

Ethical Perceptions and Trust in Artificial Intelligence: A PLS-SEM Model of Students' Intention to Use AI

Sofia Anastasiadou^{1,2}[0000-0001-6404-5003]

Ioannis Kougias¹[0000-0001-8019-456X]

¹ University of the Peloponnese, Great Alexandrou Steet, 26334, Koukouli, Patras, Greece

² University of the Western Macedonia, Active Urban Planning Zone (ZEP), 50 150, Kozani, Greece

Abstract

The rapid integration of Artificial Intelligence (AI) into higher education offers substantial opportunities while simultaneously raising critical concerns related to ethics, safety, and trust. This study investigates how key dimensions of AI ethics named Human Rights, Autonomy, Bias, Fairness, Transparency, and Safety influence Trust in AI and, subsequently, university students' Intention to Use AI. The proposed relationships are examined using Partial Least Squares Structural Equation Modeling (PLS-SEM). The measurement model is assessed in terms of internal consistency reliability, convergent validity, discriminant validity, and collinearity through Cronbach's alpha, rho_A, rho_C, Average Variance Extracted (AVE), HTMT, the Fornell–Larcker criterion, and VIF. The findings highlight Safety as a particularly influential determinant of Trust in AI, while Trust in AI positively contributes to students' intention to use AI technologies. At the same time, the measurement assessment reveals discriminant validity concerns between Transparency and Intention to Use AI, indicating substantial empirical overlap that warrants further investigation. The study contributes to the emerging literature by proposing a human-centered and ethics-oriented AI adoption framework, positioning trust as a critical mechanism linking students' ethical perceptions of AI systems with their behavioral intention to adopt them in higher education.

Keywords: Ethical Perceptions, Trust, AI, PLS-SEM Model, Students, Intention, Use AI

1. Introduction

Artificial Intelligence (AI) is one of the most important pillars of modern digital transformation, influencing education, work, health, the economy, and everyday decision-making. In higher education, applications such as generative AI, digital teaching assistants, personalized learning systems, and tools to support academic work create significant opportunities to enhance learning, broaden access to knowledge, and improve the educational process. However, the increasing use of ICT is accompanied by significant ethical, social, and institutional challenges. Issues related to respect for human rights, human autonomy, algorithmic bias, fairness, transparency, and safety take on particular importance (Seremeti & Anastasiadou, 2025; Seremeti et al., 2025). These dimensions lie at the core of responsible and human-centered AI, as acceptance of a system depends not only on its usefulness and technological effectiveness, but also on whether users perceive it as safe, fair, transparent, and compatible with human autonomy and fundamental rights.

In this context, trust in AI (Trust in AI) takes on particular importance, as it reflects the belief that an AI system is reliable, safe, and predictable, and operates in a manner consistent with users' values and interests. Trust can, therefore, serve as a critical mechanism through which perceptions of AI's ethical operation are transformed into a positive or negative intention to use AI.

Exploring these relationships is of particular interest to the student population, which is coming into increasingly frequent contact with AI applications in the context of their studies and daily lives (Karachristos et al., 2026). Understanding the factors that shape trust and intention to use can contribute to the creation of a responsible framework for integrating AI into higher education.

From this perspective, the present study examines the relationships between the six-core AI Ethical Dimensions—Human Rights, Autonomy, Bias, Fairness, Transparency, and Safety—Trust in AI, and Intention to Use AI, utilizing the Partial Least Squares Structural Equation Modeling (PLS-SEM) method. The aim is to investigate, on the one hand, the extent to which the ethical dimensions of AI shape students' trust and, on the other hand, the extent to which this trust influences their intention to use AI systems and applications.

2. Theoretical Framework

The AI Ethical Dimensions constitute a multidimensional framework for evaluating the responsible development and use of AI. This study examines six key dimensions: Human Rights, Autonomy, Bias, Fairness, Transparency, and Safety. These dimensions may influence how users perceive the reliability of AI systems, shaping Trust in AI and, through it, the Intention to Use AI. Their theoretical exploration therefore forms the basis for the development and empirical testing of the proposed PLS-SEM model.

2.1 Human Rights

The rapid development and widespread integration of AI into critical areas of social, economic, and political life make the protection of human rights a fundamental aspect of contemporary technological governance. Although AI creates significant opportunities for economic and social development and can improve the effectiveness of decision-making processes, certain applications of AI may negatively affect individuals or social groups and jeopardize fundamental rights (Mantelero & Esposito, 2021). The use of AI therefore requires appropriate safeguards to ensure that technological progress does not undermine human dignity, autonomy, freedom, and equality.

Human rights can serve as a normative link between ethics and law. Hogan and Lasek-Markey (2024) argue that human rights frameworks serve as a common point of reference for the two fields and can contribute to the development of a legal governance of the TN grounded in ethical principles. Similarly, Yeung et al. (2020) view the international human rights framework as a robust set of normative standards capable of guiding the design, development, and implementation of AI toward a responsible and human-centered direction. Consequently, the protection of human rights is not an external constraint on innovation, but a principle that must be integrated throughout the entire lifecycle of AI systems.

This need is reinforced by the range of risks associated with AI applications, such as violations of privacy and personal data, algorithmic bias and discrimination, security issues, the unequal distribution of social and economic benefits, and the potential impacts on political participation and democratic processes (Stahl et al., 2023). Particularly critical is the risk that algorithmic systems will reproduce or reinforce existing social inequalities, presenting the relevant decisions as seemingly neutral and objective.

A prime example is algorithmic and predictive policing. The combined use of AI, biometric technologies, facial and voice recognition systems, GPS, mobile devices, cameras, and social media data significantly increases the potential for systematic surveillance. The widespread adoption of such practices can restrict privacy, self-determination, and free will, and may also affect the exercise of fundamental freedoms, such as the right of assembly and association, while raising broader concerns about the quality of democracy (Kanelos, pp. 226–236).

Similar challenges are found on social media, where algorithms are not limited to personalizing content but also influence the visibility, categorization, and participation of users in the digital public sphere. Algorithmic mediation can affect the organization of collective actions and the effective exercise of the right to assembly, while at the same time it can lead to the exclusion or unequal treatment of vulnerable social groups, such as minorities, people with disabilities, and people of lower socioeconomic status (Panagopoulou-Koutnatzi, 2023, p. 105).

Consequently, the relationship between AI and human rights goes beyond addressing individual technological risks and concerns the establishment of a human-centered model of AI governance, in which dignity, autonomy, privacy, equality, non-discrimination, and democratic freedoms are fundamental prerequisites for the design and socially acceptable use of technology.

A shift is therefore required from ex post protection of rights to their extant integration into the design, development, evaluation, and implementation of AI systems (Yeung et al., 2020; Stahl et al., 2023; Hogan & Lasek-Markey, 2024). In this way, technological innovation can be combined with ethical and legal responsibility, keeping humans and their fundamental rights at the center of the digital transition.

2.2 Autonomy

Human autonomy is a multidimensional concept linked to an individual's ability to self-determine and shape their life and choices in accordance with their personal values, beliefs, and aspirations. The concept has evolved historically from the self-governance of ancient Greek political thought and Kantian moral autonomy to contemporary forms of personal, political, scientific, and clinical autonomy (Colburn, 2022). Despite the different theoretical approaches, a central element remains the human capacity to act as an agent of conscious and meaningful choices rather than as a passive recipient of externally imposed.

The development of Artificial Intelligence brings autonomy back to the center of the ethical dialogue, as AI systems no longer function merely as auxiliary tools but mediate, support, or automate human decisions. The relationship between humans and AI thus exhibits a dual dynamic: technology can expand options, offer personalized support, and enhance decision-making, but at the same time it can lead to manipulation, surveillance, dependence, and the restriction of independent judgment (Catena, 2026). The critical issue, therefore, concerns the conditions under which technological mediation enhances or restricts human autonomy.

Of particular importance is the distinction between human autonomy and the functional autonomy of AI systems. These systems are considered autonomous when they can process information and generate decisions or actions without continuous human intervention. Although functional autonomy can increase efficiency and safety, it is not sufficient as a criterion for acceptance when automated decisions directly affect people's lives, rights, or opportunities. In such cases, clear limits on automation, human oversight, and a genuine ability to intervene are required (Council of Europe, 2019).

Autonomy is also shaped within a broader sociotechnical environment and is linked to the development of human potential, social recognition, self-respect, and the ability to act independently (Laitinen & Sahlgren, 2021). Manipulation, coercion, lack of transparency, and social or political forms of oppression can substantially limit autonomy (Rubel et al., 2021), making it necessary to evaluate AI within the social, cultural, and institutional context in which it is used (Calvo et al., 2020).

At the same time, autonomy depends on actual access to social, material, and economic resources. As algorithmic systems now influence decisions related to education, work, the economy, social protection, and health, they can determine not only individuals' decisions but also the range of options available to them. O'Neil (2016) points out that algorithmic decision-making can reproduce and reinforce existing social inequalities. Therefore, when access to opportunities and resources depends on opaque algorithmic processes, not only is fairness jeopardized, but so is the meaningful exercise of autonomy.

This issue takes on particular significance in the field of health care, where autonomy is a fundamental principle of bioethics. Patient autonomy refers to a patient's ability and right to make decisions about their health and care in accordance with their personal values and preferences (Beil et al., 2019). Exercising this autonomy requires a voluntary and informed choice, adequate information, and the absence of coercion (Beauchamp & Childress, 2001). The introduction of AI into diagnosis, prognosis, and therapeutic decision-making complicates this relationship, particularly when the models are characterized by limited transparency and interpretability.

