Computer Law & Security Review 53 (2024) 105986

Contents lists available at ScienceDirect

Computer Law & Security Review: The International Journal of Technology Law and Practice

journal homepage: www.elsevier.com/locate/clsr

The ALTAI checklist as a tool to assess ethical and legal implications for a trustworthy AI development in education

Andrea Fedele [a,d,1], Clara Punzi [a,b,d,*,1], Stefano Tramacere [a,c,1]

[a] Department of Computer Science, University of Pisa, Largo Bruno Pontecorvo, 3, Pisa, 56127, Italy
[b] Scuola Normale Superiore, P.za dei Cavalieri, 7, Pisa, 56126, Italy
[c] LIDER-Lab, DIRPOLIS Institute, Scuola Superiore Sant'Anna, Via Santa Cecilia, 3, Pisa, 56127, Italy
[d] KDD Lab, ISTI-CNR, Via G. Moruzzi 1, Pisa, 56124, Italy

ARTICLE INFO

Keywords:
Trustworthy AI
Education
Vulnerability
AI regulation
AI accountability
eXplainable AI

ABSTRACT

The rapid proliferation of Artificial Intelligence (AI) applications in various domains of our lives has prompted a need for a shift towards a human-centered and trustworthy approach to AI. In this study we employ the Assessment List for Trustworthy Artificial Intelligence (ALTAI) checklist to evaluate the trustworthiness of *Artificial Intelligence for Student Performance Prediction* (AI4SPP), an AI-powered system designed to detect students at risk of school failure. We strongly support the ethical and legal development of AI and propose an implementation design where the user can choose to have access to each level of a three-tier outcome bundle: the AI prediction alone, the prediction along with its confidence level, and, lastly, local explanations for each grade prediction together with the previous two information. AI4SPP aims to raise awareness among educators and students regarding the factors contributing to low school performance, thereby facilitating the implementation of interventions not only to help students, but also to address biases within the school community. However, we also emphasize the ethical and legal concerns that could arise from a misuse of the AI4SPP tool. First of all, the collection and analysis of data, which is essential for the development of AI models, may lead to breaches of privacy, thus causing particularly adverse consequences in the case of vulnerable individuals. Furthermore, the system's predictions may be influenced by unacceptable discrimination based on gender, ethnicity, or socio-economic background, leading to unfair actions. The ALTAI checklist serves as a valuable self-assessment tool during the design phase of AI systems, by means of which commonly overlooked weaknesses can be highlighted and addressed. In addition, the same checklist plays a crucial role throughout the AI system life cycle. Continuous monitoring of sensitive features within the dataset, alongside survey assessments to gauge users' responses to the systems, is essential for gathering insights and intervening accordingly. We argue that adopting a critical approach to AI development is essential for societal progress, believing that it can evolve and accelerate over time without impeding openness to new technologies. By aligning with ethical principles and legal requirements, AI systems can make significant contributions to education while mitigating potential risks and ensuring a fair and inclusive learning environment.

* Corresponding author at: Scuola Normale Superiore, P.za dei Cavalieri, 7, Pisa, 56126, Italy.
E-mail addresses: andrea.fedele@phd.unipi.it (A. Fedele), clara.punzi@sns.it (C. Punzi), stefano.tramacere@santannapisa.it (S. Tramacere).

[1] Authors contributed equally.

[2] This article analyzes the AI Act proposed by the European Commission on 21 April 2021. We would like to inform readers that during the drafting of this paper, the Proposal was the subject of a lengthy negotiation by the EU co-legislators (Parliament and EU Council), which ended with the last trilogue in December 2023. Indeed, a political agreement was reached there by the EU co-legislators and the Commission, leading to a final approval of the text in their respective bodies *in plenario* in early 2024. Therefore, at the time of writing the final text is not yet available. The authors would like to stress that the content of the final legislation might slightly differ from what is described here. For information purposes, the reader is informed of (1) the Common Position (the so-called General Approach) of the Council of the EU, finalized on 25 November 2022 available here https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf (last accessed on 29 December 2023); this was followed by (2) the Parliament's amendments on 14 June 2023, which significantly modified the Commission's proposal, available at https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf. (last accessed on 30 December 2023). Finally, for more information regarding (3) the final trilogue on 6–9 December 2023, we suggest to read https://www.consilium.europa.eu/en/press/press-releases/2023/12/09/artificial-intelligence-act-council-and-parliament-strike-a-deal-on-the-first-worldwide-rules-for-ai/.

https://doi.org/10.1016/j.clsr.2024.105986

Available online 18 May 2024

1. Introduction

In the last few years, we have witnessed a rapid spread of AI applications to numerous spheres of our daily and professional lives: from auto-generated text and social media news feeds, to virtual homes and mobile phone voice assistants. AI offers automated translation, assists shoppers buying online and recommends the fastest route on the drive home. It is also a key component of rapidly developing technologies such as facial recognition and self-driving vehicles. The new AI Act proposal [2] [1] presented by the European Commission in 2021 to lay down harmonized rules on AI systems, defines in Article 3 an AI system as a *"machine-based system that is designed to operate with varying levels of autonomy and that can, for explicit or implicit objectives, generate output such as predictions, recommendations, or decisions, that influence physical or virtual environments"*. Due to its speed and its power of self-learning, AI has the capacity to transform our societies [2]. For instance, AI is accelerating the battle against medical diseases [3] and mitigating the impact of disability [4]; it is helping to tackle climate change [5] and optimize efficiency in agriculture [6]; it can assist distribution of humanitarian aid [7]; it has enormous potential for improving access to, and quality of, education globally [8–13]; and it can transform public and private transport [14]. Moreover, the recent widespread availability of free and low-cost generative AI tools (so-called "general-purpose AI" models in the AI Act) facilitates the spread of high volumes of text, image, voice, and video content, simplifying many repetitive tasks [15]. At the same time, however, many AI developers lack legal skills or are unaware of potential legal problems, and often have far greater resources at their disposal than the authorities that should control and regulate them [16]. Additionally, public authorities often lack the appropriate knowledge and expertise to adequately regulate their use in different areas. These asymmetries cause many problems, pushing governments to prioritize innovation (however destructive in its effects) at the cost of fundamental sacrifices of social values [17]. Therefore, in such a scenario, despite the enormous potential, there are good reasons to be sceptical of AI. For instance, the Tesla crashes have delayed the dream of self-driving cars [18]. Even in areas where AI systems seems to be an unqualified good (as in machine learning to spot melanoma better), researchers worry that current data sets do not adequately represent all patients' racial backgrounds [19]. While machines are proving "better than humans" at some narrow tests, that superiority is fragile, given the dependence of many forms of AI on data sets that change over time [16,20,21]. Furthermore, in the context of generative AI, much of the content created is likely to be favorable or could be useful to a specific audience, but the risk that these systems also facilitate the dissemination of extremely harmful content is very high. Such new generative AI systems could be used to disseminate false, misleading, biased, or dangerous content. In this regard, it must be emphasized that as generative AI tools become more sophisticated, it will be faster, cheaper and easier to produce such content, and existing harmful content may serve as a basis for producing more [15]. Therefore, in light of the pros and cons of this disruptive new technology, the call for "human-centred" and "trustworthy" AI is becoming increasingly crucial [22]. Against the global scenario 2.1, the European Union (EU) is on the way to establish a comprehensive and effective regulatory framework for trustworthy AI.

In this discussion, *trustworthy AI* refers to the property of an AI-based system of being able to maximize its benefits while at the same time preventing and minimizing its risks. This concept was first defined in the *Ethic Guidelines for Trustworthy AI* [22] by the *High Level Expert Group on AI* (hereinafter AI HLEG), which is an independent group set up by the European Commission with the mandate of drafting guidelines for ethical development of AI systems, as well as policy and investment recommendations. According to the AI HLEG, trustworthy AI systems should be:

1. *lawful*, that is, complying with all applicable laws and regulations;
2. *ethical*, that is, ensuring adherence to ethical principles and values;
3. *robust*, both from a technical and social perspective, as they have the potential to inadvertently inflict harm despite good intentions [23].

The concept of *trustworthy AI* is also established in the new AI Act proposal, which, in Recital 5, states that the Regulation supports the EU objective of being a "*global leader in the development of secure, **trustworthy** and ethical artificial intelligence*". Moreover, the AI Act is part of the EU coordinated AI strategy plan, launched by the Commission in 2018. The plan vision of an ethical, secure and cutting-edge AI made in EU has been expressed through the release of policy documents such as: *Communication on Artificial Intelligence for Europe* [24], *Building Trust in Human-Centred Artificial Intelligence* [25] and the *White Paper on Artificial Intelligence - A European Approach to Excellence and Trust* [26]. Concurrently, in 2019 the AI HLEG presented the final version of the so-called "Ethics Guideline for Trustworthy AI" (hereinafter, *AI Ethics Guidelines*) [22], a document addressed to all AI stakeholders and aimed to *"provide guidance for AI applications in general, building a horizontal foundation to achieve Trustworthy AI"*. Successively, the AI HLEG also issued the Assessment List for Trustworthy AI (ALTAI) [27], an accessible and dynamic checklist that should help AI developers and deployers to put the principles outlined in the AI Ethics Guidelines into practice. Although these requirements are not binding, they formed the foundation of the proposed AI Act presented to the European co-legislators in April 2021 [1].

This study specifically examines the ethical and legal consequences of using AI in the field of education, with a particular emphasis on the potential effects on children, who are considered a vulnerable group. The employment of AI-augmented systems in classrooms across Europe and the majority of countries globally is rapidly growing, especially after the technological acceleration prompted by the COVID-19 pandemic. All of teaching, learning, assessment and school administration can potentially take advantages from the use of AI. In general, three fundamental benefits could be expected: an increased capacity of education systems and productivity of educators, the enhancement of teaching and learning in support to the learners' well-rounded development, and the delivery of autonomous learning recommendations [28]. On the other side, the harms that AI can cause to learners are not negligible as well [16]. In fact, the very recent AI Act aims at prohibiting those AI systems which pose the unacceptable level of risk of threatening the safety, livelihoods and rights of people. The usage of AI in the educational systems can trigger different drawbacks, varying from psychological harm of both students and teachers, to concerns related to privacy, surveillance, autonomy, bias, and discrimination [29]. Therefore, a higher attention must be employed within the development and usage of AI system within the educational context of application, since the psychological manipulation might be subliminal and not necessarily beneficial.

In this work, we focus on a case study in the domain of education to assess the practicality of ethical-legal frameworks for AI in realistic scenarios. Specifically, we illustrate how the ALTAI checklist can be utilized as a methodological approach to conduct a thorough evaluation of the trustworthiness of *Artificial Intelligence for Student Performance Prediction* (AI4SPP), a simple AI tool we developed as a Proof-of-Concept for an educational support system aimed at early detection of students who are at risk of school failure. From our prospective, the use of such a tool may help educators detecting patterns and spotting indicators of risk in the school performance of their students through statistical analysis. This, in turn, could facilitate the implementation of appropriate supportive interventions [10,28,30]. Our AI-system constitutes an aid tool for teachers, as an end users, in the evaluation of student performance, with the ultimate goal of enhancing the educational sector by tailoring students' training (e.g. through proactive intervention in cases of students facing critical situations at risk of

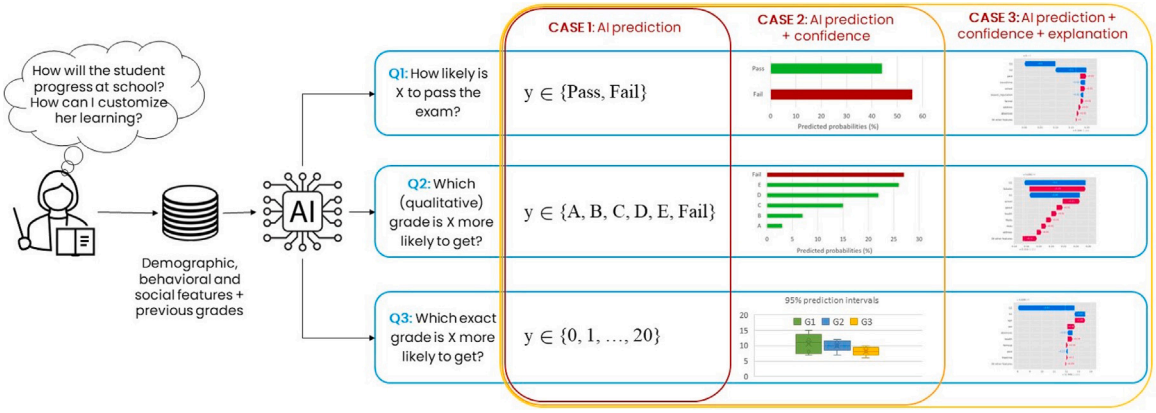


Fig. 1. Implementation design of AI4SPP. The blue boxes wrap the workflow pertaining to the three different predictive tasks (i.e., binary classification, multiclass classification and regression). The red, orange and yellow nested boxes correspond to the three possible groups of results that the system may generate in response to a request made by the end-user.. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

school failure, which would also result in a strengthened teacher–student relationship). Furthermore, AI4SPP seek to boost teachers' and students' awareness of the factors that determine particular low performance. Last but not least, we believe that our methodological approach can result in an AI system that stimulates productive reflection among educators regarding the actual biases present within the school community, thereby promoting the implementation of targeted interventions to alleviate them.

Fig. 1 presents a visual representation of the system design of AI4SPP. Nevertheless, despite these numerous advantages, in this study we also highlight that an inappropriate use of AI4SPP could generate a range of ethical and legal concerns. For instance, issues related to privacy violation may arise given the collection and analysis of data of particularly vulnerable subjects, which is necessary to train the AI model. Additionally, AI4SPP can foster the actualization of unfair actions whenever its predictions were based on unacceptable discrimination factors related to gender, ethnicity or socio-economic background. [28].

The rest of this work is structured as follows: in Section 2 we offer a comparative analysis regarding AI regulation in EU, US and China, along with a study of the AI policy framework in the context of education in EU and at international level (UNESCO). In Section 3, we examine the AI Act's regulation regarding vulnerability in prohibited and high-risk AI systems; the risk of discrimination, and the concept of vulnerability in the data protection framework. Additionally, we discuss the General Data Protection Regulation (GDPR) [31] on automated decision-making involving children. Subsequently, in Section 4 we introduce the Ethics Guidelines for Trustworthy AI and the ALTAI checklist; we describe the seven requirements for Trustworthy AI and a collection of related questions designed to assist AI designers, developers and deployer in following a multidisciplinary evaluation process. Section 5 describes our AI-based tool AI4SPP, the dataset, the implementation details, the numerical results and some qualitative examples of outcome explanations. Afterwards, in Section 6 we present an in depth assessment of the ALTAI checklist based on our case study. Finally we conclude in Section 7 suggesting possible solutions to cover gaps emerging from the ALTAI assessment process, with specific emphasis on its future alignment with the AI Act upon its adoption.

