issues should be avoided, as can happen when they are addressed by AI experts. To decide what cannot be sacrificed or what must be preferred, we must be able to offer a solid justification of what is considered "good" and what "should be done". And this is a job for ethics.

Ensuring that AIS is developed and used in education in accordance with ethical principles will also require legal and political regulation. However, such laws need to connect with the corresponding values and be in tune with people's reasons and desires. Because having a law (that can be imposed or assumed in the process of socialisation) is one thing and a person's reasons for making it their own is another (Cortina, 1997). As such, an education in values should be considered.

Secondly, the definition of a new professional teaching profile, which requires specifying the expert knowledge that is now needed to work in this profession. This should be included in initial teacher training plans. Based on the above, this should include, at least: (a) scientifically robust research training that combines scientific rigour with practical problem-solving; (b) ethical training to address the challenges of AI; and (c) training in AI that allows educators to adjust their approach and design more effective teaching strategies in order to progress towards the goal of "inclusive, fair, quality education and to promote opportunities for lifelong learning for all".

## References

Ahmed, N., Wahed, M., & Thompson, N. C. (2023). The growing influence of industry in AI research. *Science*, *379* (6635), 884-886. https://www.science.org/doi/10.1126/science.ade2420

Aiken, R. M., & Epstein, R. G. (2000). Ethical Guidelines for AI in Education: Starting a Conversation. *International Journal of Artificial Intelligence in Education*, *11*, 163-176.

Alonso, A. M. (2022). Las Ciencias de Diseño como Ciencias Aplicadas y el papel de los valores: Análisis del caso de la Educación. *Revista de Investigación Filosófica*, *9*(1), 3-27. https://doi.org/10.26754/ojs_arif/arif.202216381

Alonso, A. M. (2023). *Filosofía de la Ciencia de la Educación: Análisis como Ciencia Aplicada de Diseño*. Tirant lo Blanch.

Asamblea General de la ONU. (1948). Declaración Universal de los Derechos Humanos. Paris. http://www.un.org/en/universal-declaration-human-rights/

Bearman, M., Ryan, J., & Ajjawi, R. (2023). Discourses of artificial intelligence in higher education: a critical literature review. *Higher Education, 86*, 369–385. https://doi.org/10.1007/s10734-022-00937-2

Beauchamp, T. L., & Childress, J. F. (2019). *Principles of biomedical ethics*. Oxford University Press.

Boddington, P. (2017). *Towards a code of ethics for artificial intelligence*. Springer.

Boddington, P. (2023). Introduction, Why AI Ethics? In P. Boddington, *AI Ethics. A Textbook* (pp. 1-34). Springer.

Chu, J., Xi, L., Zhang, Q., & Lin, R. (2022). Research on Ethical Issues of Artificial Intelligence in Education. In J. Yang, *et al.* (Eds.), *Resilience and Future of Smart Learning* (pp. 101-108). Springer. https://doi.org/10.1007/978-981-19-5967-7_12

Coeckelberg, M. (2021). *Ética de la Inteligencia Artificial*. Cátedra.

Comisión Europea. (2019). Grupo de Expertos de Alto Nivel sobre IA. *Directrices Éticas para una IA fiable*. https://op.europa.eu/es/publication-detail/-/publication/d3988569-0434-11ea-8c1f-01aa75ed71a1

Cortina, A. (1997). *Ciudadanos del mundo. Hacia una teoría de la ciudadanía*. Alianza editorial.

Crawford, K. (2023). *Atlas de I. Poder, Política y Costes Planetarios de la Inteligencia Artificial*. Ned ediciones.

Executive Office of the President. (2020). *Guidance for regulation of Artificial Intelligence applications*. https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf

Flores Vivar, J. M. y García-Peñalvo, F. J. (2023). Reflexiones sobre la ética, potencialidades y retos de la Inteligencia Artificial en el marco de la Educación de calidad (ODS4). *Comunicar, 31*(74), 37-47. https://doi.org/10.3916/C74-2023-03

Floridi, L. (2014). *The Fourth Revolution - How the Infosphere is Reshaping Human Reality.* Oxford University Press.

Floridi, L (2022). *Etica dell'intelligenza artificiale. Sviluppi, opportunità, sfide*. Rafaello Cortina.

Floridi L., & Cowls, J. (2021). A Unified Framework of Five Principles for AI in Society. In L. Floridi (Ed.), *Ethics, Governance, and Policies in Artificial Intelligence* (5-17). Springer.

Gobierno de España. (2020). *Estrategia Nacional de Inteligencia Artificial*. https://www.lamoncloa.gob.es/presidente/actividades/Documents/2020/021220-ENIA.pdf

González, W. J. (2007). Configuración de las Ciencias de Diseño como Ciencias de lo Artificial: Papel de la Inteligencia Artificial y de la racionalidad limitada. En W. J. González (Ed.), *Las Ciencias de Diseño: Racionalidad limitada, predicción y prescripción* (41-69). Netbiblo.