The opacity of algorithmic recommendations can limit understanding of the reasons behind a medical decision and, consequently, undermine informed consent. At the same time, overreliance on automated systems can limit both the patient's active role and the physician's professional judgment, thereby affecting shared decision-making (Grote & Berens, 2020).

Consequently, the relationship between AI and autonomy should not be viewed as a simple conflict between humans and automation. AI can enhance human self-determination when transparency, informed consent, genuine choice, and human oversight are ensured. Conversely, when technological mediation manipulates choices, limits understanding, reproduces social inequalities, or shifts substantive decisions from humans to the system, it can undermine autonomy. The goal, therefore, is a complementary relationship between human and technological capabilities, in which the functional autonomy of AI remains oriented toward protecting and enhancing human autonomy.

2.3 Bias

Algorithmic bias is one of the most significant ethical challenges facing AI. It does not concern the algorithm alone but can arise at different stages of a system's life cycle: in the training data, in design and modeling choices, in the learning process, and in the social context of its application (Ching et al., 2018; Danks, 2017). Furthermore, it can be maintained or reinforced during the operation and adaptation of systems, a fact that refutes the notion that automated decision-making is inherently neutral or objective (Howard & Borenstein, 2018; Murphy et al., 2021). The more automated decisions are integrated into critical social functions, the more pronounced the consequences of algorithmic bias become (Ferrier, 2021).

Training data is a central source of bias. Large datasets are not socially neutral, as they may embed historical inequalities, stereotypes, and biases related to gender, race, ethnicity, religion, political beliefs, or vulnerable social groups (Zhuo et al., 2023). When machine learning models are trained on such data, there is a risk that they will reproduce or reinforce these inequalities, transforming them into predictive patterns and automated decisions.

Danks (2017) identifies three main levels at which bias can occur: modeling, training, and deployment. During modeling, the selection of variables, objectives, and criteria can introduce new forms of bias. During training, non-representative or historically biased data can lead to corresponding patterns. Finally, during deployment, a model developed for a specific population may produce problematic results when applied to different social groups. Consequently, bias must be assessed throughout the entire life cycle of the system and not just in the final result.

Of particular importance is the potential for machine learning to reinforce existing biases. Yuste et al. (2017) and Verghese et al. (2018) point out that through repeated algorithmic processing, certain biases can gain disproportionately greater weight. Thus, a model based on a limited or non-representative set of social representations may favor certain groups and disadvantage others, causing substantial harm (Yuste et al., 2017).

The consequences of bias become particularly serious when AI is involved in decisions that determine access to social goods and opportunities. Such applications are found in hiring, education, healthcare, banking, credit granting, and the justice system. In these fields, algorithmic bias can lead to unequal treatment based on gender, disability, ethnic origin, or social vulnerability, transforming a technical problem into an issue of social justice and fundamental rights (Bormila, 2021). Similarly, the underrepresentation of minority groups in the data can lead to lower accuracy and less favorable decisions for their members (Russell, 2021).

From this perspective, bias is directly linked to fairness (fairness). The absence of discriminatory intent on the part of designers does not guarantee that a system's outcomes are fair. A model may seemingly apply the same criteria to everyone and yet systematically produce less favorable outcomes for specific social groups. Assessing fairness therefore requires examining both the technical functioning and the social distribution of benefits, risks, and burdens. The perpetuation of inequalities related to gender, ethnic origin, or social class can lead to new forms of exclusion and rights violations (Panagopoulou-Koutnatzi, 2023, p. 137; Weidinger et al., 2021).

Bias also manifests in natural language processing (NLP) systems and word embedding techniques. Training models on large volumes of linguistic data obtained from the Internet and social media may result in the incorporation and reproduction of stereotypical and negative social associations. Kanellos (2020, pp. 278–279) points out that such techniques can embed biases related to social class, religious beliefs, racial groups, or age. A representative example concerns algorithmic recruitment and personnel-selection processes in the United States, where certain characteristics could lead to the exclusion of applicants with names perceived as being of foreign origin (Kanellos, 2020, pp. 278–279).

Similar challenges arise in social media environments, where AI systems influence the selection, ranking, and dissemination of content. The algorithmic amplification of content associated with racism, xenophobia, homophobia, or other forms of social hostility may contribute to the marginalization of vulnerable groups, increased polarization, and the disruption of social cohesion and democratic dialogue (Anastasiadou, 2026a). Bias therefore acquires a collective and societal dimension, as its consequences extend beyond individual users and may contribute to reshaping broader social relations and reinforcing existing inequalities.

Consequently, algorithmic bias should not be regarded merely as a problem of data quality or technical performance; rather, it constitutes a broader sociotechnical and normative challenge.

Addressing it requires the use of representative datasets, continuous assessment of algorithmic outcomes across different social groups, transparency, accountability, and meaningful opportunities for human intervention and oversight. Above all, it requires a shift from the assumption of algorithmic neutrality toward an explicit commitment to algorithmic fairness and justice.

This requirement is ultimately connected to the principle of human dignity. Given that moral autonomy lies at the core of human dignity (Tsinorema, 2018), individuals cannot be treated merely as collections of data points or as objects of statistical classification. Technological efficiency, therefore, cannot constitute the sole criterion for evaluating an AI system. Equality, non-discrimination, individual freedom, and fundamental rights must be incorporated as integral principles into both the design and implementation of AI systems.

The fundamental challenge is not simply to develop algorithms that exhibit lower levels of bias, but to create AI systems that do not transform the social inequalities of the past into the automated inequalities of the future. From this perspective, addressing algorithmic bias becomes not only a technical objective but also an ethical and societal imperative, requiring AI development to remain firmly aligned with human dignity, equality, justice, and the protection of fundamental rights.

2.4 Fairness

Justice and fairness constitute fundamental principles in the ethical development and deployment of Artificial Intelligence, as they concern the equitable distribution of the benefits, opportunities, risks, and burdens arising from the use of algorithmic systems (Li et al., 2023). Fairness is not limited to the absence of explicit discrimination; rather, it requires that algorithmic decisions do not systematically create advantages or disadvantages for particular social groups. Characteristics such as gender and ethnic or racial background are of particular importance, as algorithmic systems have been shown to produce differentiated outcomes on the basis of such factors (Grote & Berens, 2020; Ahmad et al., 2020).

Fairness is closely related to bias, although the two concepts are not synonymous. Bias refers to the mechanisms through which distortions, prejudices, or unrepresentative patterns enter datasets and computational models, whereas fairness concerns the broader question of whether the decisions and consequences generated by a system are distributed equitably across different individuals and groups. Accordingly, bias represents a major source of algorithmic unfairness, but its elimination does not automatically guarantee fair outcomes.

An AI model may demonstrate high overall accuracy while simultaneously exhibiting different error rates or producing disproportionately adverse consequences for particular population groups. Thus, evaluating fairness requires moving beyond aggregate measures of predictive performance and examining how algorithmic outcomes are distributed across relevant demographic and social groups. From this perspective, fairness is not merely a technical property of an algorithm but a multidimensional ethical principle that concerns the social consequences of AI-assisted decision-making and its implications for equality and non-discrimination.

This dimension is particularly important in the field of healthcare, where algorithmic fairness is directly linked to equitable access to diagnosis, treatment, and the allocation of healthcare resources. Health data are not collected with the same degree of completeness across all social groups, as socioeconomic inequalities, geographical constraints, unequal access to healthcare services, and the limited participation of certain populations in medical research can lead to unequal representation. When such data are used to train machine learning models, pre-existing inequalities may be transferred into the AI system itself and re-emerge in the form of differences in accuracy or decision quality (Jiang et al., 2021).

Algorithmic unfairness, therefore, does not always arise from technical errors but may reflect deeper social inequalities. When socially or economically vulnerable groups have more limited access to diagnostic testing and healthcare services, fewer data are generated for these populations. In contrast, populations with greater access to healthcare services are more extensively represented in datasets. The use of such unevenly distributed information may result in systems that perform more effectively for populations that are already better represented, thereby creating a mechanism through which existing inequalities are reproduced and reinforced.

A characteristic example is minority bias. As Rajkomar et al. (2018) point out, when minority groups are represented by only a small number of cases in training data, models may struggle to identify the patterns characterizing these groups with sufficient accuracy. Underrepresentation may therefore result in lower predictive performance for groups that are already more vulnerable to social and health inequalities.

Fairness thus also acquires a clear statistical dimension, as the overall performance of a model is not sufficient; it is necessary to examine whether its accuracy, errors, and impacts are equitably distributed across different subgroups.

Missing data bias is of similar importance and occurs when the absence of data is not random but is associated with social, economic, or geographical factors. Individuals living in remote areas or having limited access to healthcare services may have a smaller digital footprint. The absence of data, therefore, does not imply the absence of healthcare needs or health problems; rather, it may indicate unequal access to care. If this dimension is not taken into account, AI systems may generate inaccurate or unfair predictions. At the same time, *selection bias* occurs when the data selection process fails to ensure the representative participation of different population groups (Rajkomar et al., 2018).

Fairness, however, does not depend solely on the quality and representativeness of the data. The complexity of algorithmic models and the limited understanding of how their decisions are generated create additional challenges (Vellido, 2019). When it is unclear which features influenced a decision and in what way, it becomes more difficult to determine whether particular groups are systematically subjected to less favorable treatment. For this reason, transparency and interpretability are directly linked to fairness.

At the same time, data representativeness is not merely a methodological requirement but also an ethical prerequisite. A model's ability to generalize reliably depends on the extent to which the training data reflect the actual heterogeneity of the population to which it is applied (Yeng et al., 2019). Therefore, the fair operation of an AI system requires not only large volumes of data but also appropriate levels of data quality, completeness, and representation of different social groups.