2. International and european context

In this section, we first offer a comparative analysis regarding AI regulation in EU, US and China (Section 2.1). Successively, in Section 2.2 we present a detailed analysis of the AI policy framework in the field of education, focusing on the European Union and the international level (UNESCO).

2.1. Global perspective on AI regulation

From a comparative perspective, EU takes the lead in regulating AI in comparison to other countries. Firstly, the General Data Protection Regulation (GDPR) [31] already establishes specific rules for automated decision-making and profiling. Secondly, the AI Act provides an organic discipline for all industries impacted by AI, employing a risk-based approach. In contrast to the EU, the United States has no broad, federal AI-related laws, nor significant data-protection rules [32]. Nevertheless, in October 2022, the White House Office of Science and Technology Policy ("OSTP") did release a Blueprint for an AI Bill of Rights [33]: a White Paper describing five principles meant to guide the use of AI, as well as potential regulations. The document states that automated systems should be safe and effective, non-discriminatory, protective of people's privacy and transparent. In October 2022, one law did make it through Congress [34]: it requires that officials at federal agencies who procure AI products and services to be trained on how AI works. In January 2023, the National Institute of Standards and Technology (NIST) released the AI Risk Management Framework (AI RMF 1.0) [35] designed to equip organizations and individuals with approaches that increase the trustworthiness of AI systems, and to help foster the responsible design, development, deployment, and use of AI systems over time. In addition, on February 2023, President Biden signed an Executive Order that briefly mentions a requirement to *"prevent and remedy [. . .] algorithmic discrimination"* [36], which only applies to federal agencies. Lastly, on October 30 2023, the President also signed an Executive Order on Safe, Secure and Trustworthy AI *"to establish new standards for AI safety and security, protects Americans' privacy, advances equity and civil rights [...]"* [37]. At the same time, federal legislation has been put forward. Lawmakers have previously considered a bill aimed at algorithmic accountability [38] that would ask firms using automation to present impact assessments to the Federal Trade Commission (FTC), for instance. Unfortunately, the Algorithmic Accountability Act (hereinafter *AAA*) failed to make it out of the 117th Congresses. In general term, the EU and U.S. strategies share a conceptual alignment on a risk-based approach, agree on key principles of trustworthy AI, and endorse an important role for international standards, but regrettably the current US effort to regulate AI is too modest [39]. After all, not only consumer but citizens in general are increasingly affected by automated decision-making systems. Moreover, a policy is only as strong as the institutions that support it. While the EU AI Act is part of a holistic and long-term plan to shape the digital ecosystem in the EU and beyond, the US AAA constituted only a fragmented attempt.

China, another important global competitor, combines national, provincial and local regulations (e.g., those covering autonomous vehicles) with an emphasis on maintaining state power, cultural values

and technological innovation.[3] Despite this (apparent) fragmentation, China has enacted the largest number of AI laws that apply to AI systems used by companies and not by the government. Thus, the three most concrete regulations on algorithms and AI include: the 2021 regulation on recommendation algorithms,[4] the 2022 rules for deep synthesis[5] (synthetically generated content) and the 2023 draft rules on generative AI.[6] These interconnected documents contain the most targeted and impactful laws to date, creating concrete requirements for how algorithms and AI are built and deployed in China [40]. In particular, the first regulation (*Provisions on the Management of Algorithmic Recommendations in Internet Information Services*) was motivated by government fears about algorithms controlling how news and content are disseminated online. The Act includes many provisions for content control, as well as protections for workers impacted by algorithms, among others. It also created the "algorithm registry" used in future regulations. While the provisions on the *Administration of Deep Synthesis Internet Information Services* targets many AI applications used to generate text, video, and audio. It prohibits the generation of "fake news" and requires synthetically generated content to be labeled. The core motivation for the regulation is concern over deep fakes. Finally, the draft on *Measures for the Management of Generative Artificial Intelligence Services* was established in response to the explosion in popularity of AI chat-bots like ChatGPT, thus the regulation covers almost the exact same ground as the deep synthesis regulation, but with more emphasis on text generation and training data. In particular, it requires providers to ensure that both the training data and generated content be *"true and accurate"*. These laws show how Beijing is struggling to reconcile its ambition to develop cutting-edge technologies with its long-standing censorship regime [32].

Against this global scenario, in Europe, on the way to the establishment of a comprehensive, effective and trustworthy regulatory framework, the European Commission has taken further actions to address context-specific risks that derive from the use of AI-based tools in high-stake scenarios, such as those involving vulnerable groups, considering specifically the risks that AI systems may pose to the health, safety and fundamental rights of individuals [1].

2.2. The global policy frameworks for AI in education

In the EU, the awareness of the unprecedented use of technology for education and training purposes brought by the COVID-19 pandemic led to the establishment of the EU Digital Education Action Plan (2021–2027) [41], a policy initiative aimed at supporting the sustainable and effective adaptation of the education and training systems of Member States to the digital age. One of the first outcomes of such initiative has been the formulation of the AI Ethical Guidelines in Education [42], a document based on the assumption that AI-augmented educational systems can potentially enhance teaching, learning and assessment, provide better learning outcomes and improve schools administration. However, the AI Ethical Guidelines in Education also highlight that, if not carefully designed or used, the same tools run the risk of having harmful consequences on children. Additionally, it is crucial to keep a critical and supervised attitude considering that there is still limited evidence on how AI is affecting education as well as its impact on students, teachers and wider society [10,43,44]. Therefore, it is of utmost importance to empower the educators with the ability of understanding not only if the AI systems they are using are reliable, fair, safe and trustworthy, but also if the management of educational data is secure and protects the privacy of individuals. With this purpose in mind, the AI Ethical Guidelines in Education have been specifically designed to provide awareness and practical guidance to all educators who are increasingly confronted with the use of AI in their teaching practice. Notably, as soon as the proposed EU AI Act becomes legally binding, education institutions can expect to directly rely on the trustworthiness of high-risk AI systems based on the accompanying certification guaranteed by the provider. In such a case, education authorities should therefore be responsible for verifying only that AI systems are actually compliant, and will then be able to focus on the ethical concerns related to teaching, learning and assessment, while always adhering to the applicable data protection regulations [42].

Beyond EU, a parallel discourse about the challenges and opportunities posed by AI in the field of education has been carried out by the United Nations (UN), in particular by the United Nations Educational, Scientific and Cultural Organization (UNESCO) and by the United Nations Children's Fund (UNICEF). Broadly speaking, UN recognizes the potential of AI to accelerate the process towards the achievement of the Sustainable Development Goal 4 (SDG 4) of the 2030 Agenda, which specifically aims to *"ensure inclusive and equitable quality education and promote lifelong learning opportunities for all"* [45]. This will be in line with the principles stated in the UN Convention on Children's Rights (hereinafter CRC) that are relevant to analyze children's rights in the digital environment which enhance their freedoms to shape and express their opinion both as individuals and as a groups, and at the same time ensuring protection from potential harms [46]. In these regards, it is important to consider the principle of the best interests of the child which is one of the four overarching guiding principles on children's rights (right to non-discrimination, best interests, the right to life, survival and development, and the right to participation or right to express views). It is anchored in Article 3(1) of the Convention on the Rights of the Child (CRC) and in Article 24(2) of the Charter of Fundamental Rights of the European Union. Both instruments give children the right to have their best interests assessed and taken into account as a primary consideration in all actions or decisions that concern or affect children, whether undertaken by public or private social welfare institutions, courts of law, administrative authorities or legislative bodies.

However, AI in education also poses risks for child users, including those related to their privacy, safety and security. Indeed, given that AI systems can work unnoticed and on a large scale, there is a serious risk of widespread exclusion and discrimination [44]. A series of workshops conducted by UNICEF in 2020 with 245 adolescents in five countries highlighted that, in spite of their enthusiasm and expectations towards AI, most adolescents are also worried that too much data is being collected by AI systems with the result of infringements on data privacy, as well as personal data leaks, hacking and misuse [47].

In 2019 UNESCO organized the Beijing International Conference on Artificial intelligence and Education, [48] reaching the key outcome of the Beijing Consensus [8], a document offering recommendations on how to best harness AI technologies for SDG 4. Successively, it also issued specific guidelines [10] to aid policy-makers in the educational sector to enhance their capacity-building. These guidelines aim to facilitate a benefit-risk evaluation of the implementation of AI systems in school environments.

Moreover, UNESCO's guidelines explore some of the many challenges that need to be addressed to unleash the potential of AI and mitigate its downsides, among the others: ensure inclusion and equity (e.g., control over algorithmic biases), develop quality and inclusive data systems, guarantee ethics and transparency in data collection, use and dissemination (e.g., ensure data protection and privacy), prepare teachers for an AI-powered education (e.g., update curricula of both schools and teaching training appropriately) while preparing AI to support education (e.g., by specific choices of the pedagogy used in AI

[3] China AI Rules on Content Control: https://www.ft.com/content/1938b7b6-baf9-46bb-9eb7-70e9d32f4af0.

[4] China Regulation on recommendation algorithms: https://www.chinalawtranslate.com/en/algorithms.

[5] Chine rules for Synthetically Generated Contenthttps://www.chinalawtranslate.com/en/deep-synthesis/.

[6] China draft rules on Generative AI https://digichina.stanford.edu/work/translation-measures-for-the-management-of-generative-artificial-intelligence-services-draft-for-comment-april-2023/.

tools), and make research on AI in education significant (e.g., evaluate the efficacy of AI interventions, the role of teachers, and the impact on learner agency) [9,10].

Along the same line, in 2021 UNICEF published the "Policy guidance on AI for children" [44], a document that provides governments, policymakers and businesses that develop, implement or use AI systems with nine requirements for building AI policies and systems that uphold child rights. Among these, the eighth requirement (i.e., *"Prepare children for present and future developments in AI"*) drag up the topic of AI in education directly into the discussion. For this reason, UNICEF recommends to leverage the use of AI systems in education, when evidence demonstrates its benefits. In particular, it stresses the valuable help that proven AI-augmented educational platforms can provide to marginalized children, those with special needs and for personalized education for minorities [44].

In 2021, UNESCO introduced the *Recommendation on the Ethics of Artificial Intelligence* [11], providing a framework for discussions on generative AI with a specific focus on education and research. Rooted in human rights and dignity, the recommendation emphasizes a human-centered approach to AI, aiming for a sustainable and inclusive future. It highlights the importance of trustworthiness and integrity in the entire life cycle of AI systems, emphasizing monitoring by stakeholders. Education and research are identified as key policy areas, promoting global AI literacy, collaboration, and essential skills development, particularly in regions with educational disparities. The recommendation calls for AI ethics curricula, ethical research initiatives, and inclusive education, focusing on underrepresented groups and fostering interdisciplinary research and critical evaluation. Overall, it advocates for a holistic and ethical approach to AI development and deployment.

In 2023, UNESCO expanded the discourse on generative AI with the release of the *Guidance for generative AI in education and research* and a concise guide on *ChatGPT and artificial intelligence in higher education* [12,13]. Both documents aim to establish a framework that encourages the integration of AI while addressing potential ethical and legal challenges. They emphasize a nuanced approach to the use of generative AI, distinguishing between its applications in research and teaching and learning. The roles of AI in education, exemplified by ChatGPT or any other Large Language Model (LLM), span a diverse range of functions, serving from being a *Possibility Engine* generating alternative expressions of ideas to a *Study Buddy* for reflection on learning material. However, the immediate challenges and ethical concerns in higher education include issues of academic integrity, plagiarism, and cheating. Moreover, the lack of current regulations for LLM has prompted apprehension, leading to calls for a pause in its development to better understand potential risks (i.e., Italy's decision to block Chat-GPT due to issues related to the collection of personal data and ethical concerns about age-inappropriate responses.[7]). UNESCO emphasizes the need for a thoughtful integration of LLMs into education systems, emphasizing critical and creative comprehension among users. Collaborative forums are encouraged for stakeholders to collectively assess the impact of LLMs on higher education and develop adaptive strategies, with the AI audit serving as a valuable engagement tool. Providing clear guidance on LLM use, negotiated with stakeholders, is considered vital, along with training for users to refine queries and optimize LLM inputs. Staff training is recommended to align support with chat-bots and AI tools, ensuring effective technology deployment. The AI audit, orchestrated by educational institutions, involves three phases: understanding the current situation, deciding on AI usage, and monitoring performance and equity. This structured framework aids institutions in navigating AI integration, fostering understanding, strategic decision-making, and ongoing monitoring.

[7] The order of the Italian Data Protection Authority (Garante) here:https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9870832.

3. European legal context

In this section, we describe the European legal context about the protection of vulnerabilities. In Section 3.1 we introduce the AI Act regulation concerning vulnerability in prohibited and high-risk AI systems. In Section 3.2 we present the discrimination risks related to the concept of vulnerability in the data protection framework. Lastly, in 3.3 we focus on the GDPR's regulation concerning the automated decision-making involving children.

3.1. The AI act regulation about vulnerability

The purpose of the AI Act is to introduce harmonized rules in the EU that address the risks of the use of AI systems. The legal framework will apply to both public and private actors inside and outside the EU as long as the AI system is placed on the Union market or its use affects people located in the EU. It concerns both providers (e.g., a developer of a CV-screening tool) and deployers of high-risk AI systems (e.g., a bank buying this screening tool). The regulation distinguishes between unacceptable, high, medium (limited), and low (minimal) risks AI use with the most stringent requirements in the case of high-risk AI use. The four risk categories are dependent on the fields of application of AI systems and the intended use, in line with existing EU product safety legislation. For instance, AI systems for facial recognition or social scoring are deemed unacceptable and prohibited (Article 5). Similarly, AI systems used in areas such as education, healthcare, employment, migration, justice and law enforcement are considered high-risk and are therefore subject to conformity assessment procedures and require additional safeguards (Article 8 et seq.) before being placed on the EU market or otherwise put into service. This will allow providers to demonstrate that their system complies with the mandatory requirements for a trustworthy AI (e.g. data quality, documentation and traceability, transparency, human oversight, accuracy, cybersecurity and robustness). Such an assessment must be repeated if the system or its purpose are substantially modified. Within the educational context, which is the main focus of this study, the European Commission has developed specific values on the use of AI and data [42] with the support of an appointed Expert Group on Artificial Intelligence and Data in Education and Training. Such guidelines have categorized four distinct use cases of AI systems based on the intended end-user. These include AI applications designed to: teach students (e.g., language learning applications), support student learning (e.g., formative writing assessment that provide automatic feedback on the learner writing), assist the teachers (e.g., essay automatic scoring), and promote diagnostic or system-wide planning (e.g., diagnosing of learning difficulties) [22,49].