González, W. J. (2017a). Artificial Intelligence in a New Context: "Internal" and "External" Factors. *Minds & Machines*, *27*(3), 393–396. https://doi.org/10.1007/s11023-017-9444-3

González, W. J. (2017b). From Intelligence to Rationality of Minds and Machines in Contemporary Society: The Sciences of Design and the Role of Information. *Minds & Machines*, *27*(3), 397-424. https://doi.org/10.1007/s11023-017-9439-0

González, W. J. (2020). The Internet at the Service of Society: Business Ethics, Rationality, and Responsibility. *Éndoxa*, *46*, 383-412. https://doi.org/10.5944/endoxa.46.2020.28029

Google LLC. (2022). *Artificial Intelligence at Google: Our Principles*. https://ai.google/principles/

Holmes, W., & Porayska-Pomsta, K. (2023). *The Ethics of Artificial Intelligence in Education.* Routledge Taylor.

Holmes, W., Porayska-Pomsta, K., Holstein, K. et al. (2022). Ethics of AI in Education: Towards a Community-Wide Framework. *International Journal of Artificial Intelligence in Education*, *32*, 504–526. https://doi.org/10.1007/s40593-021-00239-1

Humble, N., & Mozellius, P. (2022). The threat, hype, and promise of artificial intelligence in education. *Discover Artificial Intelligence*, *2*(22). https://doi.org/10.1007/s44163-022-00039-z

Jara, I., & Ochoa, J. M. (2020). *Usos y efectos de la inteligencia artificial en educación*. Banco Interamericano de Desarrollo. http://dx.doi.org/10.18235/0002380

Linares Salgado, J. E. (2022). Principios éticos para el desarrollo de la Inteligencia Artificial y su aplicación en los sistemas de salud. *ArtefaCToS. Revista de Estudios de la Ciencia y la Tecnología*, *11*(2), 137-161.

Meta Platforms. (2021). *Facebook five pillars of responsible AI*. https://ai.facebook.com/blog/facebooks-five-pillars-of-responsible-ai/

Microsoft Corporation (2022). *Responsible IA*. https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimar yr6

Ministerio de Ciencia y Tecnología de la República Popular China. (2017). *Principios de Gobernanza para una Nueva Generación de IA: Desarrollo Responsable de la Inteligencia Artificial*. https://www.newamerica.org/cybersecurity-initiative/digichina/blog/translation-chinese-expert-group-offers-governance-principles-responsible-ai/

Moreno Padilla, R. (2019). La llegada de la inteligencia artificial a la educación. *RITI Journal*, *7*(14), 260-270. https://doi.org/10.36825/RITI.07.14.022

Nguyen A, Ngo H. N, Hong, Y., Dang, B., & Nguyen, B.T. (2023). Ethical principles for artificial intelligence in education. *Educ Inf Technol*, *28*(4), 4221-4241. https://doi.org/10.1007/s10639-022-11316-w

Nussbaum, M. (2012). *Crear capacidades. Propuesta para el desarrollo humano.* Paidós.

Ocaña-Fernandez, Y., Valenzuela-Fernandez, L., & Garro-Aburto, L. (2019). Inteligencia artificial y sus implicaciones en la educación superior. *Propósitos y Representaciones, 7*(2), 536-552. http://dx.doi.org/10.20511/pyr2019.v7n2.274

OECD (2019). *Recommendation of the Council on Artificial Intelligence*. OECD/LEGAL/0449. https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

Sætra, H.S., & Danaher, J. (2022). To Each Technology Its Own Ethics: The Problem of Ethical Proliferation. *Philosophy & Technology, 35*, 93. https://doi.org/10.1007/s13347-022-00591-7

Sánchez Vila, E. M., & Lama Penín, M. (2007). Monografía: Técnicas de la Inteligencia Artificial Aplicadas a la Educación. *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial, 11*(33), 7-12.

Simon, H. A. (1996). *The Sciences of the Artificial*. The MIT Press. 3ª ed.

Ulpiano, E. D. (1836). Digesto. En Academicun parisiense, *Corpus Iuris civilis*. https://gallica.bnf.fr/ark:/12148/bpt6k65768518/f13.item.r=digestum

Unesco. (2019). *Beijing consensus on artificial intelligence and education*. https://unesdoc.unesco.org/ask:/48223/pf0000368303

Unesco. (2022). *Recomendación sobre la ética de la inteligencia artificial*. https://unesdoc.unesco.org/ark:/48223/pf0000379920_spa.page=15.

Unión Europea. (2022). *Directrices éticas sobre el uso de la inteligencia artificial (IA) y los datos en la educación y formación para los educadores*. Publications Office of the European Union. https://data.europa.eu/doi/10.2766/898

United Nations. (2019). *The Sustainable Development Goals Report 2019*. https://bit.ly/34nbq60

Wingspread Statement on the Precautionary Principle (1998). http://www.sehn.org/sehn/the-precautionary-principle-march-1998