Fairness, therefore, cannot be treated as a technical correction applied retrospectively. Ahmad et al. (2020) point out that fairness-related issues may arise throughout the entire processing pipeline, from system design and data selection to training, processing, the representation of results, and their use in decision-making. Ensuring fairness should therefore constitute a cross-cutting principle throughout the entire lifecycle of an AI system.

From this perspective, algorithmic fairness encompasses three interconnected levels: the quality and representativeness of data, the fairness of algorithmic processes, and the fairness of social outcomes. A technically accurate system cannot be considered ethically adequate when its errors, risks, or adverse consequences disproportionately affect particular social groups or populations that are already vulnerable.

The fundamental objective of fairness in AI is, therefore, to move beyond merely equal algorithmic treatment toward fair and substantively equitable treatment, taking into account the different social conditions, needs, and vulnerabilities of the populations to which AI systems are applied.

2.5 Transparency

Transparency constitutes one of the fundamental principles of the ethical and responsible development of Artificial Intelligence systems, as it is directly associated with the ability of users and stakeholders to understand how these systems operate and the processes through which their decisions are generated (Floridi et al., 2018; Frost et al., 2022). Its importance becomes even greater in the context of big data and machine learning, where the complexity of models and the volume of information being processed create significant challenges regarding the ability to monitor and understand algorithmic processes (Bertino et al., 2019).

Data transparency is a critical prerequisite for the ethically responsible use of data (Bertino et al., 2019). This requirement also has a clear legal dimension, as, according to Article 5(1) of the GDPR, data subjects must be informed about the collection and processing of their personal data by the data controller (Panagopoulou-Koutnatzi, 2023, p. 167). Transparency, therefore, is not limited to the technical operation of an algorithm but also encompasses the provision of adequate information regarding the data being used, the purpose of their processing, and the potential consequences for individuals.

In the context of AI, transparency refers to providing users and stakeholders with access to sufficient information concerning the structure, operation, and decision-making processes of a system. For this reason, it is considered a fundamental prerequisite for the development of trustworthy AI (Balasubramaniam et al., 2023). According to Radanliev (2025), transparency enables stakeholders to better understand how AI systems operate and, consequently, to evaluate their decisions and the resulting consequences.

Transparency is directly linked to AI governance and to the development of trust in the way machine learning algorithms generate decisions across different scientific and social domains (Balasubramaniam et al., 2023). At the same time, the ability to understand algorithmic decisions strengthens accountability, as it makes it possible to identify the factors that led to a particular outcome and to assign responsibility when that outcome is problematic or unfair (Radanliev, 2025).

The central role of transparency is also confirmed by international ethical guidelines for AI. Recognized expert groups have emphasized the need for AI systems to operate in a sufficiently understandable and well-documented manner, enabling users and stakeholders affected by their decisions to evaluate their functioning and consequences (Balasubramaniam et al., 2023). Transparency, therefore, constitutes a fundamental characteristic of responsible and human-centered AI rather than merely a technical property of the system.

This need is particularly pronounced in high-risk domains, such as healthcare, judicial practice, and banking transactions, where algorithmic decisions may have direct and significant consequences for individuals' lives, rights, and opportunities. In such contexts, the ability to understand how an algorithmic decision is reached constitutes a fundamental prerequisite for assessing its accuracy, fairness, and legality.

A particular challenge is posed by so-called "black-box" systems, in which the internal decision-making process is not easily understandable, even by experts. The limited transparency of such systems makes it difficult to identify errors, biases, or unfair decisions and may constrain accountability. For this reason, the requirement to make the operation of black-box algorithms more transparent and interpretable constitutes a central issue in the ethical and responsible development of AI (Haresamudram et al., 2023).

Consequently, transparency should not be understood solely as the disclosure of technical information about an algorithm. Rather, it constitutes a broader principle that encompasses user information, personal data protection, understanding of algorithmic decisions, the capacity for oversight, and accountability. From this perspective, the greater the perceived transparency of an AI system, the stronger the conditions for evaluating its trustworthiness and, consequently, for developing trust in AI.

2.6 Safety

Artificial Intelligence (AI) is bringing about significant transformations across multiple dimensions of human activity, while simultaneously raising critical challenges concerning the safety and ethics of its development and use (Salhab et al., 2024). AI Safety concerns the development and operation of systems in a reliable, controllable, and robust manner against errors, failures, malicious interventions, and other potential risks. Its key dimensions include explainability and interpretability, model reliability and robustness, the mitigation of bias, and protection against adversarial or malicious attacks (adversarial attacks) that may compromise the proper functioning of AI systems (Salhab et al., 2024).

AI Safety has emerged as a particularly important field of scientific research, as the rapid development and expanding applications of AI are accompanied by complex and interconnected risks that require systematic investigation and effective mechanisms for prevention and control (Rudner & Toner, 2021). Safety, therefore, is not limited to preventing technical failures but more broadly concerns the ability of AI systems to operate in a predictable and reliable manner without creating unacceptable risks for users and society.

Data protection constitutes a critical dimension of safety, particularly in applications that process large volumes of personal and sensitive information. This issue becomes even more important in healthcare, where AI systems and machine learning algorithms are often trained on extensive patient datasets. Protecting such information is an essential prerequisite for the safe and socially acceptable use of AI in healthcare (Murphy et al., 2021).

At the same time, AI systems are vulnerable to cyberattacks, making the development of robust cybersecurity mechanisms necessary, particularly when health data are involved. Unauthorized access to, alteration of, or theft of medical records may violate patient privacy, undermine data reliability, and, in certain cases, even affect the safety of clinical care. Effective cybersecurity systems are therefore essential for protecting the confidentiality, integrity, and secure management of information (Puri & Gochhait, 2023).

Within this context, Goel et al. (2021) identify three fundamental security principles: confidentiality, integrity, and availability. Confidentiality refers to ensuring that data are accessible exclusively to authorized individuals; integrity concerns protecting data against unauthorized modification or deletion; and availability refers to ensuring that information remains accessible to legitimate users whenever required. These principles constitute fundamental prerequisites for the secure management of data in AI environments.

The importance of these principles extends throughout the entire data lifecycle, from data creation and collection to storage, processing, reuse, and final destruction (Kousouni-Pantazopoulou, 2022, p. 45). The reuse of data and computational resources may create confidentiality risks, while unauthorized modification or deletion may compromise data integrity. Similarly, the inability to access data in a timely manner may create availability problems even when the information is protected through encryption. These risks may also be associated with issues concerning the protection of intellectual property (Kousouni-Pantazopoulou, 2022, p. 45).

Effective AI safety therefore requires an approach that encompasses the entire lifecycle of AI systems rather than focusing solely on their final deployment. Wang et al. (2024) emphasize the need to develop appropriate security mechanisms and architectures, as well as procedures for monitoring all stages of the development and operation of algorithms and their different use cases. At the same time, an appropriate legislative and regulatory framework is required so that technical safety safeguards are accompanied by clear institutional requirements and accountability mechanisms.

Consequently, AI Safety constitutes a multidimensional ethical and technological requirement that integrates system reliability and robustness, protection against cyber threats, secure data management, and the prevention of potential harm (Anastasiadou, 2026a;2026b;2026c). H). An AI system cannot be considered responsible merely because it is technically effective; it must also operate in a safe, predictable, and controllable manner throughout its entire lifecycle. From this perspective, perceived safety constitutes a critical prerequisite for AI acceptance, as users' belief that a system protects their data, mitigates risks, and operates reliably may substantially strengthen trust in AI.

2.7 Research Gap

Despite the rapid growth of literature on the acceptance and use of Artificial Intelligence in education, a substantial proportion of existing research has focused primarily on traditional determinants of technology acceptance, such as perceived usefulness, ease of use, effectiveness, performance, and adoption intention. In contrast, the role of users' ethical perceptions as determinants of trust and intention to use AI has been investigated less systematically.

More specifically, critical dimensions of AI Ethics, such as Human Rights, Autonomy, Bias, Fairness, Transparency, and Safety, are often examined separately or predominantly at theoretical and normative levels. Consequently, empirical research investigating their simultaneous effects within an integrated structural model that connects AI ethical principles with behavioral acceptance remains limited.

At the same time, the role of Trust in AI as an intermediate mechanism between ethical perceptions and intention to use AI has not been sufficiently investigated. The critical question is not only whether users perceive AI as useful and effective, but also whether their perception that an AI system is safe, fair, transparent, free from bias, and compatible with human autonomy and fundamental rights shapes their trust and, through that trust, their intention to use it.

This research gap is particularly significant in Higher Education, where the rapid expansion of generative AI has substantially transformed the ways in which students search for information, learn, and complete academic assignments, while the ethical and social implications associated with its use continue to be investigated.

There is, therefore, a need to develop an integrated empirical model that links AI Ethical Dimensions with the conceptual constructs of Trust in AI and Intention to Use AI, and identifies which specific ethical dimensions contribute most strongly to the development of trust and, consequently, to students' intention to use AI.

2.8 Novelty of Research

The main novelty of the present study lies in the development of an ethically oriented and human-centered model of Artificial Intelligence acceptance. In contrast to traditional technology acceptance approaches, which primarily focus on usefulness, ease of use, and effectiveness, the present study examines AI acceptance through the ethical evaluation of AI systems and the trust that such an evaluation may generate.

More specifically, the novelty of the study lies in the simultaneous integration of six key dimensions of AI Ethics, namely Human Rights, Autonomy, Bias, Fairness, Transparency, and Safety, within a single structural model. In this way, the ethical dimensions are not examined in isolation but as distinct yet complementary components of students' overall perceptions of the responsible functioning of AI.