Vulnerability currently stands as one of the main concerns in the creation of an AI legal framework capable of conjugating innovation and development of technologies with the safeguarding of fundamental rights (protected in the EU Charter of Fundamental rights), in order to lay dawn an embankment to the risk of manipulation [50]. The AI Act proposal pays particular attention to children and their rights because of their particular vulnerability. In these regards, the regulation bans AI practices that exploit vulnerabilities of children and others for the purpose of materially distorting their behavior in a way that *"causes or is likely to cause that person or another person physical or psychological harm"* (Article 5 (1)(b)). This provision deserves specific attention in the context of education where the line between beneficial and harmful use of (subliminal) manipulation becomes difficult to draw. However, the norm does not refer to children as such, but rather solely to the concept of *"age"*. This may mean that it can refer not only to children, but also to young adults (and to older people). Alternatively, it may mean that not all children are covered. On the basis of Recital 16 proposed by the Commission, the first explanation seems the most obvious, given that it speaks of *"children and people due to their age"*. Therefore, it is recommended that the wording of Article 5(1)(b) be

aligned with the Recital to eliminate ambiguities. However, it should be noted that on 14 June 2023, the European Parliament, in its Amendment n.38 to the AI Act, replaced the explicit reference to children by inserting: *"[…] Such AI systems deploy subliminal components individuals cannot perceive or exploit vulnerabilities of individuals and specific groups of persons due to their known or predicted personality traits, age, physical or mental incapacities, social or economic situation"*. This regulatory choice could be due to a desire to bring this provision closer to what is provided for in Article 5 (1)(b) with Amendment n.216. In fact, the latter leaves age as the criterion characterizing vulnerability, but at the same time widens the net by adding other types of vulnerability, such as *"social or economic situation"*, which make the actual assessment more complicated in practical terms. In these regards, interestingly the Regulation in its Article 5 mentions *"people with vulnerabilities"*, instead of *"vulnerable people"*. This idea of vulnerabilities as not a static attribute of a category of individuals (e.g., women), but a transient and contextual situation depending on the specific circumstances of the data processing can be found also in guidance from data protection authorities. For example, the Spanish DPA [51], in its official list of high-risk data processing practices, mentioned: *"data processing regarding vulnerable subjects or those who are at risk of social exclusion, including […] the victims of gender-related violence, as well as their descendants and persons who are in their guardianship or custody"*. Victims of gender-related violence might be regarded as vulnerable due to their gender, but it is not here a general assumption according to which women are vulnerable: it is rather a contextual evaluation (gender can be a source of domestic violence or similar forms of violence and so a source of vulnerabilities) [52]. Another passage of the AI Act proposal also refers to vulnerabilities, specifically focusing on the potential "vulnerable position" in which adversely impacted persons might find themselves in relation to the user of an AI system, in particular *due to an imbalance of power, knowledge, economic or social circumstances, or age* (Recital 16). This vulnerable position of impacted persons need to be taken into account by the Commission for assessing the possible future qualification of additional systems as high-risk, under the Article 7 (2)(f). The acceptance of contextual vulnerabilities becomes evident in this provision, thus it includes the position of the user in relation to the digital service providers *"due to status, authority, knowledge, economic or social circumstances, or age"*, as amended by EU Parliament (n.250) as a disadvantaging factor. This is by far the broadest normative definition of vulnerability in the field [50]; nevertheless, it applies only where AI systems might pose high-risk to fundamental rights (and health and safety), thus severely restraining its use. In fact, proposed Article 9 on risk management systems stipulated that when implementing such a system, specific consideration needs to be given to whether the high-risk AI system is likely to be accessed by or have an impact on children. High-risk are, for instance, systems used for *"assessing students in educational and vocational training institutions"*, which may or may not include children (i.e., under 18s) (Annex III).

Another considerable accomplishment in term of vulnerability may also be found in the type of control that the Proposal establishes: it requires, at least for high-risk systems, an *ex-ante* control, which appears to, at least partially, redress the problem of systemic vulnerability, by transferring a part of the burden from the recipient (as an active actor who personally experiences harm) to the provider (as an active actor who must control the rightness of his activities) [50]. These requirements include the implementation of a risk management system (Article 9), of accurate data governance principles (Article 10), the duty of technical documentation (Article 11) and record-keeping (Article 12), transparency measures for companies that will use those AI systems (Article 13), comprehensive human oversight duties (Article 14) and, most importantly, accuracy, robustness and cybersecurity standards (Article 15). For the purposes of this study, only the most relevant ones will be illustrated. Risk management is one of the first requirements that basically defines the obligation for companies to test their systems in advance for their impact. In fact, the aim is to put the least harmful system on the market. As already mentioned, Article 9 explicitly considers the impact of high-risk systems on children. Moreover, for that purpose, a data-sets should also be controlled in view of the elimination of possible biases and assessments on the relevant design choices, as prescribed in Article 10 (2). Providers must inform the recipient about the intended purpose, circumstances that may lead to foreseeable risks and its performance with respect to the intended recipients of such system (Article 13). In this respect, the intervention of human oversight is thus not only strongly recommended, but mandated (Article 14). The AI providers need to comply with these requirements and "justify" their compliance through a conformity assessment. In case the conformity assessment is not carried out, or is carried out incorrectly or irregularly, the market surveillance authority can *"take all appropriate measures to restrict or prohibit the high-risk AI system being made available on the market or ensure that it is recalled or withdrawn from the market"* (Article 68(2)) in addition to monetary sanctions (Article 71). The main core of the AI Act proposal is based on an ex ante "licensure" approach, where the AI providers need to "justify", through some technical documentations, that their system is adequate according to specific principles (transparency, accountability, human oversight, accuracy, security) [16].

In June 2023, the European Parliament adopted its negotiating position. Many recitals and some articles have been added to emphasize the legislative need to protect fundamental rights. For instance, Article 29a introduced the *Fundamental impact assessment for high-risk AI systems*. In this context, deployers that are bodies governed by public law or private operators providing public services, and operators providing high-risk systems shall perform an assessment of the impact on fundamental rights and notify the national authority of the results. The proposed Article at the paragraph 1 (f) established that the assessment shall include *"specific risks of harm likely to impact marginalized persons or vulnerable groups"*. Even though children are not mentioned, we should interpret them as belonging to vulnerable groups. Interestingly, if the provider already met this obligation through the data protection impact assessment, the fundamental rights impact assessment shall be conducted in conjunction with that data protection impact assessment (Article 35 GDPR).

In conclusion of this analysis, we highlight that one significant important issues in the future application of the AI Act will be the definition of high-risk systems. Actually, the rules that apply to them form the bulk of the Regulation. For this reason, according to some authors, it would be preferable for the law to define high-risk AIs by means of a risk assessment based on reviewable criteria, rather than on an arbitrary list of existing use cases, so as to make the legislation future-proof [53]. Hence, it is important to note that children were indirectly included by the Parliament on the list of considerations that the Commission should take into account when assessing the risk posed by a system in its role as designating use-cases of AI systems as high-risk.[8] According to the Parliament, the Commission's assessment should evaluate the extent to which there is an imbalance of power, or the potentially harmed or adversely impacted persons are in a vulnerable position in relation to the user of an AI system due to, among other things, knowledge, economic or social circumstances or age. Although it would be better for children to be mentioned explicitly for their protection to be effective, we may assume that the inclusion of vulnerability and age at least offers some leeway to interpret childhood as a relevant factor.

[8] https://europeanlawblog.eu/2023/09/12/children-and-the-artificial-intelligence-act-is-the-eu-legislator-doing-enough/

3.2. Risks of discrimination — the concept of vulnerabilities in the EU data protection framework

In many EU legal fields, there are clearer definitions of different individuals involved and of possible vulnerable categories [54] among them. This is the case of, for example, EU private law, consumer law [55], car insurance regulation [56], regulation of scientific research [57], AI regulation [1]. In these fields, there are, for instance, descriptions of average subjects and separate descriptions of vulnerable individuals (classifying either generally or on the basis of the specific groups they belong to) [58]. In the data protection framework the concept of vulnerability emerges in the form of harms to which individuals are exposed. AI data-driven systems can serve as tools of potential discrimination, manipulation or may lead to physical and psychological harms (also Article 5 (1)(b) of the AI Act proposal). Although the GDPR does not contain an explicit definition of vulnerable data subjects, there are at least two indirect references to vulnerable individuals: the protection for children (Article 8 and 12) and the notion of risk to fundamental rights and freedoms, including *inter alia* the analysis of whether data subjects are vulnerable as regards their rights and freedom (Recital 75). The rational for protecting children is their decisional vulnerability during the data processing, the implications, risk and related rights, as well as the incapability to give valid consent and to exercise the data protection rights. Such form of decisional vulnerabilities concerns not only children, but also other data subjects. The other rationale for protecting vulnerable data subjects, as suggested at Recital 75, is the higher capability in incurring high risks to fundamental rights and freedoms. In particular, we need to take into account rights and freedoms protected in the EU Charter of Fundamental Rights [58].

Different examples from law enforcement, welfare, banking or housing are showing that those technologies can reinforce social inequalities and lead to discrimination in the access to services and goods [59,60]. On these ground, the rights to privacy and data protection are fundamental rights which must be respected at all times (Article 8, European Convention of Human Rights; Article 7 and Article 8 of European Charter of Fundamental Rights - hereinafter ECFR). AI systems must be built in a way that embeds the principles of data minimization (Article 5) and data protection by design and by default (Article 25) as prescribed by the EU's General Data Protection Regulation (hereinafter GDPR) [61]. Privacy rights must be safeguarded by data governance models that ensure data accuracy and representativeness; protect personal data and enable humans to actively manage their personal data and the way the system uses it. Appropriate personal data protection can help developing trust in data sharing and facilitate data sharing models uptake. Moreover, data minimization and data protection should never be leveraged to hide bias or avoid accountability, and these should be addressed without harming privacy rights, especially in case of vulnerable subjects such as children. In this respect, the notion of vulnerability plays a significant role in the human right discussion. Although the concept of vulnerability is neither present in the European Convention on Human Rights (hereinafter ECHR) nor in the ECFR, scholars, human rights institutions and organizations refer to it as an imperative that entails special protection of socially marginalized groups like women, people with disabilities, children, or ethnic minorities (refer to [62] for a general discussion and to [63] for a focus on the US case). Therefore, not surprisingly, the ECHR Court has firstly addresses the idea of vulnerable persons in 1981, referring specifically to children. In Dudgeons v. UK the Court it referred to the *"moral interests and welfare of certain individuals who are in need of special protection for reasons such as lack of maturity, mental disability or state dependence"* [64]. In this judgment, the Court adopted the idea of inherent vulnerability based on (age as an index of) weakness, inexperience and dependence [65]. Notwithstanding, the ECHR recognizes vulnerable situations of particular groups, but unfortunately it never employed the notion of vulnerability in the field of private life, privacy or data protection (Article 8 ECHR) [65]. In the digital era, vulnerable data subject require special attention when determining whether a data processing operation poses a high risk to their rights and freedoms. Recital 75 of GDPR establishes that those risks *"may result from personal data processing which could lead to physical, material or non-material damage, in particular (…) where personal data of vulnerable natural persons, in particular children, are processed"*. In other words, some subjects should be protected not only because of their limited capacity to understand and give consent, but from higher risks of material or non-material damages. The examples in the digital arena might be several: risk of discrimination during an automated data processing, including profiling; risk to be more easily impaired in one's freedom of thought when data are processed for direct marketing; risk to have bigger physical or psychological damages in case of data breach; risk to suffer from higher risks of damages from a data processing, etc. In light of this, it is evident that children represent a category that is more vulnerable and exposed to greater risks of harm. Therefore, in these situations, an AI system must *process personal data in a lawful, fair and transparent manner* (Article 5 (1)(a) GDPR).

In the data protection landscape, there are two primary dichotomies in human vulnerability theories. The first pertains to the characterization of subjects who are vulnerable and involves a distinction between the concepts of *universality*, which posits that all individuals are equally vulnerable, and *particularity*, which instead suggests that certain subjects may be more vulnerable than others. The second dichotomy relates to the various forms in which vulnerability can manifest. Specifically, vulnerability can arise during the data processing stage, which includes decisional vulnerability risks that are associated with data collection, consent provision, and inappropriate exercise of data protection rights; alternatively, vulnerability can also manifest as a result of the outcomes of the processing, whereby certain data processing activities may give rise to discrimination, manipulation, or secondary harms such as physical or psychological harm. According to some scholars, there is no single definition of a vulnerable individual in the EU [65]. Moreover, although specific lists of vulnerable individuals exist in several areas, the general framework reveals a highly contextual and relational comprehension of vulnerability, particularly based on power imbalance capable of creating harm (this line of interpretation seems also adopted by the GDPR [46]). That is to say, being vulnerable – in different legal domain – generally means to be more exposed to harm (compared to other individuals) *in some particular contexts* [66,67]. In such reconstructions, power imbalance and context are an incontrovertible fact for understanding the phenomenon: e.g., if the data controller (Article 4(7)) want to process personal data on the basis of legitimate interests (Article 6(1)(f) of the GDPR), it need to consider the nature and source of the legitimate interest, if there are additional safeguards and what is the impact on data subject, considering in particular *"the status of the data controller and data subject, including the balance of power between the data subject and the data controller, or whether the data subject is a child or otherwise belongs to a more vulnerable segment of the population"* [68]. Moreover, *"the question whether the data subject is an employee, a student, a patient, or whether there is otherwise an imbalance in the relationship between the position of the data subject and the controller must certainly be also relevant. It is important to assess the effect and risk of actual processing on particular individuals"* [68]. In these regards, the European Data Protection Supervisors (EDPB), commenting in June 2020 on the European Commission's approach to AI [69], suggested in that *"in the absence of a formally adopted legal definition of vulnerable groups"*, a *"context-specific, pragmatic approach"* should be adopted [52].

In the GDPR the notion of threats to fundamental rights and freedom is pivotal. In particular, according to the risk-based approach in the GDPR (Article 24), the data controller is obliged to implement appropriate technical and organizational measures to ensure the compliance with the data protection principles: *"taking into account the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for the rights and freedoms of natural persons"*. In particular, when assessing such risks, the controller should take into account

scenarios in which specific individuals, who may be more vulnerable, could be disproportionately impacted by the processing of certain data [65]. Therefore, the risk-base approach can play a significant role in mitigating potentially harmful outcomes of data-driven technologies. Moreover, according to the principle of data protection by design (the above mentioned Article 25), the controller shall, both at the time of the determination of the means for processing and at the time of the processing itself, implement appropriate technical and organizational measures, which are designed to implement data-protection principles. Even in this case the controller should take into account *"the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons"*, but also *"the state of the art [and] the cost of implementation"*. Essentially, the main difference between Article 24 and Article 25 lies in the fact that the former requires the data controller to demonstrate adherence to data protection principles, whereas the latter mandates the implementation of said principles [70]. In both cases, the attention to vulnerable data subjects and the implementation of specific safeguards to protect their rights and freedoms (i.e. to mitigate factors of vulnerability) seems necessary [65].

3.3. The GDPR's regulation concerning the use of automated decision systems involving children

The GDPR has significant implication for algorithmic decision-making [71]. At first, the legal debate focused on whether the GDPR created an individual right to an explanation of an individual algorithmic decision [72–75]. Subsequent legal analysis, however, began to focus instead on other accountability tools [59], either required by the text of the GDPR or recommended by the interpretative Guidelines on Automated Individual Decision-making and Profiling (hereinafter Guideline on ADM) from the Article 29 Working Party (now the European Data Protection Board) [76]. These tools include third-party auditing, the appointment of Data Protection Officers (hereinafter DPO) (Article 37), and the requirement of Data Protection Impact Assessment (DPIA) (Article 35) [76].

According to some authors, the GDPR combines a series of individual rights (Articles 12-23) with a systemic governance regime overseen by regulators, targeted at more comprehensive oversight over the algorithm and the people around it (Articles 24-43 & throughout) [77]. These two systems interact and overlap. An individual right is often also a company's duty. But even if individuals (data subjects) fail to invoke their rights, companies (data controller) have significant obligations – both procedural and substantive – under the GDPR [71]. Indeed, the GDPR has tried to provide a concrete solution to risks of automated decision-making through different legal provisions: a right to receive and access meaningful information about logic, significance and envisaged effects of automated decision-making processes (Articles 13(2)(f); 14 (2)(g); and 15 (1)(h); the right not to be subject to automated decision making (Article 22) with several safeguards and restrains for the limited cases in which automated decision making is permitted [73,78]. The Article 22 (1) states that *"the data subject shall have the right not to be subjected to a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affect him or her"*. This right shall apply almost always in case of sensitive data (Article 22 (4)). For other personal data, shall not apply in only three cases: the decision is authorized by Union or Member State law [79]; it is necessary for a contract; or is based on the data subject's explicit consent (Article 22 (2)). In the last two cases, *the data controller shall implement suitable measures to safeguard the data subject's right and freedom and legitimate interests, at least the right to obtain the human intervention on the part of the controller, to express his or her point of view and to contest the decision"* (Article 22 (3)). In addition, Recital 71 explains that such suitable safeguards *"should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision"*. According to some authors [72,73], automated decision with significant effects must be "legible" to individuals, in the sense that individuals must be able to understand enough about the decision-making process to be able to invoke their other rights under the GDPR, including the right to contest the decision [80,81].

As we already highlight, the GDPR creates additional obligations for data controllers when they are processing children's personal data. Article 22 itself makes no distinction as to whether the automated decision processing concerns adults or children. However, Recital 71 says that solely automated decision-making, including profiling, with legal or similarly significant effects should not apply to children. Given that this wording is not reflected in the Article itself, the Guideline on ADM does not consider that this represents an absolute prohibition on this type of processing in relation to children. However, in the light of this Recital, the Guideline on ADM recommends that, as a rule, controllers should not rely upon the exceptions in Article 22(2) to justify it. There may nevertheless be some circumstances in which it is necessary for controllers to carry out solely automated decision-making, including profiling, with legal or similarly significant effects in relation to children, for example to protect their welfare or, as in our case scenario, to improve the student's performance. If so, the automated processing may be carried out on the basis of the exceptions in Article 22(2)(a), (b) or (c) as appropriate. Therefore, we strongly believe in our case that the consent from the student's parents to use the AI4SPP system should be mandatory, as well as their right to receive and access meaningful information about the logic involved (Article 13(2)(f); 14 (2)(g); and 15 (1)(h). Indeed, the Guideline of ADM declares: *"in those cases there must be suitable safeguards in place, as required by Articles 22(2)(b) and 22(3), and they must therefore be appropriate for children. The controller must ensure that these safeguards are effective in protecting the rights, freedoms and legitimate interests of the children whose data they are processing"*. For this scope, one possible legal protection for vulnerable data subjects is the Data Protection Impact Assessment (DPIA) [65]. In particular, Article 35 (as interpreted by Recital 75 and by Guidelines on Data Protection Impact Assessment) [76] requires performing a DPIA in case of high-risk data processing, including the case where the data subjects can be considerable vulnerable. The DPIA is based on several steps (Article 35 (7) and Recitals 84, 90): i.e., the systematic description of the processing, the assessment of necessity and proportionality, the assessment of risks and the description of measures envisaged to mitigate such risks. In other words, even according to the accountability principle, it is the controller who should autonomously determine measures for protecting vulnerable individuals. Data controllers may suggest mitigation measures for specific vulnerable groups: for instance, in case of decisional vulnerability, the data controller could implement specific forms of consent or information disclosure measures; in case of individuals that might be easily discriminated, the data controller could implement periodical audits against discrimination, or regular quality assurance checks, or establishing data minimization and clear retention periods, using pseudonymization techniques, certification mechanisms, etc. Therefore, according to [71], the DPIA serves a dual purpose: it is a tool in the GDPR's systemic (and collaborative) governance regime, and it is an element of the GDPR's protection of individual rights. Firstly, the DPIA can function as a resource for the provision of information to individuals regarding algorithmic decision-making, specifically with regards to individual notification and access rights. As an example, individuals have the right to receive "meaningful information" concerning *"logic involved, as well as the significance and the envisaged consequences"* of automated decision-making processes (as stated in Articles 13, 14, and 15 of the GDPR). A DPIA should include a comprehensive description of the planned processing activities and their purposes (as outlined in Article 35(7)). If organizations already document automated decision-making processes at a systemic level during the DPIA process, these internal descriptions could be shared with individuals or used as a basis for such disclosures. Additionally,

the GDPR and the accompanying Guidelines on ADM suggest that DPIA can serve as a way to demonstrate a commitment to protecting and facilitating individual rights in algorithmic due process, especially in the context of algorithmic decision-making. DPIA play a role in linking the GDPR's collaborative governance system with its individual rights framework by imposing systemic accountability measures like audits or external reviews. Consequently, companies can develop concrete methodologies to mitigate risks such as unfairness, errors, bias, discrimination, and so on [71].

The ethics guidelines for trustworthy AI and the ALTAI checklist consider the seven requirements and a collection of questions designed to assist AI designers, developers and deployers in following a multidisciplinary evaluation process.

4. The ethics guidelines for trustworthy AI and the ALTAI checklist

In 2018 the European Commission appointed a panel of experts to offer guidance on its AI strategy, namely, the *High Level Expert Group on AI* (AI HLEG). Operating as an independent group, AI HLEG is responsible for defining principles for the ethical advancement of AI systems, along with providing policy and investment suggestions. In 2019 the AI HLEG presented the final version of the so-called "Ethics Guideline for Trustworthy AI" (*AI Ethics Guidelines*) [22]. This document addresses all AI stakeholders, seek to present a human-centric perspective on AI and lists seven criteria that AI systems must fulfill in order to be considered trustworthy. Based on this deliverable, the AI HLEG successively presented the ALTAI checklist [27], an additional resource that can be utilized to provide guidance on ensuring a trustworthy approach during the design, development, deployment, and usage of AI-based solutions that uphold fundamental rights and freedoms. More precisely, the ALTAI checklist serves a dual purpose in facilitating risk assessment and the implementation of corresponding mitigation measures, as well as identifying specific roles and responsibilities at each stage of the design process for a given AI-based technology. Moreover, it raises awareness of the potential impact of AI on society, the environment, consumers, workers and citizens (in particular children and vulnerable people belonging to marginalized groups), encouraging the involvement of all relevant stakeholders in the process. A trustworthy approach is crucial in facilitating "responsible competitiveness" by determining the fundamental basis on which all those utilizing or impacted by AI systems can trust that their design, development and use are lawful, ethical and robust [27].

In operational terms, ALTAI is a collection of questions designed to assist AI designers, developers and deployers in following a multidisciplinary evaluation process. This approach aims to tackle the seven challenges related to ethical, legal, and robust compliance that were identified in the Ethical Guidelines on Trustworthy AI endorsed by the EU Commission [22]. According to such guidelines, an AI system attains trustworthiness when it adheres to the three key criteria introduced in Section 1: legality (i.e., compliance with relevant legal regulations), ethics (i.e., compliance with applicable ethical standards), and robustness (i.e., compliance with relevant safety standards). The relationship among these three pillars is influenced by the fundamental rights impact assessment and the following seven factors:

1. **Human Agency and Oversight** encompasses both ethical and legal aspects, emphasizing the protection of fundamental rights (EU Charter of Fundamental Rights) and the delicate equilibrium between human control and technological progress. It prioritizes the well-being of individuals and groups, which is achieved by facilitating the oversight over the system and the limitation of automation bias, user manipulation and illegal discrimination.
2. **Technical Robustness and Safety** pertain to the system's ability to withstand attacks and ensure security. It involves the implementation of fallback plans and adherence to the highest standards of general safety, accuracy, reliability, and reproducibility.
3. **Privacy and Data Governance**: the rights to privacy and data protection are fundamental rights which must be respected at all times. AI systems must be built in a way that embeds the principles of data minimization and data protection by design and by default as prescribed by GDPR.
4. **Transparency** is a principle put in place to ensure the traceability, explainability, and effective communication of the methods, objectives, and outcomes of an AI system.
5. **Diversity, Non-discrimination, and Fairness** encompass the interdisciplinary measures that should be implemented to prevent the misuse or unfair use of AI. This includes addressing issues such as bias, accessibility, and promoting universal design to ensure equitable and unbiased outcomes.
6. **Societal and Environmental Well-being** emphasizes that AI systems should be introduced into the market as sustainable solutions, taking into account environmental, social, and societal perspectives. This includes aligning with the democratic values embedded within the framework of the European Union.
7. **Accountability** is the central principle that facilitates the compliance process by proactively allocating responsibilities through a risk-based approach. This includes elements such as auditability, minimizing and reporting negative impacts, considering trade-offs, and providing avenues for redress.

These seven key requirements have been transformed into a comprehensive set of 80 questions and sub-questions that necessitate interdisciplinary expertise for successful solution. A mere binary yes or no response is insufficient because a comprehensive analysis is necessary to form a deliberate and informed opinion regarding the trustworthiness level of the designed tool [82]. To provide an example, the initial set of questions (Q1-Q4) aims to evaluate the impact of the designed application on European Union fundamental rights [83]. Consequently, a thorough understanding of fundamental rights protection and the ability to assess the associated impact must be incorporated into the evaluation process workflow [84].

The second set of question (Q5–Q16) is designed to promote human agency and decision-making, in accordance with the notion of respecting individual autonomy. AI systems must serve as facilitators for a democratic, prosperous, and fair society by promoting the user's autonomy, while also safeguarding fundamental rights through human supervision.

Moreover, to adequately address the third set of questions (Q17–38) about technical robustness and safety, a comprehensive understanding of cyber-security and safety standards is essential. It is crucial for the AI designer to explain the reasoning behind their choice of specific technical measures and the implementation of particular safeguards, considering various aspects such as human safety, animal protection, environment, security, and misuse. Additionally, the designer must establish a "fallback plan" to ensure an acceptable level of risk management in case of unforeseen circumstances. These evaluations also have a proactive dimension, as they contribute to answering questions in the seventh set (Q63–Q71) that pertain to societal and environmental well-being [85]. Therefore, trustworthiness encompasses not only the prevention of harm but also the ability to address the multifaceted challenges associated with empowering society. In particular, individual well-being means people can live fulfilling lives, in which they are able to pursue their own needs and desires in mutual respect. Social well-being refers to the flourishing of societies, whose basic institutions, such as healthcare and politics, function well, and where sources of social conflict are minimized. Environmental well-being refers to the well-functioning of ecosystems, sustainability, and the minimization of environmental degradation. For that reasons, AI systems should not contribute to any harm to individual, societal or environmental well-being, but instead AI systems should strive to make a positive contribution to these forms of well-being [61].

The fourth set of questions (Q39–45) emphasizes the importance of GDPR compliance. The results obtained from the DPIA conducted

in accordance with the well-know Article 35 of the GDPR must be incorporated into the broader framework of the AI-based system. In addition to the points discussed in the preceding paragraphs, it is worth emphasizing that the involvement of a DPO is not only recommended but also regarded as an organizational measure that plays a significant role in attaining the desired level of reliability for the developed ecosystem. In fact, the presence of a DPO contributes to ensuring compliance with data protection regulations and promoting a culture of trust and accountability within the organization [86]. As an initial step, it is crucial for the AI designer/developer to assess the need for consulting and appointing a DPO and/or a privacy expert. This role is closely intertwined with the first set of questions, as the impact on other fundamental rights is directly connected to ensuring the confidentiality, availability, and integrity of personal data in a cause-and-effect relationship. For instance, in our case study, a data breach by the AI system could pose risks to the dignity, privacy, and right to education of vulnerable individuals, such as students. Furthermore, such a breach may lead to significant instances of unacceptable discrimination. The evaluation of fairness, encompassing these critical considerations, is specifically addressed in the sixth set of questions Q52–Q62 within the evaluation checklist. According to the ALTAI checklist, the concept of fairness involves various dimensions, including equity, impartiality, egalitarianism, non-discrimination, and justice. While this definition is still open to debate, fairness entails that all people are entitled to the same fundamental rights and opportunities. This does not require identical outcomes, i.e., that people must have equal wealth or success in life. However, there should be no discrimination on the basis of the fundamental aspects of one's own identity which are inalienable and cannot be taken away. Various legislations already acknowledge a number of them, such as gender, race, age, sexual orientation, national origin, religion, health and disability. Broadly speaking, fairness comprises both "substantive" and "procedural" fairness. Procedural fairness requires that the procedure was not designed in a way that disadvantages single individuals or groups specifically. Substantive fairness entails that the AI does not foster illegal discrimination patterns that unduly burden individuals and/or groups for their specific vulnerability [61].

The fifth set of questions Q46–Q51 pertains to transparency and encompasses inquiries regarding the design and development of an AI-based system. Transparency requires that the purpose, inputs, and operations of AI programs are knowable and understandable to its stakeholders [87]. In fact, transparency and explanation have a significant impact on all aspects of an AI system, including the data, the system itself, and the processes involved in its design and operation. Stakeholders must be empowered with the ability to comprehend the key concepts underlying the AI system, such as how it functions and arrives at its decisions, and its purpose [88]. Claims related to IP rights, confidentiality, or trade secrets should not impede transparency as long as they are appropriately preserved. Selective transparency, such as confidentially sharing information with trustworthy third parties, technology-based solutions, and confidentiality commitments, can be utilized to achieve this [89,90]. Moreover, transparency is essential to realize other principles: respect for human agency, privacy and data governance, accountability, and oversight [91]. Without transparency (meaningful information about the purpose, inputs, and operations of AI programs), AI outputs cannot be understood, much less contested. [59,92]. This would make it impossible to correct errors and unethical consequences [93]. Therefore, when building an AI solution, one must consider what measures will enable the traceability of the AI system during its entire life-cycle, from initial design to post-deployment evaluation and audit or in case its use is contested [94,95]. Indeed, explainability is a particularly relevant requirement for systems that make decisions or recommendations or perform actions that can cause significant harm, affect individual rights, or significantly affect individual or collective interests [96].