At the same time, Trust in AI is not treated merely as another independent variable, but as a central mechanism linking the ethical evaluation of AI systems to behavioral intention to use AI. This approach makes it possible to investigate not only whether ethical perceptions influence trust, but also whether trust functions as the mechanism through which these perceptions are translated into intention to use AI.

Furthermore, the study seeks to determine which of the individual ethical dimensions exerts the strongest relative effect on Trust in AI. Identifying, for example, Safety as a particularly strong determinant could provide a more differentiated understanding of the criteria on the basis of which students evaluate and develop trust in AI systems. Finally, the application of Partial Least Squares Structural Equation Modeling (PLS-SEM) enables the simultaneous assessment of the measurement and structural models, as well as the comparison of the relative strength of the individual relationships within the proposed model.

The proposed approach therefore shifts the research focus from the traditional pathway of Perceived Usefulness → Technology Acceptance toward a more human-centered model: AI Ethical Dimensions → Trust in AI → Intention to Use AI.

This shift constitutes a central element of the study's novelty, as it conceptualizes AI acceptance not merely as a technological choice, but as the outcome of a preceding process of ethical evaluation and trust formation.

2.9. Significance of the Research

The significance of the present study extends across theoretical, empirical, methodological, and practical dimensions. At the theoretical level, the study brings together two fields that are often examined separately: Artificial Intelligence ethics and technology acceptance. In doing so, it supports a broader perspective according to which AI adoption is not exclusively a decision based on usefulness and effectiveness, but also a process involving ethical evaluation, the formation of risk perceptions, and the development of trust.

At the empirical level, the study makes it possible to determine whether the different ethical dimensions have equal or varying levels of importance in shaping Trust in AI. The potential identification of Safety as the strongest determinant may indicate that students attach particular importance to whether an AI system is perceived as safe, controllable, and reliable. In this way, the study can provide a more nuanced understanding of the ethical factors underlying trust in and acceptance of AI.

At the methodological level, the use of Partial Least Squares Structural Equation Modeling (PLS-SEM) enables the simultaneous estimation of a complex model involving multiple latent variables, the assessment of the reliability and validity of the conceptual constructs, and the investigation of both direct and potentially indirect effects among them. Furthermore, it provides the opportunity to compare the relative strength of the different ethical dimensions in shaping trust.

At the practical level, the findings can be utilized by universities, educators, educational policymakers, and AI application developers to formulate strategies for the responsible integration of AI into Higher Education. The effective adoption of AI systems should not focus exclusively on their technological capabilities but should also involve creating conditions that promote safety, transparency, fairness, the protection of autonomy and human rights, and, ultimately, trust.

Consequently, the study shifts the research focus from the simple question, "*Will students use AI?*", toward a deeper and more human-centered question: "*Under what ethical conditions do students trust AI, and how does this trust translate into their intention to use it?*" Addressing this question can make a substantial contribution to the development of a model for the responsible, trustworthy, and human-centered adoption of AI in Higher Education.

3. Methodology

3.1 Research Instrument

The instrument entitled "Human Rights, Autonomy, Bias, Fairness, Transparency, Safety, and AI" was developed by the author and consists of 29 items organized into six conceptual constructs.

The initial instrument had 68 items organized into six conceptual constructs. In the present study we kept only those items with the most significant items.

Thus, the first conceptual construct, Human Rights, comprises 3 items (e.g., Hur1: AI algorithms, through social media platforms, allow the dissemination of homophobic biases). The second construct, Autonomy, consists of 3 items (e.g., Aut1: Machine learning algorithms do not undermine human autonomy). The third construct, Bias, comprises 6 items (e.g., Bias1: There are issues of bias in AI). The fourth construct, Fairness, consists of 4 items (e.g., Fair3: Neural networks used for text representation reproduce social stereotypes).

The fifth conceptual construct, Transparency, comprises 7 items (e.g., Tran2: *There are issues related to a lack of transparency regarding who works behind social media platforms*). Finally, the sixth conceptual construct, Safety, consists of 6 items (e.g., Sec1: *AI technologies are vulnerable to cyberattacks*).

Two more conceptual constructs named “Trust in AI” and “Behavioral Intention to Use AI” (InU) were used for the purpose of the study. The conceptual construct “Trust in AI” consists of 3 items (Trust1–Trust3), capturing three complementary aspects of trust: reliability, competence/accuracy, and confidence/reliance.

The first item, Trust1 (I trust AI systems to provide reliable results), measures Reliability. More specifically, it assesses the extent to which users believe that AI systems generate results that can be trusted, focusing on the consistency and reliability of AI-generated outcomes.

The second item, Trust2 (I trust AI systems to perform their intended functions accurately), measures the perceived Competence/Accuracy of AI systems. More specifically, it captures the belief that an AI system is technically capable of performing its intended functions correctly and accurately. Thus, this item reflects the perceived technical competence of the system.

The third item, Trust3 (*I feel confident relying on AI systems when I use them*), refers to Confidence/Reliance. More specifically, it assesses the extent to which trust translates into users’ willingness to rely on AI. It concerns not only the belief that the system performs effectively but also the extent to which users feel confident relying on its output. Finally, the conceptual construct “Behavioral Intention to Use AI” (InU_i) consists of 6 items (InU_i) (e.g., InU2: *I intend to use AI in my studies as often as needed*). Based on InU1–InU6, this conceptual construct reflects individuals’ behavioral intention to use, continue using, and recommend the use of Artificial Intelligence both in their studies and in their daily lives.

The conceptual construct “Behavioral Intention to Use AI” (InU_i) refers to the extent to which individuals intend to use, continue using, and increase their use of AI applications in the future, both within academic contexts and in their broader everyday lives. The construct initially encompasses the intention to use AI directly and systematically for academic purposes. InU1 and InU2 examine the extent to which users plan to use AI during their studies and how often they consider it necessary. These items therefore capture users’ intention to integrate AI into their routine academic activities. At the same time, InU3 and InU4 broaden the concept of behavioral intention by incorporating users’ intention to recommend AI to others, such as fellow students and relatives. This dimension reflects not only users’ personal acceptance of the technology but also their willingness to encourage its adoption within their social environment. InU5 captures the intention to continue and increase the future use of AI technology. More specifically, it reflects the extent to which users intend to expand their use of AI and remain motivated to choose AI technologies over alternative technological solutions. Finally, InU6 reflects the intention to integrate AI into everyday life, thereby extending the construct beyond the strict academic context.

Overall, Behavioral Intention to Use AI refers to an individual’s willingness and planned commitment to use, continue using, and increase the use of AI technologies in academic and everyday activities, as well as their willingness to recommend the use of AI to others.

3.2 Statistical Methodology

Partial Least Squares Structural Equation Modeling (PLS-SEM) was employed to analyze the proposed research model and assess the relationships among the latent constructs. PLS-SEM is a variance-based structural equation modeling technique that enables the simultaneous estimation of complex relationships between observed indicators and latent variables, as well as among the latent constructs themselves. The analytical procedure was conducted in two consecutive stages.

In the first stage, the measurement model was evaluated to assess the reliability and validity of the constructs. Specifically, indicator reliability and internal consistency reliability were examined, followed by the assessment of convergent and discriminant validity. These procedures were performed to ensure that the measurement scales adequately represented their respective latent constructs before proceeding to the evaluation of hypothesized relationships.

In the second stage, the structural model was assessed to examine the magnitude, direction, and statistical significance of the hypothesized relationships among the constructs. The evaluation focused on the estimated path coefficients and the model's explanatory and predictive capabilities. This stage provided the empirical basis for testing the proposed research hypotheses and determining the extent to which the exogenous constructs explained variation in the endogenous constructs.

PLS-SEM was selected because it is particularly suitable for prediction-oriented research and for investigating models involving multiple latent constructs and simultaneous relationships. Moreover, it provides a flexible analytical framework when the primary objective is to estimate the effects of predictor constructs on endogenous variables. Nevertheless, although PLS-SEM effectively captures the specified structural relationships, complex patterns of concurrent or alternative causality among variables should be interpreted with appropriate caution (Rasoolimanesh, et al., 2021).

3.3 Statistical Hypotheses

Based on the theoretical framework and the proposed structural model, the following null hypotheses (H_0) are formulated:

H_{01} : Perceived protection of and respect for Human Rights by Artificial Intelligence systems has no statistically significant effect on Trust in AI (Human Rights $\rightarrow$ Trust in AI).

H_{02} : Perceived Autonomy has no statistically significant effect on Trust in AI (Autonomy $\rightarrow$ Trust in AI).

H_{03} : Perceived Bias in AI systems has no statistically significant effect on Trust in AI (Bias $\rightarrow$ Trust in AI).

H_{04} : Perceived Fairness of AI systems has no statistically significant effect on Trust in AI (Fairness $\rightarrow$ Trust in AI).

H_{05} : Perceived Transparency of AI systems has no statistically significant effect on Trust in AI (Transparency $\rightarrow$ Trust in AI).

H_{06} : Perceived Safety of AI systems has no statistically significant effect on Trust in AI (Safety $\rightarrow$ Trust in AI).

H_{07} : Trust in AI has no statistically significant effect on Intention to Use AI (Trust in AI $\rightarrow$ Intention to Use AI).

H_{08} : Trust in AI does not significantly mediate the relationship between perceived Safety and Intention to Use AI (Safety $\rightarrow$ Trust in AI $\rightarrow$ Intention to Use AI).