The final set of questions Q72–Q80 aims to evaluate the overall accountability of the process. This principle in the context of AI applications involves the acknowledgment and acceptance of responsibility by those involved in their development or operation [97,98]. It encompasses transparency and oversight as fundamental prerequisites [99]. To be accountable, developers and operators of AI systems should be capable of providing justification regarding the system's behavior and the resulting outcomes [100]. Human oversight is crucial, as it ensures that human actors can comprehend, supervise, and regulate the design and functioning of the AI system [71]. Indeed, accountability relies on effective oversight, enabling developers and operators to assume responsibility and take appropriate actions based on their understanding and control of the system's operations [101]. Thus, developers must possess the ability to elucidate the rationale behind the system's characteristics to ensure accountability [102]. These measures are crucial for ensuring that the AI system operates responsibly and that appropriate solutions are in place to handle potential problems. It is imperative to thoroughly document how any ethically and socially undesirable effects, such as discriminatory outcomes or lack of transparency, will be identified, prevented, and addressed within the system [95,103]. Human oversight and control over decision cycles and operations of AI systems must be facilitated, unless there are compelling justifications demonstrating that such oversight is unnecessary. Any such justifications should include an explanation of how humans will interpret the system's decisions and the mechanisms in place for human intervention [104]. The proposal should include an evaluation of the ethical risks associated with the proposed AI system, including risk assessment procedures and post-deployment mitigation measures. Additionally, consideration should be given to how end-users, data subjects, and other parties can report complaints, ethical concerns, or adverse events, and how these reports will be evaluated, addressed, and communicated back to the relevant parties [105]. As a general principle, all AI systems should be auditable by independent third parties, including the procedures and tools used during the development process (i.e. Explainable AI approach (XAI) [106,107]). In conclusion, human-accessible logs of the AI system's internal processes should be generated, where applicable.

5. The AI4SPP tool: design and implementation

This section outlines our implementation of the Proof-of-Concept for AI4SPP, an AI-driven application designed to forecast students' grades using interpretable and explainable features. The AI4SPP tool aim is to assist educators in monitoring the academic progress of their students, with the added benefit of early identification of those at risk of school failure. While developing the tool, our focus was on highlighting the essence of a trustworthy AI development process, diverging from conventional practices to incorporate both legal and ethical dimensions. Thus, this case study serves as a prime illustration of how the ALTAI checklist can be employed as a methodical means to comprehensively assess the trustworthiness of AI systems.

5.1. Data description and preprocessing

AI4SPP was trained on the *Student Performance* dataset,[9,10] which is a [open-access dataset] that provides details regarding the learning achievements of students enrolled in two Portuguese public secondary education institutions during the 2005–2006 school year. Despite the

[9] Dataset available at:https://archive.ics.uci.edu/ml/datasets/student+performance.

[10] Due to the extremely sensitive nature of the data involved (i.e. highly confidential information on vulnerable individuals), it is very complex to have children's data available, so for this study we selected a public dataset that relates to slightly older students, i.e. between the ages of 15 and 22.

data not being very recent, we chose to utilize this particular dataset for our Proof-of-Concept due to its open-access nature, its representation of vulnerable subjects (i.e., school students rather than university students), and its inclusion of sensitive attributes that necessitate compelling ethical evaluation. As reported in [108], the data in the *Student Performance* dataset was gathered through school reports and questionnaires, with the latter being used to collect individual attributes including demographic, social, emotional and additional school related features. The dataset provides the students' grades corresponding to three different periods (first semester, second semester and final grade) in two classes, namely Portuguese language (649 students) and mathematics (395 students). Specifically, we used data from the Portuguese class to train the machine learning models and we successively tested them for the prediction of grades in mathematics. This choice allowed us to evaluate the robustness of the AI systems when employed to predict grades for different school subject.

In accordance to the Portuguese grading system, all grades in the dataset are given on a scale from 0 to 20, with 10 being the minimum passing. Among the other attributes, 17 are categorical, while the remaining 13 are numerical, either binary or discrete. Some of these attributes are sensitive and might enhance human biases when fed through an AI system. Illustrative examples we want to underline are features like sex, address type (rural or urban), number of family members, parent's jobs and educational level and romantic relationship of the student. Before fitting the models, we encoded all categorical features: we converted the four ones having only two values to numerical binary variables and we one-hot encoded the remaining 13 nominal attributes. Since the dataset has no missing values, there was no need to employ any data imputation procedure. In order to overcome the high imbalanced distribution of students' grades, we employed the Synthetic Minority Over-sampling Technique (SMOTE) [109] for both the binary and multiclass classification tasks described in Section 5.2.

5.2. System design and implementation details

In order to foster a human-centric adoption of our AI system, we deliberately built AI4SPP to offer assistance in a customized manner based on the amount of information required by the end-user, i.e., a teacher or educator. First of all, the AI system can be queried to provide a recommendation for the partial grades of the first (G1) and second semester (G2), as well as for the final grade at the end of the school year (G3). It is important to note that the predictions for G1 are solely based on the dataset's non-grading attributes, hence a lower level of performance can be expected in this case. On the other hand, the estimate of subsequent grades also takes into account the prior marks given by the end-user: while predicting G2 grades, the system will consider both non-grading features together with G1 grade score. Similarly, while predicting G3, the system will consider non-grading features, G1 and G2 previous grades. As a result, we would expect the model gradually adjusting to the students' real academic achievements and improving its predicted performance.

AI4SPP has been designed to return its grade estimate for three distinct predictive tasks, namely:

1. *Binary classification*: in this case, AI4SPP only predicts whether a student will pass or fail the class, where a positive evaluation correspond to grades from 10 to 20. Note that this type of results is only useful for the prediction of the final grade.
2. *Multiclass classification*: the result of this task is the prediction of a qualitative evaluation from a 6-levels grading system. The outcomes are encoded as follows: Fail (grades from 0 to 9), E (sufficient, 10-11), D (satisfactory, 12-13), C (good, 14-15), B (very good, 16-17), and A (excellent, 18-20 grades).
3. *Regression*: in this task, the exact grade in the range 0-20 is predicted.

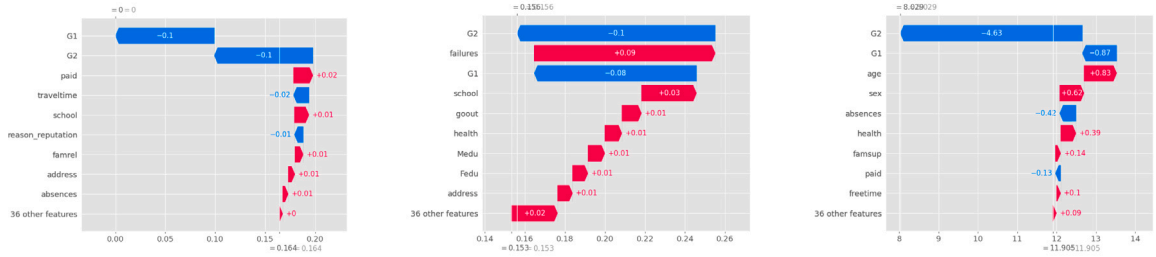
To find the best fitting model, we performed an extensive grid search for hyper-parameter tuning for each of the three tasks. Specifically, the comparison for tasks (1) and (2) covered the following models: Logistic Regression, Random Forest, K-Nearest Neighbors, Support Vector Machines [110], CatBoost [111] and XGBoost [112]. Performance were compared between Linear Regression, Random Forest, K-Nearest Neighbors, Support Vector Regression Machine [113], CatBoost [111] and XGBoost [112] for task (3) instead.

Ultimately, a crucial aspect of our implementation involved empowering AI4SPP with supplementary functionalities that enable it to report not only its point-estimate, but also its level of confidence and explanations in terms of *eXplainable AI* (XAI) outcomes. Indeed, these additional pieces of information have been argued to be effective in enhancing the performance of human-AI assisted decision-making settings such as the one described in our case study. This is due to the fact that they enable the end-users to accurately gauge their level of trust towards the AI system [114–117].

All together, by querying AI4SPP an end-user can get access to a number of information, that we synthesize in the following and exemplify through a concrete example in Box 5.2 and Fig. 2:

1. Point-estimate predictions of the students' grades. In particular, AI4SPP provides its recommendations with respect to the following questions:
   (a) Will the student pass the class at the end of the year?
   (b) Which qualitative evaluation will the student get at the end of each semester and as a final evaluation (i.e., excellent, very good, good, satisfactory, sufficient or fail)?
   (c) Which exact grade in the 0-20 scale will the student get at the end of each semester and as a final evaluation?
2. Confidence of the model expressed in terms of:
   (a) Overall performance metrics of the machine learning models (i.e., accuracy, F1-score and ROC-AUC score for the classification models, $RMSE$ and R^2 for regression). Such scores are conveyed to the end-user in a textual form and are accompanied by a definition to facilitate their correct comprehension.
   (b) Prediction probabilities, for instance: the student x could be predicted to have 60% of probability of passing the final exam and the remaining 40% of failing.
3. Explanations of the underlying black-box model, expressed in terms of
   (a) *Global explanations* via feature importance scores computed via impurity-based methods (e.g., through the built-in methods provided in gradient boosting algorithms) and Shapley values [118].
   (b) *Local explanations* via feature importance scores computed with SHAP [118] and counterfactuals rules computed with DiCE [119].

   It is worth mentioning that there are numerous explainable AI strategies available, but we have chosen not to do a comprehensive review of all the possibilities as it is outside the scope of this study. Likewise, in this Proof-of-Concept we refrain from performing a quantitative or qualitative assessment of explainable AI techniques, while acknowledging their importance in deployed AI systems [120].



(a) Binary classification (b) Multiclass classification (c) Regression task

Fig. 2. SHAP local explanations for the student with $id = 367$. *G1* and *G2* stand for the grade obtained in the first and second semester, respectively. Other acronyms are defined as follows: *absences* = number of school absences, *address* = student's home address type, *age* = student's age, *failures* = number of past class failures, *famrel* = quality of family relationships, *famsup* = family educational support, *Fedu* = father's education, *freetime* = how much free time after school, *goout* = how often go out with friends, *health* = current health status, *Medu* = mother's education, *paid* = extra paid classes within the course subject, *school* = student's school, *school_reputation* = school chosen by virtue of its reputation, *sex* = student's sex, and *traveltime* = home to school travel time.

A use case of AI4SPP

We exemplify the expected behavior of the AI4SPP tool by showing its possible outputs when queried with respect to the student with $id = 367$. We assume that a teacher decides to use AI4SPP to obtain a suggestion regarding the final grade to assign to the selected student. This situation is particularly critical since a mark below 10 would mean an unsuccessful achievement, which consequently entails the repetition of the school year. As described in Section 5.2, the teacher can query the AI-system in the following three different modalities:

1. **Prediction only.** In this case, the system only returns the point-estimate prediction of each class. In this specific case, both the binary and multiclass classifiers predict that the student will fail (class 0), while the regressor forecast the grade 8.
2. **Predictions + probability estimates.** The binary classifier predicts that the student will fail with 89% of probability; the multiclass classifier predicts that that he or she will fail with 75% of probability, while there is 19% of chance that he or she will get the score E and a 4% probability that he or she will get the score D.
3. **Predictions + probability estimates + eXplanations.** In addition to the information described in the previous points, in this case the AI-system also provides SHAP local explanation as shown in Fig. 2. Both classification models assign roughly the same importance score to the grades G1 and G2, while the regressor largely relies on G2 only. Among the few other features that negatively impact the prediction, of a particular note is the long travel time (between 30 min and 1 h) highlighted in the binary case. The regressor assigned the highest importance scores to the sex and age features, albeit small in magnitude. This finding is significant as it indicates that the black-box model may be utilizing sensitive features to generate its prediction. In the absence of any bias mitigation measures in the AI4SPP system, this could potentially heighten the likelihood of discrimination when the system is accessed through either of the two modalities described above, since they do not provide any explanation. Additionally to feature importance explanations, AI4SPP employs DiCE to produce counterfactual explanations that can be understood by educators as actionable measures to enhance a student's school performance and possibly avoid an insufficient final grade. For instance, as regards to the student taken here as an example, DiCE suggested to increase the weekly study time, but also, and most interestingly, to improve the current health status of the student.

5.3. Student performance prediction: evaluation of the AI-based tool

The results of the evaluation of the best fitted models for the three predictive tasks are reported in Table 1 (binary and multiclass classification) and Table 2 (regression). Based on such scores, we selected for AI4SPP the following models:

1. *Binary classification*: Support Vector Machine for the prediction of G3 ($accuracy = 0.950$, F_1-$score = 0.946$, $ROC - AUC$ score= 0.996).
2. *Multiclass classification*: CatBoost for the prediction of G1 ($accuracy = 0.629$, F_1-score= 0.615, ROC-AUC score= 0.883), Random Forest for G2 ($accuracy = 0.754$, F_1-score= 0.744, ROC-AUC score= 0.947) and G3 ($accuracy = 0.835$, F_1-score= 0.827, ROC-AUC score= 0.977).
3. *Regression*: Random Forest for the prediction of G1 ($RMSE = 2.183$, $R^2 = 0.290$) and XGBoost for G2 ($RMSE = 1.468$, $R^2 = 0.770$) and G3 ($RMSE = 1.269$, $R^2 = 0.851$).

Overall, on the evaluation dataset the AI-system achieved a very high accuracy and F_1-score in the binary classification task, while a gradual increase in performance from G1 to G2 can be observed in the regression and multiclass classification task. This suggests that particular care should be taken when employing the AI-system at early stages of the school year. The same trend, but with significantly lower performance, can be observed on the test dataset corresponding to the mathematics class, as reported in Table 3 and Table 4.

The reason for such an increase in performance can be easily understood by analyzing global explanations. The findings indicate that the black-box models that underlie AI4SPP are predominantly reliant on the previous grades that have been assigned to the students. In particular, the feature importance scores computed through impurity-based, permutation and SHAP methods all reveal significant higher values in correspondence to G1 for the prediction of G3 and in both G1 and G2 for the prediction of G3. Other features that all models identifies as having high predictive value are the school chosen (suggesting that one of the two school is either characterized by a lower educational offer or its teachers are particularly tight on grades), the number of previous failures and absences. Notably, the two sensitive features encoded in the dataset (i.e., sex and age) were not scored to have particularly high importance scores, suggesting that our AI-system does not discriminate towards these protected groups, at least in this pilot study.

6. Self-assessment of AI4SPP: the ALTAI cheklist

The ALTAI checklist has been employed as a guiding framework for the development of AI4SPP. Additionally, a fundamental rights impact assessment (FRIA) has also been conducted prior to utilizing the ALTAI checklist [22]. In the paragraphs below, we describe the results of

Table 1
Evaluation of the binary and multiclass classification models. In the latter case, the F1-score is computed as the macro-average of the per-class F1-scores.

Model	Binary classification		Multiclass classification					
	G3		G1∗		G2		G3	
	Accuracy	F_1-score	Accuracy	F_1-score	Accuracy	F_1-score	Accuracy	F_1-score
Logistic Regression	0.936	0.936	0.522	0.509	0.719	0.715	0.777	0.773
Random Forest	0.945	0.943	0.616	0.595	**0.754**	**0.744**	**0.835**	**0.827**
Support Vector Machine	**0.950**	**0.946**	0.549	0.543	0.696	0.692	0.810	0.805
K-Nearest Neighbors	0.927	0.919	**0.634**	0.588	0.728	0.711	0.814	0.803
CatBoost	0.941	0.938	0.629	**0.615**	0.746	0.737	0.810	0.802
XGBoost	0.941	0.938	0.594	0.583	0.701	0.692	0.806	0.800

∗ *CatBoost model has been preferred over KNN as the best model for the multiclass prediction of G1 as it achieves both higher* F_1 *and ROC-AUC scores.*

Table 2
Evaluation of the regression models.

Regression	G1		G2		G3	
Model	RMSE	R^2	RMSE	R^2	RMSE	R^2
Linear Regression	2.291	0.217	1.509	0.757	1.271	0.851
Random Forest	**2.183**	**0.290**	1.819	0.647	1.635	0.753
Support Vector Regression	2.279	0.226	1.535	0.749	1.200	0.867
K-Nearest Neighbors	2.515	0.057	1.969	0.642	1.492	0.795
CatBoost	2.194	0.282	1.468	0.770	1.456	0.774
XGBoost	2.194	0.282	**1.468**	**0.770**	**1.269**	**0.851**

both the FRIA and ALTAI evaluation and also draw practical recommendations that the AI-based system should consider when deployed in real-world applications. This application of the ALTAI checklist fosters semantic alignment, raises awareness, and promotes interdisciplinary training, thereby influencing the potential standardization of skills and competencies necessary for AI compliance processes. The checklist serves as a valuable self-assessment tool during the system's design phase, highlighting commonly overlooked weak points by AI developers. However, it is important to note that the answers and practical suggestions must be accompanied by a set of good practices, which may initially appear as barriers to the unrestricted development of AI technology. We argue that adopting a critical approach to AI development is a societal goal, and we firmly believe that it can evolve and accelerate over time without impeding the progress and openness towards new technologies.

6.1. Fundamental rights (FRIA)

Both students and teachers have fundamental rights at stake, including dignity, data protection for students, and work-life for teachers. Our proposed system upholds these rights and aims to assist students throughout their academic journey. However, it is essential for teachers to use the system in an informed manner to prioritize the best interests of the children. To ensure this, both teachers and students should receive comprehensive training and information about the AI system, including clear *Terms & Conditions* statements and dedicated training sessions. Notably, during the initial phase of the AI system development, we have intentionally chosen not to implement bias mitigation techniques in the data processing phase, but instead adopted XAI techniques to spot unfair behaviors [96]. This was done with the aim of disclosing potential data bias, which is a recognized indicator of historical societal inequalities. This decision aligns with the overall framework, which emphasizes the importance of fostering critical thinking skills among educators. By being aware of potential discriminatory factors, teachers can gain a deeper understanding and sensitivity towards these issues, thereby enabling them to identify and address any inherent biases and take preventive measures. However, in later phases of the AI system's life cycle, the inclusion of bias mitigation techniques may become necessary if mandated by school solicitations or if supported by evidence of biased behavior.

6.2. Human agency and oversight

AI4SPP is specifically designed to assist (but not *replace*) teachers in grading students, a task that has a significant impact on the lives of children. Therefore, it is crucial to address both excessive reliance and complete distrust in the AI system through technical and organizational measures. The design of the system is intended to promote critical thinking among teachers during its utilization. Alongside providing prior training to users, disclaimers are included to describe the quality of the output generated by the system. These may include the accuracy of the predictive model, or the ranking of the available features with respect to their predictive importance. The role of human empowerment is crucial in ensuring proper usage of the system. In AI4SPP the machine's self-learning process is supplemented by additional explanations of the model's outcomes, which enhance trust calibration and enable more informed human oversight. The appointment of a designated external person or body charged with monitoring and reporting teachers' feedback to the development team can make a substantial contribution to the continuous enhancement of the AI system. This figure could also engage with students to gather their opinions on the use of the AI4SPP tool. Additionally, implementing a peer evaluation system among teachers could help identify potential issues or areas of concern. Both teachers and the external representative could serve as a *"human stop-button"* to address any problematic usage of the system.

6.3. Technical robustness and safety

With respect to safety to ensure the protection of sensitive data throughout the different phases of our system's life cycle, compliance with the EU Cyber-Security Act [121] is essential, regardless of where the data is stored (e.g., internally within the school or with the AI developers). Before launching any development, it is imperative for the school and the development team to engage in discussions and reach agreement on the implementation of data protection-by-design and data protection-by-default safeguards (GDPR Article 25, [122]). In this regard, it is crucial to prioritize the analysis of security measures, as privacy cannot be ensured without adequate security measures in place.

In light of the potential consequences of adopting AI recommendations generated via underperforming models, we enhanced the technical robustness of AI4SPP by giving the end-user the possibility of further analyzing the performance of each single prediction, both in terms of predictive probabilities (i.e., *what was the probability assigned by the model to the predicted outcome?*), and of local explanation (i.e., *which attributes mostly determined the predicted outcome?*). Furthermore, in order to assess the extent to which our system can be applied to different contexts, we conducted grade predictions for a discipline (specifically, mathematics) that was distinct from the one utilized during training (namely, Portuguese language). Our findings indicated a notable reduction in predictive accuracy, which suggests that the generalizability of AI4SPP may be limited. Through the promotion of

Table 3
Performance of the final model on the test dataset of mathematics class (binary and multiclass classification tasks).

Task	Period	Accuracy	F_1-score (macro)	F_1-score (weighted)	ROC-AUC score (macro)
Binary classification	G3	0.82	0.78	0.81	0.90
Multiclass classification	G1	0.27	0.27	0.29	0.62
Multiclass classification	G2	0.46	0.35	0.47	0.84
Multiclass classification	G3	0.53	0.48	0.52	0.91

Table 4
Performance of the final model on the test dataset of mathematics class (regression task).

Task	Period	RMSE	R^2
Regression	G1	2.96	0.20
Regression	G2	2.08	0.69
Regression	G3	2.14	0.78

open science principles, we have made the source code of our Proof-of-Concept publicly available. Moreover, we guarantee reproducibility and transparency by providing comprehensive documentation and version management using Git. The source code can be found at: https://github.com/andreafedele/student-performance.

Finally, we note that improving data quality over time is crucial. In the initial release, historical data from students within the same school can be used for training. However, subsequent system updates should incorporate real-time data from enrolled students, which can be given higher importance through appropriate fine-tuning of model parameters. Deployment processes should include fallback plans, such as distributed recurrent backups of the AI system and its state. Furthermore, comprehensive testing should be conducted to ensure accurate communication of the system's performance to end-users.

6.4. Privacy and data governance

The AI-assessment concerning privacy and data governance encompasses all activities related to GDPR compliance. For detailed information on this topic, please refer to paragraph 3.3. In our specific context, we classify the school as the *data controller*, the AI-developers as the *data processor*, and the students as the *data subjects*. The legal basis for data processing through the tool is obtained through consent from the individuals responsible for each student (GDPR Article 6(1)(a)). It is important to pay a particular attention to special categories of personal data (GDPR Article 9(1)), for which specific consent should be obtained. Our AI system functions as an automated decision-making system governed by Article 22 of GDPR, thus posing a high risk to the rights and freedoms of individuals. For this reason, a comprehensive data protection impact assessment should be conducted in accordance with GDPR Article 35, and the school should designate a Data Protection Officer (DPO) as required by GDPR Article 37. It is crucial to implement technical and organizational measures that enable individuals to exercise their rights once they have given consent, including the right to withdraw consent (GDPR Article 7(3)), the right to object (GDPR Article 21), and the right to be forgotten (GDPR Article 17). These training activities can be conducted annually or more frequently, potentially by the same external representative mentioned earlier in the discussion about *Human Agency and Oversight*. The data controller should also implement appropriate technical and organizational measures, such as encryption, pseudonymization, aggregation, anonymization, and minimization, to achieve data protection by design and default (GDPR Article 25). These measures are essential to prevent re-identification of data subjects. For instance, differential privacy algorithms can be utilized to facilitate secure information sharing without disclosing private information (e.g., sensitive attributes) of individuals whose data is stored in the database [123]. The data controller has access to non-protected data and is responsible for performing the aforementioned activities, as well as communicating the resulting protected data to the data processor. The data processor processes the protected data using the AI solution and communicates back the (protected) results to the data controller. It is important to note that the data controller is the only entity holding the encrypted function that maps the outcomes of the AI system for each individual to the corresponding real data subjects. Additionally, the data controller and data processor have access to separate servers and data repositories.

6.5. Transparency

During the pre-processing phase in the training stage, the input data is assessed, which involves quantifying missing values and detecting erroneous, incorrect, or inaccurate data in the specified format. It is crucial to emphasize that this pre-processing stage must be carried out during each system update, usually on an annual basis, when new data is utilized for additional refinement steps. During inference time, data quality is also assessed using standard procedures, ensuring that prediction scores fall within expected ranges. Logging procedures are implemented to record the recommendations provided by the AI system, while explainable AI techniques enable traceability of the logical steps taken by the AI system. Our AI system offers end users the ability to access both the output probability scores and explanations for each grade prediction. To encourage critical usage, users can choose the level of information they wish to obtain from the AI system, ranging from output alone to output with confidence scores and explanations.

Prior to using the AI system, it is crucial for end-users to undergo training that enables them to understand the recommendations provided. Users should be provided with appropriate technical documentation, training materials and disclaimers, ensuring they are well-prepared to use the AI system effectively. It is recommended to conduct follow-up sessions periodically to gather feedback from teachers regarding their preferences for different types of outputs, especially with respect to the quality of explanations. This feedback can contribute to ongoing research efforts aimed at assessing the significance of explanations in hybrid decision-making processes. Additionally, the feedback obtained from these follow-up training sessions can be used to enhance the quality of disclaimers and explanations, both in terms of their content and graphical interface.

6.6. Diversity, non-discrimination, and fairness

Continuous assessment and monitoring of sensitive features within the dataset should be conducted at various stages throughout the lifecycle of the AI system. This analysis should not be limited to the initial stages but should be performed regularly, such as on an annual basis during the fine-tuning process. By conducting descriptive analysis of sensitive features, potential biases or discriminatory patterns can be identified and addressed, ensuring fairness in the AI system's outcomes. This ongoing monitoring helps maintain accountability and enables necessary adjustments to be made to mitigate any potential adverse impacts on individuals or specific groups. It promotes transparency and ensures that the AI system operates in a manner that upholds ethical standards and respects the rights of the individuals involved.

During the preliminary development stage, bias mitigation techniques are not initially implemented. However, as already mentioned in the discussion about *Human Agency and Oversight*, this "flaw" can serve two important purposes: (1) it can bring attention to existing biases and contribute to societal change by highlighting the need for

fairness, and (2) it allows for the monitoring of potential algorithmic biases that may arise as a result of incorporating teacher scores in subsequent versions of the system. If the system reveals significant data or algorithmic biases, appropriate measures for bias mitigation should be implemented. Explanations provided by the system can serve as a useful tool for identifying issues related to bias, discrimination, or poor performance. Additionally, it is essential to conduct further comparisons to evaluate the significance of sensitive features in model performance. If certain features do not contribute significantly to overall performance, they should be abandoned or no longer collected, thereby reducing the risk of biased outcomes. Lastly, prior to finalizing the design of the AI system, conducting a survey to understand the current status of school systems and gather insights about special needs from potential end-users can ensure that the AI system is accessible to all without discrimination. This survey helps promote inclusivity and fairness in system implementation.

6.7. Societal and environmental well-being

AI tools can be limited by the substantial hardware and computational power required, particularly during the training process of deep learning models. Additional expenses such as GPUs, cloud services, and hardware may arise as a result. Despite these limitations, utilizing a tool with these characteristics can yield significant beneficial results for society. It aids in the detection of patterns and risk indicators, facilitating personalized interventions and fostering improved teacher–student relationships. By enabling proactive interventions for students at risk of school failures and increasing awareness of factors influencing low performance, the tool contributes to overall educational enhancement. Its implementation also encourages introspection within the teaching community, leading to the adoption of measures to address biases within the school environment. However, caution must be exercised to avoid over-reliance on AI systems and potential human deskilling. Prioritizing reproducibility, transparency, and explainability in our system represents the direction in which AI development should strive.

In order to mitigate energy consumption without significantly compromising performance, specific technical methods can be implemented. In our case study, extending the training phase without relying on GPUs is unlikely to impact the requirements of the school. Indeed, given the tabular nature of the data and the tool's usage, teachers should just anticipate slight delays when querying the AI system on traditional CPUs.

6.8. Accountability

The implemented measures of reproducibility, traceability, and explainability in our AI system enhance its auditability. These design features should be upheld by the data controller and made available to external auditors upon request. Additionally, developers can utilize these auditing techniques as debugging tools to enhance the performance and quality of the AI system. The data controller should establish an AI ethics review board within the school, consisting of referees with expertise in technology and AI ethics. This board is responsible for monitoring and evaluating the AI system's compliance with ALTAI principles. It should also engage in ongoing discussions regarding accountability and ethical practices, and serve as a point of contact for external parties, including whistle-blowers, NGOs, and trade unions, who raise valid concerns about the AI system. If adverse effects on data subjects are reported, the AI ethics review board must promptly notify the data controller and data processor, urging them to rectify the system's functioning. Simultaneously, the board should inform end users about the issue to prevent its recurrence and protect other data subjects. Regular communication between the data processor and controller is crucial for assessing the AI system's behavior regarding risk management and ensuring vulnerability protection. End users should receive comprehensive training on risk awareness before deploying the AI system, and follow-up training sessions should address potential negative impacts and the relevant legal framework.