3.4 Participants

The study of a pilot sample consisted of 111 participants, of whom 41 (36.9%) were women and 70 (63.1%) were men. Regarding the age of the respondents, 57 (51.4%) were between 18 and 20 years old, 27 (24.3%) were between 20 and 22 years old, 13 (11.7%) were between 22 and 24 years old, and finally, 14 (12.6%) were over 24 years old. Regarding place of residence, 86 of the 111 respondents, or 77.5%, reported living in an urban area, 22 (19.8%) in a semi-urban area, and finally, 3 (2.7%) in a rural area.

Regarding their school of study, 69 (62.2%) were from the School of Engineering, 30 (27%) were from the School of Natural Sciences, 7 (6.3%) were from the School of Health Sciences, 3 (2.7%) were from the School of Economics and Social Sciences, and finally, 2 (1.8%) were from the Schools of Humanities.

Regarding their year of study, 42 (37.8%) were in their first year, 26 (23.4%) were in their second year, 36 (23.4%) were in their third year, 13 (11.7%) were in their fourth year, 7 (6.3%) were in their fifth year, 7 (6.3%) were in their sixth year, 2 (1.8%) were pursuing a bachelor's degree, and 3 (2.7%) were enrolled in a master's or doctoral program.

4. Results

4.1 PLS-SEM Measurement Model

The PLS-SEM Measurement Model (Figure 1) presents and evaluates the quality of the latent constructs included in the proposed model. The assessment of the measurement model focuses on four key criteria: indicator reliability, internal consistency reliability, convergent validity, and discriminant validity.

These criteria are examined in the following sections to establish the reliability and validity of the measurement model before proceeding with the evaluation of the structural model.

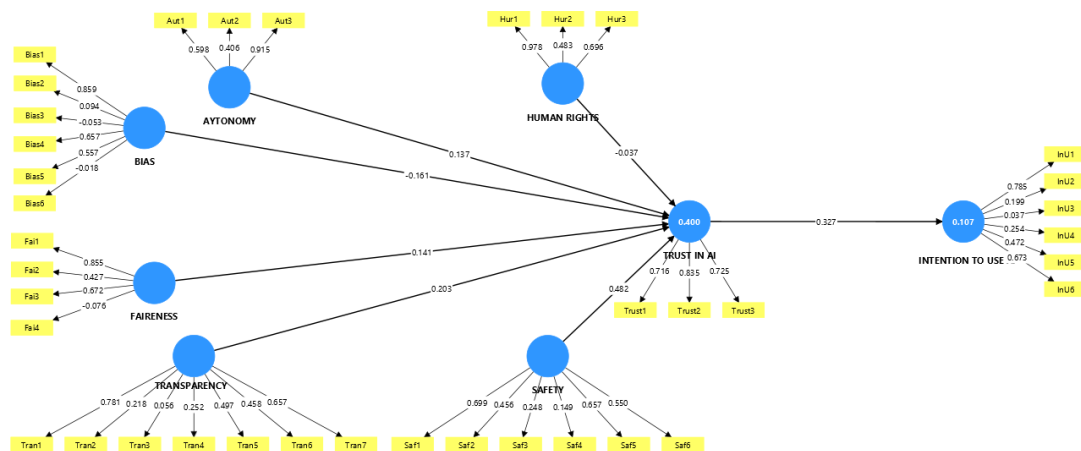


Figure 1: PLS-SEM Measurement Model

The evaluation of the measurement model is conducted prior to examining the structural relationships among the latent variables. More specifically, the internal consistency of the scales is assessed using Cronbach's α and Composite Reliability (CR), convergent validity is evaluated through the Average Variance Extracted (AVE) (Anastasiadou & Anastasiadis, 2019), and discriminant validity is examined using the HTMT criterion.

The Cronbach's α values indicated satisfactory internal consistency for most of the latent variables (Anastasiadou, 2011;2019; Thapa et al., 2016; Taraza & Anastasiadou, 2019). More specifically, the Human Rights ($\alpha = 0.756$), Transparency ($\alpha = 0.880$), Safety ($\alpha = 0.825$), Trust in AI ($\alpha = 0.837$), and Intention to Use AI ($\alpha = 0.777$) scales demonstrated acceptable to very good reliability levels. Fairness showed marginally acceptable reliability ($\alpha = 0.699$), while lower values were observed for Autonomy ($\alpha = 0.696$) and particularly for Bias ($\alpha = 0.689$) (Table 1).

Composite Reliability was generally satisfactory. The values were 0.860 for Human Rights, 0.789 for Autonomy, 0.751 for Bias, 0.810 for Fairness, 0.913 for Transparency, 0.878 for Safety, 0.902 for Trust in AI, and 0.844 for Intention to Use AI. Since most values exceed the recommended threshold of 0.70, it can be concluded that the majority of the constructs demonstrate adequate internal consistency. Particularly high reliability was observed for the Transparency, Safety, and Trust in AI constructs (Table 1).

Regarding convergent validity, the AVE values were above the recommended threshold of 0.50 for most of the variables. Specifically, AVE values of 0.672 for Human Rights, 0.554 for Autonomy, 0.517 for Fairness, 0.601 for Transparency, 0.548 for Safety, and 0.755 for Trust in AI were recorded (Table 1). These findings indicate that the corresponding latent variables explain more than 50% of the variance in their respective indicators.

Table 1. Reliability and Validity of the Measurement Model

	Cronbach's alpha	Composite reliability	Average variance extracted (AVE)
AYTONOMY	0.696	0.789	0.554
BIAS	0.689	0.751	0.435
FAIRENESS	0.699	0.810	0.517
HUMAN RIGHTS	0.756	0.860	0.672
INTENTION TO USE AI	0.777	0.844	0.494,
SAFETY	0.825	0.878	0.548
TRANSPARENCY	0.880	0.913	0.601
TRUST IN AI	0.837	0.902	0.755

4.2 Discriminant Validity – Heterotrait–Monotrait Ratio (HTMT)

The discriminant validity of the model was assessed using the Heterotrait–Monotrait Ratio (HTMT). The HTMT criterion examines whether the latent constructs of the model are sufficiently distinct from one another. As a general rule, HTMT values below 0.90 are considered acceptable, while the more stringent threshold of 0.85 provides stronger evidence of discriminant validity.

The results indicate that the HTMT values for most relationships among the constructs are substantially below the acceptable thresholds. More specifically, the values range from 0.116 to 0.850, suggesting that the dimensions of Autonomy, Bias, Fairness, Human Rights, Intention to Use AI, Safety, Transparency, and Trust in AI generally represent distinct theoretical concepts.

The highest HTMT value was observed between Safety and Trust in AI (HTMT = 0.850). This value lies exactly at the more stringent threshold of 0.85 but remains below the more widely used threshold of 0.90 (Table 2). Therefore, there is no clear violation of discriminant validity. At the same time, this relatively high value is theoretically plausible, as the perceived safety of an AI system is expected to be closely associated with the trust that users develop toward it. However, since the value lies at the boundary of the stricter criterion, this relationship should be reported with appropriate caution.

Another relatively high value was observed between Autonomy and Bias (HTMT = 0.802). Despite the relatively strong association between these two constructs, the value remains below both the 0.85 and 0.90 thresholds and, therefore, supports their conceptual distinctiveness.

Notable but clearly acceptable values were also recorded between Transparency and Intention to Use AI (HTMT = 0.707), Safety and Transparency (HTMT = 0.659), and Safety and Intention to Use AI (HTMT = 0.647). These findings suggest that these constructs are related to one another without exhibiting such a degree of overlap that would raise concerns regarding discriminant validity (Table 2).

In contrast, particularly low HTMT values were observed between Human Rights and Trust in AI (0.116), Autonomy and Trust in AI (0.190), Human Rights and Safety (0.252), Human Rights and Intention to Use AI (0.264), and Human Rights and Transparency (0.267). These low values provide strong evidence that the corresponding constructs capture clearly distinct dimensions of the theoretical model (Table 2).

Overall, the HTMT results provide satisfactory support for the discriminant validity of the measurement model, as none of the values exceeds the threshold of 0.90 (Table 2). Therefore, the eight latent variables can be considered empirically distinct and can be retained as separate constructs in the proposed PLS-SEM model.

Table 2: Discriminant validity - Heterotrait-monotrait ration (HTMT)-Matrix

	AYTONO MY	BIA S	FAIREN ESS	HUM AN RIGH TS	INTENTI ON TO USE AI	SAFE TY	TRANSPARE NCY	TRU ST IN AI
AYTONOMY								
BIAS	0.802							
FAIRENESS	0.419	0.399						
HUMAN RIGHTS	0.422	0.394	0.329					
INTENTION TO USE AI	0.482	0.358	0.379	0.264				
SAFETY	0.330	0.488	0.350	0.252	0.647			
TRANSPARENCY	0.437	0.375	0.366	0.267	0.707	0.659		
TRUST IN AI	0.190	0.290	0.341	0.116	0.435	0.850	0.409	

Particular attention is required only for the Safety–Trust in AI relationship (HTMT = 0.850), due to its borderline value according to the more stringent criterion. Nevertheless, this value does not, by itself, constitute a discriminant validity problem, particularly when HTMT < 0.90 is adopted as the primary assessment threshold. Moreover, the finding is theoretically consistent with the proposed model, in which Safety serves as a determinant of Trust in AI.

4.3 Discriminant validity – Fornell-Larcker criterion

Discriminant validity is a fundamental requirement for the evaluation of the measurement model in PLS-SEM, as it assesses whether each latent construct in the model is empirically distinct from the other constructs. In other words, the aim is to establish that the constructs Autonomy, Bias, Fairness, Human Rights, Safety, Transparency, Trust in AI, and Intention to Use AI represent different theoretical concepts and do not essentially measure the same phenomenon.