7. Conclusive remarks

The present work utilized the ALTAI checklist to assess an AI application designed for the educational domain. Specifically, the tool in question was intended to facilitate the timely detection of learning failure in schools, with the ultimate goal of implementing appropriate interventions. Our study adopted an interdisciplinary method, emphasizing the significance of a mixed knowledge background for the development of human-centered AI. We advocate maintaining this approach throughout the entire life cycle of such AI applications, particularly when operating in sensitive contexts. While conducting the ALTAI self-assessment, a number of technical and organizational measures emerged as being essential factors in the development of an AI system that manages data pertaining to vulnerable subjects. For instance, it is crucial to correctly identify the roles of the data controller and data processor, not only for GDPR accountability but also for technical reasons. Defining who will collect, maintain, and anonymize the data must be specified prior to the development of the AI tool. We propose assigning these tasks to the school or a representative after appropriate training. Additionally, AI developers and data processors should work with anonymized data on dedicated servers separate from the school's infrastructure. We recommend establishing a school-internal ethical review board to monitor the AI system's compliance with ALTAI principles throughout its life cycle. This board should foster critical thinking among teachers and students regarding accountability and ethical practices, and serve as a point of contact for external auditors who can provide feedback to the development team. We emphasize the crucial role of teachers' training in ensuring the success of the platform. Teachers should engage in critical thinking before and during system usage. Clear and easily understandable explanations of model decisions and model output confidence are essential, with opportunities for improvement based on feedback from external representatives and the development team. In conclusion, it is important to clarify that the aim of this study was not to suggest the most optimal AI tool as a readily available product, but rather to develop and analyze a Proof-of-Concept of an AI-driven solution that may potentially be implemented in a vulnerable setting.

At the time of writing, the AI Act has not yet been adopted, so our goal was to offer a methodology for creating an AI system operating in a vulnerable context that can highlight and subsequently meet the ethical and legal requirements. To this end, our study made use of the ALTAI checklist, which, with its seven requirements, largely formed the basis for the drafting of the AI Act, thus attempting to mitigate the risks and serious concerns about the use of AI in education (i.e., privacy, surveillance, autonomy, bias, and discrimination). According to the AI Act, in order to be placed on the EU market, a high-risk system, such as one used in education, must comply with a whole series of obligations that indirectly echo the ALTAI ethical guidelines. For instance, for high-risk AI systems the *Fundamental Rights Impact Assessment* was introduced by the European Parliament at Article 29a; *Human Agency* and *Human Oversight* are laid down in Article 14; *Technical Robustness and Safety* are provided for in Article 15; *Privacy and Data Governance* are mainly prescribed in Article 10; *Transparency* (also to achieve *Diversity, Non-Discrimination and Fairness*) is laid down in Articles 11, 12, 13; *Social and Environmental Well-being* is reported in numerous Recital (especially after the parliamentary amendments). Finally, *Accountability* is codified in Article 9 which – together with the *Transparency* provisions mentioned – at most constitutes a *fil rouge* that binds all the requirements of Chapter 2. Therefore, once the AI Act is passed and enters into force developers and deployers will mainly follow the requirements therein, and the ALTAI checklist will at most constitute an auxiliary interdisciplinary ethical-legal support methodology capable of complementing regulatory obligations and the interpretation given by national and European courts.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

I have shared the link to the paper code at the "Attach file" step.

Acknowledgments

We sincerely thank Dr. Denise Amram (LIDER-Lab, DIRPOLIS Institute, Scuola Superiore Sant'Anna, Pisa, Italy) for her invaluable guidance and insightful discussions.

The research was carried out with the contribution of the research groups linked to the European Union under ERC-2018-ADG GA 834756 (XAI), the HumanE-AI-Net GA 952026, the Partnership Extended PE00000013 - "FAIR - Future Artificial Intelligence Research" - Spoke 1 "Human-centered AI", and the "SoBigData RI Preparatory Phase Project" GA 10107904. It has been realised also thanks to the computational resources of "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (http://www.sobigdata.eu) GA 871042, and NextGenerationEU - National Recovery and Resilience Plan (PNRR) - Project: "SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics" - Prot. IR0000013 - Notice n. 3264 of 12/28/2021.

References

[1] European Union: European Commission. Proposal for a regulation of the European Parliament and of the Council: Laying down harmonized rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts. 2021, cOM(2021) 206 final [Accessed 4 March 2023]. URL https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021PC0206.

[2] Jones K. AI governance and human rights: Resetting the relationship. 2023, http://dx.doi.org/10.55317/9781784135492.

[3] Soomro TA, Zheng L, Afifi AJ, Ali A, Yin M, Gao J. Artificial intelligence (ai) for medical imaging to combat coronavirus disease (covid-19): A detailed review with direction for future research. Artif Intell Rev 2021;55(2):1409–39. http://dx.doi.org/10.1007/s10462-021-09985-z.

[4] EPR. Artificial intelligence and service provision for people with disabilities. Tech. rep., European Platform for Rehabilitation (EPR); 2020.

[5] Rolnick D, Donti PL, Kaack LH, Kochanski K, Lacoste A, Sankaran K, et al. Tackling climate change with machine learning. ACM Comput Surv 2022;55(2):1–96. http://dx.doi.org/10.1145/3485128.

[6] Cline T. Making the most of machine learning on farm, SPORE. 2019.

[7] Beduschi A. Harnessing the potential of artificial intelligence for humanitarian action: Opportunities and risks. Int Rev Red Cross 2022;104(919):1149–69.

[8] UNESCO. Beijing consensus on artificial intelligence and education. 2019, available at: https://unesdoc.unesco.org/ark:/48223/pf0000368303. [Accessed 5 March 2023].

[9] Pedró F, Subosa M, Rivas A, Valverde P. Artificial intelligence in education: Challenges and opportunities for sustainable development. UNESCO; 2019, eD-2019/WS/8, available at: https://unesdoc.unesco.org/ark:/48223/pf0000366994. [Accessed 5 March 2023].

[10] Miao F, Holmes W, Huang R, Zhang H. AI and education: Guidance for policy-makers. UNESCO; 2021, available at: https://unesdoc.unesco.org/ark:/48223/pf0000376709. [Accessed 5 March 2023].

[11] UNESCO. Recommendation on the ethics of artificial intelligence. 2022, sHS/BIO/PI/2021/1, available at: https://unesdoc.unesco.org/ark:/48223/pf0000386693. [Accessed 29 December 2023].

[12] Emma S, Arianna V. Guidance for generative AI in education and research. UNESCO; 2023, available at: https://unesdoc.unesco.org/ark:/48223/pf0000386693. [Accessed 29 December 2023].

[13] Emma S, Arianna V. Chatgpt and artificial intelligence in higher education: quick start guide. 2023, eD/HE/IESALC/IP/2023/12, available at: https://unesdoc.unesco.org/ark:/48223/pf0000385146. [Accessed 29 December 2023].

[14] Niestadt M, Debyser A, Scordamaglia D, Pape M. Artificial intelligence in transport. 2019.

[15] Fergusson G, Fitzgerald C, Frascella C, Iorio M, McBrien T, Schroeder C, Winters B, Zhou E. Generating harms: Generative ai's impact & paths forward. Tech. rep., Electronic Privacy Information Center (epic.org); 2023.

[16] Malgieri G, Pasquale F. Licensing high-risk artificial intelligence: Toward ex ante justification for a disruptive technology. Comput Law Secur Rev 2024;52:105899. http://dx.doi.org/10.1016/j.clsr.2023.105899.

[17] O'Neil C. Weapons of math destruction: How big data increases inequality and threatens democracy. USA: Crown Publishing Group; 2016.

[18] Broussard M. Artificial unintelligence. London, England: The MIT Press; 2019.

[19] Lashbrook A. Ai-driven dermatology could leave dark-skinned patients behind. 2018.

[20] Marcus G, Davis E. Rebooting AI. New York, NY: Ballantine Books; 2019.

[21] Topol E. Deep medicine: How artificial intelligence can make healthcare human again. London, England: Basic Books; 2019.

[22] European Commission and Directorate-General for Communications Networks, Content and Technology. Ethics guidelines for trustworthy AI. Publications Office of the European Union; 2019, URL https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[23] Chatila R, Dignum V, Fisher M, Giannotti F, Morik K, Russell S, Yeung K. Trustworthy AI. Springer International Publishing; 2021, p. 13–39. http://dx.doi.org/10.1007/978-3-030-69128-8_2.

[24] European Union European Commission. Communication from the Commission to the European Parliament, the European Council, The Council, the European Economic and Social Committee and the committee of the regions: Artificial intelligence for Europe. 2018, cOM(2018) 237 final, available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018DC0237. [Accessed 4 March 2023].

[25] European Union European Commission. Communication from the commission to the European Parliament, the Council, the European Economic and Social Committee and the committee of the regions: Building trust in human-centric artificial intelligence. 2019, cOM(2019) 168 final, available at: https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX%3A52019DC0168. [Accessed 29 December 2023].

[26] European Union European Commission. White paper: On artificial intelligence - A European approach to excellence and trust. 2020, cOM(2020) 65 final, available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0065, [Accessed 29 December 2023].

[27] European Commission and Directorate-General for Communications Networks, Content and Technology. The assessment list for trustworthy artificial intelligence (ALTAI) for self assessment. Publications Office of the European Union; 2020, URL https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment.

[28] Institute of Ethical AI in Education. Interim report: towards a shared vision of ethical ai in education. Tech. rep., University of Buckingham; 2020, URL https://www.buckingham.ac.uk/wp-content/uploads/2020/02/The-Institute-for-Ethical-AI-in-Educations-Interim-Report-Towards-a-Shared-Vision-of-Ethical-AI-in-Education.pdf.

[29] Akgun S, Greenhow C. Artificial intelligence in education: Addressing ethical challenges in k-12 settings. AI Ethics 2021;2(3):431–40. http://dx.doi.org/10.1007/s43681-021-00096-7.

[30] Herodotou C, Rienties B, Boroowa A, Zdrahal Z, Hlosta M, Naydenova G. Implementing predictive learning analytics on a large scale. In: Proceedings of the seventh international learning analytics & knowledge conference. ACM; 2017, http://dx.doi.org/10.1145/3027385.3027397.

[31] European Commission. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (general data protection regulation) (text with EEA relevance). 2016, URL https://eur-lex.europa.eu/eli/reg/2016/679/oj.

[32] Hutson M. Rules to keep ai in check: Nations carve different paths for tech regulation. Nature NEWS FEATURE 2023.

[33] The White House Office of Science and Technology Policy. The blueprint for an ai bill of rights: Making automated systems work for the American people. 2022, available at https://www.whitehouse.gov/ostp/ai-bill-of-rights/.

[34] The US Government Publishing Office. Artificial intelligence training for the acquisition workforce act, public law 117-207 (10/17/2022), 117th congress public law 207. 2022, available at https://www.congress.gov/bill/117th-congress/senate-bill/2551/text.

[35] Tabassi E. Artificial intelligence risk management framework (ai rmf 1.0) (2023-01-26 05:01:00 2023). http://dx.doi.org/10.6028/NIST.AI.100-1. URL https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=936225.

[36] The White House Presidential Actions. Executive order on further advancing racial equity and support for underserved communities through the federal government. 2023, available at https://www.whitehouse.gov/briefing-room/presidentialactions/2023/02/16/executive-order-on-further-advancing-racial-equity-and-support-for-underserved-communities-through-the-federal-government/.

[37] The White House Statements and Releases. Fact sheet: President biden issues executive order on safe, secure, and trustworthy artificial intelligence. 2023, available at https://www.whitehouse.gov/ostp/ai-bill-of-rights.

[38] Senate, H. of representatives of the United States of America in congress, algorithmic accountability act of 2022, s.3572 — 117th congress (2021–2022). 2022, available at: https://www.congress.gov/bill/117th-congress/senate-bill/3572/text.

[39] Mökander J, Juneja P, Watson DS, Floridi L. The US algorithmic accountability act of 2022 vs. the EU artificial intelligence act: What can they learn from each other? Minds Mach 2022;32(4):751–8. http://dx.doi.org/10.1007/S11023-022-09612-Y.

[40] Sheehan M. China's ai regulations and how they get made, carnegie endowment for international piece. 2023, URL https://carnegieendowment.org/files/202307-Sheehan_Chinese%20AI%20gov.pdf. [Accessed 2 August 2003].

[41] European Union European Commission. Communication from the commission to the European Parliament, the council, the European Economic and Social Committee and the committee of the regions: Digital education action plan 2021–2027. resetting education and training for the digital age. 2020, cOM(2020) 624 final, available at: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0624. [Accessed 4 March 2023].

[42] European Commission and Directorate-General for Education, Youth, Sport and Culture. Ethical guidelines on the use of artificial intelligence (AI) and data in teaching and learning for educators. Publications Office of the European Union; 2022, URL https://op.europa.eu/en/publication-detail/-/publication/d81a0d54-5348-11ed-92ed-01aa75ed71a1/language-en.

[43] Zawacki-Richter O, Marín VI, Bond M, Gouverneur F. Systematic review of research on artificial intelligence applications in higher education - Where are the educators? Int J Educ Technol Higher Educ 2019;16(1). http://dx.doi.org/10.1186/s41239-019-0171-0.

[44] Dignum V, Pigmans K, Vosloo S, Penagos M. Policy guidance on AI for children, United Nations Children's Fund. UNICEF, 2021, available at: https://www.unicef.org/globalinsight/media/2356/file. [Accessed 5 March 2023].

[45] Assembly UG. Transforming our world: the 2030 agenda for sustainable development, a/RES/70/1. 2015, available at: https://www.refworld.org/docid/57b6e3e44.html. [Accessed 5 March 2023].

[46] Amram D. Children (in the digital environment). In: Comande G, editor. Elgar encyclopedia of law and data science. Cheltenham, England: Edward Elgar Publishing; 2022.

[47] Isaacs S. Adolescent perspectives on artificial intelligence. A report on consultations with adolescents across the world. United Nations Children's Fund (UNICEF; 2021, available at: https://www.unicef.org/globalinsight/sites/unicef.org.globalinsight/files/2021-02/UNICEF_AI_AdolescentPerspectives_20210222.pdf. [Accessed 5 March 2023].

[48] UNESCO. International conference on artificial intelligence and education, planning education in the ai era: Lead the leap, ed/PLS/ict/2019/13. 2019, available at: https://unesdoc.unesco.org/ark:/48223/pf0000370967. [Accessed 5 March 2023].

[49] Baker T, Smith L, Anissa N. Educ-ai-tion rebooted? Exploring the future of artificial intelligence in schools and colleges. Tech. rep., NESTA; 2019, URL https://www.nesta.org.uk/report/education-rebooted.

[50] Lanza C, Francq S, Fallon M. Vulnerability and ai-based technologies: European protection of vulnerable consumers in the digital market. 2023.

[51] Agencia Española de Protección de Datos. List of the types of data processing that require a data protection impact assessment under art 35.4. English version available avaiable at: https://www.aepd.es/documento/listas-dpia-en-35-4.pdf.

[52] Malgieri G, Fuster GG. The vulnerable data subject: A gendered data subject? Eur J Law Technol 2022;13(2). URL https://ejlt.org/index.php/ejlt/article/view/843.

[53] Edwards L. Regulating ai in Europe: Four problems and four solutions. 2022, Retrieved March 15 2022.

[54] Malgieri G. Vulnerability and data protection law. Oxford University Press; 2023.