The Fornell–Larcker criterion is based on comparing the square root of the Average Variance Extracted ($\sqrt{AVE}$) of each construct with its correlations with the other latent variables. To establish discriminant validity, the ($\sqrt{AVE}$) value of each construct should be greater than the absolute value of its correlation with any other construct. The underlying rationale is that a construct should share more variance with its own indicators than it shares with other constructs in the model.

In the present study, the diagonal values of the Fornell–Larcker matrix are 0.673 for Autonomy, 0.660 for Bias, 0.586 for Fairness, 0.747 for Human Rights, 0.703 for Intention to Use AI, 0.740 for Safety, 0.775 for Transparency, and 0.761 for Trust in AI. In all cases, these values are greater than the corresponding correlations among the different constructs, providing clear evidence of discriminant validity (Table 3).

For Autonomy, the $\sqrt{AVE}$ value is 0.673, while its highest correlation is observed with Bias ($r = 0.313$). The remaining correlations are even lower, such as those with Human Rights ($r = 0.285$), Fairness ($r = 0.160$), and Intention to Use AI ($r = 0.100$). Since $0.673 > 0.313$, Autonomy demonstrates clear discriminant validity with respect to all the other constructs (Table 3).

Similarly, for Bias, the $\sqrt{AVE}$ value is 0.660. Its highest absolute correlation is with Autonomy ($r = 0.313$), while its correlations with Human Rights and Trust in AI are $r = 0.247$ and $r = -0.203$, respectively. Therefore, the value of 0.660 substantially exceeds all the corresponding correlations, confirming that Bias represents a distinct conceptual construct (Table 3).

For Fairness, the $\sqrt{AVE}$ value is 0.586, while its correlations with the other constructs are relatively low. The highest correlation is observed with Trust in AI ($r = 0.200$), followed by Intention to Use AI ($r = 0.190$) and Transparency ($r = 0.186$). The substantial difference between the $\sqrt{AVE}$ value of 0.586 and the highest correlation of 0.200 provides strong evidence that Fairness captures a distinct dimension of the ethical evaluation of AI (Table 3). The Human Rights construct has a $\sqrt{AVE}$ value of 0.747. Its highest correlation is with Autonomy ($r = 0.285$), while its relationships with the remaining variables are even weaker. For example, the correlation with Bias is $r = 0.247$, with Fairness $r = 0.136$, with Safety $r = -0.102$, and with Trust in AI $r = -0.073$. Therefore, Human Rights demonstrates particularly clear empirical distinctiveness from the other ethical dimensions (Table 3).

Of particular interest is Intention to Use AI, for which the $\sqrt{AVE}$ value is 0.703. Its highest correlation is observed with Transparency ($r = 0.451$), followed by Trust in AI ($r = 0.327$). Although the Transparency–Intention relationship is relatively stronger compared with several other relationships in the model, the value of 0.451 remains clearly lower than the $\sqrt{AVE}$ value of 0.703 (Table 3). Therefore, Intention to Use AI remains sufficiently distinct from Transparency as well as from all the other constructs.

Safety is of particular importance to the theoretical model, as it represents the strongest determinant of Trust in AI in the structural model. The $\sqrt{AVE}$ value for Safety is 0.740, while its highest correlation is with Trust in AI ($r = 0.539$). Although this is the highest correlation in the entire matrix, it remains lower than both the $\sqrt{AVE}$ of Safety (0.740) and the $\sqrt{AVE}$ of Trust in AI (0.761) (Table 3). Therefore, the stronger Safety–Trust relationship does not indicate conceptual overlap. Rather, it demonstrates that the two constructs are substantively related while remaining empirically distinct.

For Transparency, the corrected $\sqrt{AVE}$ value is 0.775, representing the highest diagonal value among the eight constructs. Its strongest correlation is observed with Intention to Use AI ($r = 0.451$), followed by Trust in AI ($r = 0.316$), Fairness ($r = 0.186$), and Safety ($r = 0.130$). Since $0.775 > 0.451$, Transparency demonstrates clear discriminant validity (Table 3). This finding is particularly important, as it indicates that perceptions of the transparency of AI systems are conceptually distinct from both trust in these systems and users' intention to use them.

Finally, for Trust in AI, the $\sqrt{AVE}$ value is 0.761. Its strongest correlation is with Safety ($r = 0.539$), followed by Intention to Use AI ($r = 0.327$) and Transparency ($r = 0.316$) (Table 3). Since all these values are lower than 0.761, the results confirm that Trust in AI represents a distinct conceptual construct and does not overlap with perceptions of safety, transparency, or intention to use AI.

Overall, the Fornell–Larcker criterion is satisfied for all eight latent constructs. None of the inter-construct correlations exceed the corresponding square root of the AVE. This finding indicates that each construct shares more variance with its own indicators than with the other constructs included in the model.

This finding is particularly important for the present study because the six dimensions of Human Rights, Autonomy, Bias, Fairness, Transparency, and Safety represent related aspects of the broader concept of AI Ethics. Despite their theoretical interrelatedness, the results indicate that participants perceive them as distinct and empirically distinguishable dimensions. At the same time, the two constructs, Trust in AI and Intention to Use AI, are sufficiently differentiated both from each other and from the six ethical dimensions.

Consequently, the results of the Fornell–Larcker criterion strengthen the measurement validity of the proposed PLS-SEM model and provide evidence that the relationships examined in the structural model do not merely result from excessive conceptual overlap among the constructs. Combined with the HTMT results, these findings provide stronger overall evidence supporting the discriminant validity of the model.

Table 3: Discriminant validity – Fornell-Larcker criterion

	AYTON OMY	BIAS	FAIREN ESS	HUMA N RIGHT S	INTENTIO N TO USE AI	SAFE TY	TRANSPA RENCY	TRU ST IN AI
AYTONOMY	0.673							
BIAS	0.313	0.660						
FAIRENESS	0.160	0.133	0.586					
HUMAN RIGHTS	0.285	0.247	0.136	0.747				
INTENTION TO USE AI	0.100	-0.048	0.190	-0.040	0.703			
SAFETY	-0.054	-0.170	0.054	-0.102	0.139	0.740		
TRANSPARENCY	0.096	-0.063	0.186	-0.026	0.451	0.130	0.775	
TRUST IN AI	0.092	-0.203	0.200	-0.073	0.327	0.539	0.316	0.761

4.4 Collinearity Statistics (VIF) – Inner Model

Prior to evaluating the structural relationships of the proposed PLS-SEM model, the potential presence of multicollinearity among the predictor constructs was examined using the Variance Inflation Factor (VIF). Assessing inner multicollinearity is essential because high levels of correlation among predictor variables may affect the stability and interpretation of the estimated path coefficients in the structural model.

The results indicate that all VIF values are particularly low, ranging from 1.000 to 1.196. Specifically, for the paths predicting Trust in AI, the values are as follows: Autonomy → Trust in AI: VIF = 1.196, Bias → Trust in AI: VIF = 1.187, Fairness → Trust in AI: VIF = 1.084, Human Rights → Trust in AI: VIF = 1.138, Safety → Trust in AI: VIF = 1.053, and Transparency → Trust in AI: VIF = 1.070. For the relationship Trust in AI → Intention to Use AI, the VIF value is 1.000 (Table 4).

All values are substantially below the commonly accepted thresholds for assessing multicollinearity. Even when applying the more stringent criterion of $VIF < 3.3$, none of the examined relationships shows evidence of problematic collinearity.

The highest VIF value is observed for the Autonomy → Trust in AI relationship (VIF = 1.196), followed by Bias → Trust in AI (VIF = 1.187) (Table 4). However, even these values are very close to 1 and substantially below the critical thresholds. Therefore, there is no evidence of substantial overlap among the six ethical dimensions when considered as predictors of Trust in AI.

In particular, the low level of collinearity among Human Rights, Autonomy, Bias, Fairness, Transparency, and Safety is important for the proposed theoretical model, as it suggests that these individual ethical dimensions can be simultaneously examined as distinct predictors of Trust in AI without the estimated path coefficients being substantially affected by multicollinearity.

The VIF value of 1.000 for the Trust in AI → Intention to Use AI relationship is also expected, since Trust in AI is the only direct predictor of Intention to Use AI in this particular part of the structural model. Therefore, there is no competition with other predictor variables in explaining this endogenous construct.

Overall, the VIF results confirm the absence of problematic multicollinearity within the inner structural model. Consequently, the relevant prerequisite for the subsequent evaluation of the structural model is satisfied, and the path coefficients can be examined without substantial concerns regarding collinearity among the predictor constructs.

The ranking of the VIF coefficients is as follows:

Autonomy (1.196) > Bias (1.187) > Human Rights (1.138) > Fairness (1.084) > Transparency (1.070) > Safety (1.053) (Table 4).

Nevertheless, all values are very low in practical terms. This further supports the argument that the six AI Ethical Dimensions can function as distinct predictors of Trust in AI.

There is also an interesting connection with the preceding results: Safety has the lowest VIF among the six predictors (1.053), while in the Total Effects analysis it demonstrated the strongest positive effect on Trust in AI ($\beta = 0.482$) (Table 4). Therefore, the strong effect of Safety does not appear to be attributable to multicollinearity with the other ethical dimensions.

Table 4: Collinearity statistics (VIF)-Inner model –List

VARIANCE INFLATION FACTOR (VIF)	VIF
AYTONOMY → TRUST IN AI	1.196
BIAS →TRUST IN AI	1.187
FAIRENESS →TRUST IN AI	1.084
HUMAN RIGHTS→TRUST IN AI	1.138
SAFETY →TRUST IN AI	1.053
TRANSPARENCY →TRUST IN AI	1.070
TRUST IN AI → INTENTION TO USE AI	1.000

4.5 Interpretation of Total Effects

The Total Effects table presents the overall effects of the constructs Autonomy, Bias, Fairness, Human Rights, Safety, and Transparency on Trust in AI and Intention to Use AI, as well as the total effect of Trust in AI on intention to use. Total effects include both direct and indirect effects operating through the mediating variables of the model.

The results initially identify Safety as the strongest positive determinant of Trust in AI, with a total effect of $\beta = 0.482$ (Table 5). This finding suggests that the more students perceive AI systems as safe, reliable, and protected against potential risks and failures, the greater their trust in these systems. Safety exhibits a clearly stronger effect on Trust in AI compared with the other ethical dimensions.

The second strongest positive effect on Trust in AI is observed for Transparency ($\beta = 0.203$). This result suggests that the ability to understand how AI systems operate and make decisions may contribute positively to building trust. Fairness ($\beta = 0.141$) and Autonomy ($\beta = 0.137$) follow, demonstrating positive but clearly weaker effects on Trust in AI (Table 5).

In contrast, Bias exhibits a negative total effect on Trust in AI ($\beta = -0.161$) (Table 5). This direction is theoretically expected, provided that higher scores on the construct represent greater perceived bias. Therefore, the more students perceive AI systems as potentially producing biased or prejudiced decisions, the lower their trust in these systems tends to be.

Of particular interest is the Human Rights dimension, which exhibits a very small negative effect on Trust in AI ($\beta = -0.037$) (Table 5). The magnitude of this coefficient is extremely low, suggesting that, within the specific model, this construct has only a limited overall relationship with trust. However, the negative direction should not be substantively interpreted before examining the bootstrapping results and, in particular, the statistical significance of this relationship.

Regarding Intention to Use AI, Trust in AI exhibits a positive total effect of $\beta = 0.327$ (Table 5). This finding is particularly important for the theoretical rationale of the proposed model, as it indicates that higher levels of trust in AI systems are associated with a stronger intention to use them. In other words, trust appears to represent an important mechanism through which students' perceptions may translate into behavioral intention.

Among the six ethical dimensions, Safety also exhibits the strongest total effect on Intention to Use AI ($\beta = 0.157$). This is followed by Transparency ($\beta = 0.066$), Fairness ($\beta = 0.046$), and Autonomy ($\beta = 0.045$). Bias demonstrates a negative total effect on intention to use ($\beta = -0.053$), whereas Human Rights exhibits a very small negative effect ($\beta = -0.012$) (Table 5).

The effects on Trust in AI, ranked according to their absolute magnitude, are as follows:

Safety (0.482) > Transparency (0.203) > Bias (-0.161) > Fairness (0.141) > Autonomy (0.137) > Human Rights (-0.037). Similarly, regarding Intention to Use AI, the total effects of the ethical dimensions are:

Safety (0.157) > Transparency (0.066) > Bias (-0.053) > Fairness (0.046) > Autonomy (0.045) > Human Rights (-0.012), while Trust in AI exhibits a total effect of $\beta = 0.327$ on Intention to Use AI (Table 5).

The results therefore reveal an interesting hierarchy among the ethical factors. Safety emerges as the most influential dimension in building Trust in AI and, through trust, in strengthening intention to use AI. Transparency represents the second strongest positive factor, whereas Bias operates in the opposite direction, reducing trust and, consequently, intention to use AI.

Overall, the findings appear to support the fundamental theoretical logic of the proposed model: AI Ethical Dimensions $\rightarrow$ Trust in AI $\rightarrow$ Intention to Use AI. In particular, the pathway Safety $\rightarrow$ Trust in AI $\rightarrow$ Intention to Use AI exhibits the largest overall positive effect among the ethical dimensions examined, highlighting safety as a potentially pivotal factor in fostering trust and AI acceptance among students.

An additional noteworthy pattern emerges from the results. Given that Safety $\rightarrow$ Trust in AI = 0.482 and Trust in AI $\rightarrow$ Intention to Use AI = 0.327, the corresponding indirect effect is approximately: $0.482 \times 0.327 \approx 0.157$.

which is essentially identical to the Safety $\rightarrow$ Intention to Use AI = 0.157 value reported in the Total Effects table. This provides strong evidence that, within the specified model, the effect of Safety on Intention to Use AI is transmitted through Trust in AI. A similar pattern appears to characterize the other ethical dimensions.

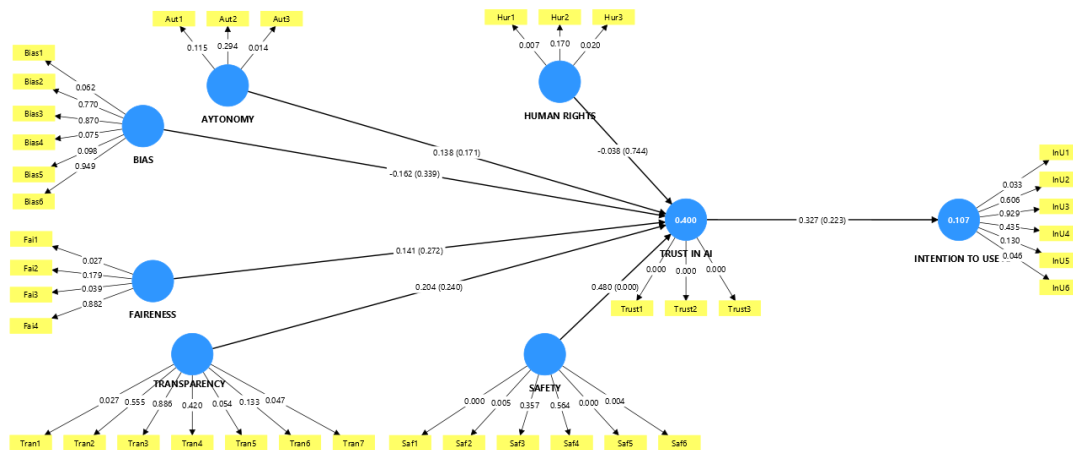
Table 5: Total effects

RELATIONS	Total effects
AYTONOMY → INTENTION TO USE AI	0.045
AYTONOMY → TRUST IN AI	0.137
BIAS → INTENTION TO USE AI	-0.053
BIAS → TRUST IN AI	-0.161
FAIRENESS → INTENTION TO USE AI	0.046
FAIRENESS → TRUST IN AI	0.141
HUMAN RIGHTS → INTENTION TO USE AI	-0.012
HUMAN RIGHTS → TRUST IN AI	-0.037
SAFETY → INTENTION TO USE AI	0.157
SAFETY → TRUST IN AI	0.482
TRANSPARENCY → INTENTION TO USE AI	0.066
TRANSPARENCY → TRUST IN AI	0.203
TRUST IN AI → INTENTION TO USE AI	0.327

4.6 Interpretation of the PLS-SEM Structural Model

Figure 2 presents the bootstrapping results for the proposed PLS-SEM model, in which the six ethical dimensions, namely Human Rights, Autonomy, Bias, Fairness, Transparency, and Safety, are examined as determinants of Trust in AI, while Trust in AI is examined as a determinant of Intention to Use AI.

Using $p < 0.05$ as the criterion for statistical significance, among the six ethical dimensions, only Safety has a statistically significant effect on Trust in AI ($p < 0.001$). In contrast, the relationships Autonomy → Trust in AI ($p = 0.208$), Bias → Trust in AI ($p = 0.344$), Fairness → Trust in AI ($p = 0.305$), Transparency → Trust in AI ($p = 0.220$), and Human Rights → Trust in AI ($p = 0.745$) are not statistically significant.

**Figure 2:** Structural model PLS-SEM

This finding identifies Safety as the only one of the six AI Ethical Dimensions that makes a statistically significant contribution to the prediction of Trust in AI. Combined with the coefficients reported in the previous table, where the effect of Safety was positive and relatively strong ($\beta \approx 0.48$), the results indicate that perceived safety constitutes the most important factor in building trust in AI within the proposed model.

The relationship between Trust in AI → Intention to Use AI is also particularly noteworthy. In the new figure, $p = 0.208$, indicating that the effect is not statistically significant at the $\alpha = 0.05$ level. Although the previous table shows that the path coefficient is positive ($\beta = 0.327$), there is insufficient statistical evidence to support the conclusion that Trust in AI significantly predicts Intention to Use AI in the specific sample.

4.7 Testing the Research Hypotheses

Regarding H6, the results provide sufficient statistical evidence to reject H_{06} and support H_{16} . In contrast, for H1–H5 and H7, the findings do not provide sufficient statistical evidence to reject the corresponding null hypotheses.

Table 6: Testing of the Research Hypotheses

Hypothesis	Structural Relationship	p-value	Decision
H1	Human Rights → Trust in AI	0.745	Not supported
H2	Autonomy → Trust in AI	0.208	Not supported
H3	Bias → Trust in AI	0.344	Not supported
H4	Fairness → Trust in AI	0.305	Not supported
H5	Transparency → Trust in AI	0.220	Not supported
H6	Safety → Trust in AI	<0.001	Supported
H7	Trust in AI → Intention to Use AI	0.208	Not supported

4.8 Interpretation of R^2 Values

The model yields an R^2 value of 0.400 for Trust in AI. This indicates that the six ethical dimensions jointly explain 40.0% of the variance in Trust in AI. This represents a meaningful level of explanatory power, although, at the level of the individual predictors, only Safety demonstrates a statistically significant effect.