[55] European Parliament and Council. Directive 2005/29/ec of 11 May 2005 concerning unfair business-to-consumer commercial practices in the internal market (unfair commercial practices directive). 2005, avaiable at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32005L0029.

[56] European Parliament and Council. Directive (eu) 2021/2118 of the European Parliament and of the Council of 24 November 2021 amending directive 2009/103/ec relating to insurance against civil liability in respect of the use of motor vehicles, and the enforcement of the obligation to insure against such liability. 2021, avaiable at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32021L2118.

[57] European Parliament and Council. Regulation regulation 536/2014 of 16 April 2014 on clinical trials on medicinal products of human use. 2014, avaiable at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32014R0536.

[58] Comandé G, editor. Elgar encyclopedia of law and data science. Cheltenham, England: Edward Elgar Publishing; 2022.

[59] Edwards L, Veale M. Slave to the algorithm? Why a right to explanationn is probably not the remedy you are looking for. Duke Law Technol Rev 2017;16(1). http://dx.doi.org/10.2139/ssrn.2972855.

[60] Pasquale F. Introduction: The need to know. Harvard University Press; 2015, p. 1–18, URL http://www.jstor.org/stable/j.ctt13x0hch.3.

[61] European Commission (DGResearch and Innovation). Ethics by design and ethics of use approaches for artificial intelligence. 2021, URL https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf.

[62] Ippolito F, Sanchez S Iglesias, editors. Protecting vulnerable groups. Modern studies in European law, Oxford, England: Hart Publishing; 2015.

[63] TURNER BS. vulnerability and human rights. Penn State University Press; 2015, http://dx.doi.org/10.5325/j.ctt7v124.

[64] Dudgeon v. United Kingdom, no. Appl. No. 7525/76, Council of Europe: European court of human rights. 1981, available at: https://www.refworld.org/cases,ECHR,47fdfaf7d.html. [Accessed 29 May 2023].

[65] Malgieri G, Niklas J. Vulnerable data subjects. Comput Law Secur Rev 2020;37:105415. http://dx.doi.org/10.1016/j.clsr.2020.105415, URL https://www.sciencedirect.com/science/article/pii/S0267364920300200.

[66] Peroni L, Timmer A. Vulnerable groups: The promise of an emerging concept in European human rights convention law. Int J Const Law 2013;11(4):1056–85. http://dx.doi.org/10.1093/icon/mot042.

[67] Gennet É, Andorno R, Elger B. Does the new EU regulation on clinical trials adequately protect vulnerable research participants? Health Policy 2015;119(7):925–31. http://dx.doi.org/10.1016/j.healthpol.2015.04.007.

[68] Article 29 Data Protection Working Party. Opinion 06/2014 on the notion of legitimate interests of the data controller under article 7 of directive 95/46/ec, 844/14/EN WP 217. 2014, available at: https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf. [Accessed 29 May 2023].

[69] Supervisor European Data Protection. Edps opinion on the European Commission's white paper on artificial intelligence – A European approach to excellence and trust, opinion 4/2020. 2020, available at: https://edps.europa.eu/data-protection/our-work/publications/opinions/edps-opinion-european-commissions-white-paper_en. [Accessed 03 January 2024].

[70] Jasmontaite L, Kamara I, Zanfir-Fortuna G, Leucci S. Data protection by design and by default:. Eur Data Prot Law Rev 2018;4(2):168–89. http://dx.doi.org/10.21552/edpl/2018/2/7.

[71] Kaminski ME, Malgieri G. Algorithmic impact assessments under the GDPR: Producing multi-layered explanations. Int Data Privacy Law 2020;11(2):125–44. http://dx.doi.org/10.1093/idpl/ipaa020.

[72] Kaminski ME. The right to explanation. explained 2019;34(1). http://dx.doi.org/10.15779/Z38TD9N83H.

[73] Malgieri G, Comandé G. Why a right to legibility of automated decision-making exists in the general data protection regulation. Int Data Privacy Law 2017;7(4):243–65. http://dx.doi.org/10.1093/idpl/ipx019.

[74] Selbst AD, Powles J. Meaningful information and the right to explanation. Int Data Privacy Law 2017;7(4):233–42. http://dx.doi.org/10.1093/idpl/ipx022.

[75] Wachter S, Mittelstadt B, Floridi L. Why a right to explanation of automated decision-making does not exist in the general data protection regulation. Int Data Privacy Law 2017;7(2):76–99. http://dx.doi.org/10.1093/idpl/ipx005.

[76] Article 29 Data Protection Working Party. Guidelines on data protection impact assessment (dpia) and determining whether processing is likely to result in a high risk for the purposes of regulation 2016/679. 2017, 17/EN WP 248 rev.01, available at: https://ec.europa.eu/newsroom/article29/items/611236.

[77] Kaminski ME. Binary governance: Lessons from the gdpr's approach to algorithmic accountability. S Cal L Rev 2018;92:1529.

[78] Wachter S, Mittelstadt B. A right to reasonable inferences: Re-thinking data protection law in the age of big data and ai. Columbia Bus Law Rev 2019;494.

[79] JUDGMENT OF THE COURT (First Chamber). Reference for a preliminary ruling – protection of natural persons with regard to the processing of personal data – regulation (eu) 2016/679 – article 22 – automated individual decision-making – credit information agencies – automated establishment of a probability value concerning the ability of a person to meet payment commitments in the future ('scoring') – use of that probability value by third parties, case C-634/21, document ECLI:EU:C:2023:957. 2023, ECLI:EU:C:2023:957. URL https://curia.europa.eu/juris/document/document.jsf?text=&docid=280426&pageIndex=0&doclang=EN&mode=lst&dir=&occ=first&part=1&cid=11861879.

[80] Malgieri G. Automated decision-making in the eu member states: The right to explanation and other suitable safeguards in the national legislations. Comput Law Secur Rev 2019;35(5):105327. http://dx.doi.org/10.1016/j.clsr.2019.05.002, URL https://www.sciencedirect.com/science/article/pii/S0267364918303753.

[81] Kaminski ME, Urban JM. The right to contest ai. Columbia Law Rev 2021;121(7):1957–2048.

[82] Rockwell G, Black E, Selinger E, Davola A, Seide E, Gulson K. From shortcut to sleight of hand: Why the checklist approach in the eu guidelines does not work. 2019, URL https://escholarship.org/uc/item/12s9x39n.

[83] European Commission (DGResearch and Innovation). Charter of fundamental rights of the european union, document 12012p/TXT. 2000, URL http://data.europa.eu/eli/treaty/char_2012/oj.

[84] Ciacchi AC, Brüggemeier G, Comandé G, editors. Fundamental rights and private law in the European union: volume 2, comparative analyses of selected case patterns. Cambridge, England: Cambridge University Press; 2010.

[85] Scherer MU. Regulating artificial intelligence systems: Risks, challenges, competencies, and strategies. Harv JL Tech 2015;29:353.
[86] Amram D, Cignoni A, Banfi T, Ciuti G. From p4 medicine to p5 medicine: transitional times for a more human-centric approach to AI-based tools for hospitals of tomorrow. Open Res Europe 2022;2:33. http://dx.doi.org/10.12688/openreseurope.14524.1.
[87] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Artif Intell 2019;267:1–38. http://dx.doi.org/10.1016/j.artint.2018.07.007.
[88] Hacker P, Passoth J-H. Varieties of ai explanations under the law. From the gdpr to the aia, and beyond. In: XxAI-beyond explainable AI: international workshop, held in conjunction with ICML 2020, July 18 2020, Vienna, Austria, revised and extended papers. Springer; 2022, p. 343–73.
[89] La Diega GN. Against the dehumanisation of decision-making. J Intell Prop Info Tech Elec Com L 2018;9:3.
[90] Perel M, Elkin-Koren N. Black box tinkering: Beyond disclosure in algorithmic enforcement. Fla L Rev 2017;69:181.
[91] Kaminski ME. Understanding transparency in algorithmic accountability. In: Barfield W, editor. The Cambridge handbook of the law of algorithms, Cambridge law handbooks. Cambridge University Press; 2020, p. 121–38. http://dx.doi.org/10.1017/{9781108680844}.006.
[92] Wachter S, Mittelstadt B, Russell C. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. Harv JL Tech 2017;31:841.
[93] Loi M, Ferrario A, Viganò E. Transparency as design publicity: Explaining and justifying inscrutable algorithms. Ethics Inf Technol 2020;23(3):253–63. http://dx.doi.org/10.1007/s10676-020-09564-w.
[94] Hacker P. A legal framework for AI training data—from first principles to the artificial intelligence act. Law, Innov Technol 2021;13(2):257–301. http://dx.doi.org/10.1080/17579961.2021.1977219.
[95] Kroll JA. Outlining traceability. In: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency. ACM; 2021, http://dx.doi.org/10.1145/3442188.3445937.
[96] Guidotti R, Monreale A, Giannotti F, Pedreschi D, Ruggieri S, Turini F. Factual and counterfactual explanations for black box decision making. IEEE Intell Syst 2019;34(6):14–23.
[97] Comandé G. Multilayered (accountable) liability for artificial intelligence. Nomos Verlagsgesellschaft mbH; 2019, p. 165–84.
[98] Novelli C, Taddeo M, Floridi L. Accountability in artificial intelligence: What it is and how it works. AI & SOCIETY; 2023, http://dx.doi.org/10.1007/s00146-023-01635-y.
[99] Lepri B, Oliver N, Letouzé E, Pentland A, Vinck P. Fair, transparent, and accountable algorithmic decision-making processes. Philos Technol 2017;31(4):611–27. http://dx.doi.org/10.1007/s13347-017-0279-x.
[100] Vedder A, Naudts L. Accountability for the use of algorithms in a big data environment. Int Rev Law, Comput Technol 2017;31(2):206–24. http://dx.doi.org/10.1080/13600869.2017.1298547.
[101] Kroll JA. Accountable algorithms (Ph.D. thesis), Princeton University; 2015.
[102] Cobbe J, Veale M, Singh J. Understanding accountability in algorithmic supply chains. 2023, arXiv preprint arXiv:2304.14749.
[103] Malgieri G, Pasquale FA. From transparency to justification: Toward ex ante accountability for AI. SSRN Electron J 2022. http://dx.doi.org/10.2139/ssrn.4099657.
[104] Williams R, Cloete R, Cobbe J, Cottrill C, Edwards P, Markovic M, Naja I, Ryan F, Singh J, Pang W, et al. From transparency to accountability of intelligent systems: Moving beyond aspirations. Data Policy 2022;4:e7. http://dx.doi.org/10.1017/dap.2021.37.
[105] Raji ID, Smart A, White RN, Mitchell M, Gebru T, Hutchinson B, Smith-Loud J, Theron D, Barnes P. Closing the AI accountability gap. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. ACM; 2020, http://dx.doi.org/10.1145/3351095.3372873.
[106] Ehsan U, Liao QV, Muller M, Riedl MO, Weisz JD. Expanding explainability: Towards social transparency in AI systems. In: Proceedings of the 2021 CHI conference on human factors in computing systems. ACM; 2021, http://dx.doi.org/10.1145/3411764.3445188.
[107] Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D. A survey of methods for explaining black box models. ACM Comput Surv 2018;51(5):1–42. http://dx.doi.org/10.1145/3236009.
[108] Cortez P, Silva AMG. Using data mining to predict secondary school student performance. 2008.
[109] Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. Smote: Synthetic minority over-sampling technique. J Artif Int Res 2002;16(1):321–57.
[110] Cortes C, Vapnik V. Support-vector networks. Mach Learn 1995;20(3):273–97.
[111] Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. Catboost: Unbiased boosting with categorical features. In: Proceedings of the 32nd international conference on neural information processing systems. NIPS'18, Red Hook, NY, USA: Curran Associates Inc.; 2018, p. 6639–49.
[112] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. KDD '16, New York, NY, USA: Association for Computing Machinery; 2016, p. 785–94. http://dx.doi.org/10.1145/2939672.2939785.
[113] Drucker H, Burges CJC, Kaufman L, Smola A, Vapnik V. Support vector regression machines. In: Mozer M, Jordan M, Petsche T, editors. In: Advances in neural information processing systems, vol. 9, MIT Press; 1996, URL https://proceedings.neurips.cc/paper_files/paper/1996/file/d38901788c533e8286cb6400b40b386d-Paper.pdf.
[114] Bansal G, Nushi B, Kamar E, Lasecki WS, Weld DS, Horvitz E. Beyond accuracy: The role of mental models in human-AI team performance. In: Proceedings of the AAAI conference on human computation and crowdsourcing, vol. 7, 2019, p. 2–11. http://dx.doi.org/10.1609/hcomp.v7i1.5285.
[115] Cabitza F, Campagner A, Ronzio L, Cameli M, Mandoli GE, Pastore MC, Sconfienza LM, Folgado D, Barandas M, Gamboa H. Rams, hounds and white boxes: Investigating human–AI collaboration protocols in medical diagnosis. Artif Intell Med 2023;138:102506. http://dx.doi.org/10.1016/j.artmed.2023.102506.
[116] Ribeiro MT, Singh S, Guestrin C. Why should I trust you? In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM; 2016, http://dx.doi.org/10.1145/2939672.2939778.
[117] Zhang Y, Liao QV, Bellamy RKE. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. FAT* '20, New York, NY, USA: Association for Computing Machinery; 2020, p. 295–305. http://dx.doi.org/10.1145/3351095.3372852.
[118] Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 2017;30.
[119] Mothilal RK, Sharma A, Tan C. Explaining machine learning classifiers through diverse counterfactual explanations. In: Proceedings of the 2020 conference on fairness, accountability, and transparency. FAT* '20, New York, NY, USA: Association for Computing Machinery; 2020, p. 607–17. http://dx.doi.org/10.1145/3351095.3372850.
[120] Amann J, Vetter D, Blomberg SN, Christensen HC, Coffee M, Gerke S, Gilbert TK, Hagendorff T, Holm S, Livne M, Spezzatti A, Strümke I, Zicari RV, Madai VI. To explain or not to explain?—Artificial intelligence explainability in clinical decision support systems. PLOS Digit Health 2022;1(2):e0000016. http://dx.doi.org/10.1371/journal.pdig.0000016.
[121] European Union The European Parliament and the Council of the European Union. On ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing regulation (EU) no 526/2013 (Cybersecurity Act) (Text with EEA relevance), pE/86/2018/REV/1. 2019, URL https://eur-lex.europa.eu/eli/reg/2019/881/oj. [Accessed 30 May 2023].
[122] Cavoukian A, et al. Privacy by design: The 7 foundational principles. Inf Priv Comm Ontario, Canada 2009;5:12.
[123] Dwork C. Differential privacy. In: Automata, languages and programming: 33rd international colloquium, ICALP 2006, venice, Italy, July (2006) 10-14, proceedings, part II 33. Springer; 2006, p. 1–12.