For Intention to Use AI, the R^2 value is 0.107, indicating that Trust in AI explains 10.7% of the variance in Intention to Use AI. This relatively limited proportion of explained variance suggests that intention to use AI is likely to be influenced by additional factors that have not been incorporated into the present model.

Regarding the mediation hypothesis H8: Safety → Trust in AI → Intention to Use AI, the results show that although the first path, Safety → Trust in AI, is clearly statistically significant ($p < 0.001$), the subsequent path, Trust in AI → Intention to Use AI, is not statistically significant ($p = 0.208$).

5. Conclusions

The present study sought to approach the acceptance of Artificial Intelligence in Higher Education from an ethically oriented and human-centered perspective, investigating the contribution of six key dimensions of AI Ethics, namely Human Rights, Autonomy, Bias, Fairness, Transparency, and Safety, to the formation of Trust in AI and, subsequently, Intention to Use AI. This approach shifts the focus from the exclusive examination of traditional technology acceptance factors toward the role that ethical perceptions play in building trust in AI systems.

The most important finding of the study is that, among the six ethical dimensions examined, Safety emerged as the only dimension with a statistically significant effect on Trust in AI. Moreover, the six ethical dimensions jointly explain 40% of the variance in Trust in AI ($R^2 = 0.400$), indicating that the ethical evaluation of AI systems constitutes a meaningful component of the trust-building process.

The prominent role of Safety suggests that, within the specific student sample, trust in AI is primarily associated with the perception that AI systems are safe, reliable, controllable, and capable of mitigating potential risks. In contrast, the constructs of Human Rights, Autonomy, Bias, Fairness, and Transparency, although representing fundamental normative principles of responsible and human-centered AI, did not demonstrate statistically significant independent effects on Trust in AI.

This finding has particular theoretical significance, as it highlights the distinction between the normative importance of an ethical principle and its empirical ability to predict users' trust. The absence of statistically significant effects for the remaining ethical dimensions does not diminish their importance; rather, it suggests that students may evaluate broad ethical principles differently from more immediate and experientially salient considerations, such as the safety and reliability of an AI system.

At the same time, Trust in AI exhibited a positive but statistically non-significant effect on Intention to Use AI, while accounting for 10.7% of its variance ($R^2 = 0.107$). This finding suggests that trust, although conceptually important, is not sufficient on its own to explain students' intention to use AI. Future models could therefore incorporate additional technological, educational, and individual-level factors to provide a more comprehensive explanation of AI adoption intentions.

Overall, the study demonstrates that the responsible integration of AI into Higher Education cannot be based solely on technological effectiveness. It requires the development of an environment in which students perceive AI as safe, reliable, transparent, fair, and human-centered. From this perspective, the critical question shifts from "Can students use AI?" to the more substantive question, "Under what ethical conditions can students trust AI?" Within the proposed model, the Safety construct appears to represent the first and strongest step toward building such trust.

References

- Ahmad, M., Ahmed, I. & Jeon, G. (2025). A sustainable advanced artificial intelligence-based framework for analysis of COVID-19 spread. *Environ Dev Sustain* 27, 24783–24798 (2025).
<https://doi.org/10.1007/s10668-022-02584-0>.
- Anastasiadou, S. (2011). Reliability and validity testing of a new scale for monitoring attitudes toward learning statistics with technology. *Acta Didactica Napocensia*, 4(1), 1–10. <http://adn.teaching.ro/>.
- Anastasiadou, S. 2019. Comparison of contemporary advanced statistical methods regarding construct validity evaluation of TEIQue-SF instrument: Statistical Implicative Analysis vs. Principal Components Analysis. 9ème Colloque International sur Analyse Statistique Implicative (ASI 10). Belfort – France. pp. 148-163.
- Anastasiadou, S. D. (2026a). H-IoT Ethics Instrument (H-IoTEI): Development, Validity, and Reliability. In S. Anastasiadou, L. Seremeti, & S. Papalexandris (Eds.), *Bioethics and Legal Aspects of AI in Healthcare and Life Sciences* (pp. 225-236). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3373-5646-4.ch010> Scopus
- Anastasiadou, S. D. (2026b). Attitudes Toward AI, Neuroscience, and Bio(ethical) Dilemmas: Development, Validity, and Reliability. In S. Anastasiadou, L. Seremeti, & S. Papalexandris (Eds.), *Bioethics and Legal Aspects of AI in Healthcare and Life Sciences* (pp. 161-174). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3373-5646-4.ch007>
- Anastasiadou, S. (2026c). Exploring AI ethics through factorial correspondence analysis. In S. D. Anastasiadou, T. G. Chadjipadelis, N. Koutsoupas, A. Markos, & G. Menexes (Eds.), *Modern International Developments in Data Analysis and Multivariate Statistical Analysis* (pp. 495–528). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3693-9400-7.ch016>
- Anastasiadou, S., & Anastasiadis, L. (2019). Quality assurance in education in the light of the effectiveness of transformational school leadership. In N. Sykianakis, P. Polychronidou, & A. Karasavoglou (Eds.), *Economic and financial challenges for Eastern Europe* (pp. 323–344). Springer. https://doi.org/10.1007/978-3-030-12169-3_21
- Attard-Frost, A. De los Ríos, D.R. (2022). Walters, The ethics of AI business practices: a review of 47 AI ethics guidelines, *AI Ethics*, <https://doi.org/10.1007/s43681-022-00156-6>.
- Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023). Transparency and explainability of AI systems: From ethical guidelines to requirements. *Information and Software Technology*, 159, 107197. <https://doi.org/10.1016/j.infsof.2023.107197>
- Beauchamp, T.L., Childress, J.F. (2001): *Principles of biomedical ethics*. Oxford University Press, Oxford (2001)
- Beil, M., Proft, I., van Heerden, D., Sviri, S. & van Heerden, P.V. (2019). Ethical Considerations about Artificial Intelligence for Prognostication in Intensive Care. *Intensive Care Med. Exp.* 2019, 7, 70.
- Bertino, E. Merrill, S. Nesen, A. & C. Utz, C. (2019). Redefining Data Transparency: A Multidimensional Approach. *Computer*, vol. 52, no. 1, pp. 16-26, Jan. 2019, doi: 10.1109/MC.2018.2890190.
- Bormida, Marina Da. (2021). The Big Data World: Benefits, Threats and Ethical Challenges. *Ethical Issues in Covert, Security and Surveillance Research*, (2021): 8, 71–91. doi:10.1108/S2398-601820210000008007.

- Calvo, R. A., Peters, D., Vold, K., and Ryan, R. M. (2020). Supporting Human Autonomy in AI Systems: A Framework for Ethical Enquiry, in *Ethics of Digital Well-Being: A Multidisciplinary Approach*. Editors C. Burr and L. Floridi (Cham: Springer), 31–54. doi:10.1007/978-3-030-50585-1_2.
- Catena, E. (2026). AI and human autonomy: a literature review. *AI Ethics* 6, 126. <https://doi.org/10.1007/s43681-025-00958-4>.
- Ching, T., Himmelstein, DS., Beaulieu-Jones, BK., Kalinin, AA., Do BT., Way, GP., et al. (2018). Opportunities and obstacles for deep learning in biology and medicine. *J R Soc Interface*. 2018;15(141):20170387. <https://doi.org/10.1098/rsif.2017.0387>.
- Colburn, B. (2022). *The Routledge Handbook of Autonomy*. Routledge, London.
- Council of Europe (2019). Commissioner for Human Rights: Unboxing Artificial Intelligence: 10 Steps to Protect Human Rights.
- Danks, D. & London, A. (2017). Algorithmic Bias in Autonomous Systems, *IJCAI*.
- Ferrier, P.M. (2021). Detecting origin fraud with trade data: the case of U.S. honey imports. *Australian Journal of Agricultural and Resource Economics*, Special Issue - Food Fraud, <https://doi.org/10.1111/1467-8489.12412>.
- Ferrer Aran, X., van Nuenen, T., Such, J., Coté, M., & Criado Pacheco, N. (2023). (Accepted/In press). Bias and Discrimination in AI: a cross-disciplinary perspective. *IEEE Technology and society magazine*, 40(2), 72. <https://doi.org/10.1109/MTS.2021.3056293>
- Floridi, L., Cowls J, Beltrametti M, et al. (2018). AI4 people—an ethical framework for a good AI society: opportunities, risks, principles, and recommendations. *Minds Mach*. 2018, 28:689–707. doi: 10.1007/s11023-018-9482-5.
- Goel, P., Patel, R., Garg, D. & A. Ganatra, A. (2021). A Review on Big Data: Privacy and Security Challenges, 3rd International Conference on Signal Processing and Communication (ICSPC), Coimbatore, India, 2021, pp. 705-709, doi: 10.1109/ICSPC51351.2021.9451749.
- Grote, T. & Berens, P. (2020). On the Ethics of Algorithmic Decision-Making in Healthcare. *J. Med. Ethics*, 46, 205–211.
- Haresamudram, K., Larsson, S. & Heintz, F. (2023). Three Levels of AI Transparency. *Computer*, 56(2), 93-100. <https://doi.org/10.1109/MC.2022.3213181899>.
- Hogan, L.; Lasek-Markey, M. (2024). Towards a Human Rights-Based Approach to Ethical AI Governance in Europe. *Philosophies* 2024, 9, 181. <https://doi.org/10.3390/philosophies9060181>.
- Howard, A. & Borenstein, J. (2018). The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Sci Eng Ethics*, 24(5):1521–36. <https://doi.org/10.1007/s11948-017-9975-2>.
- Jiang, T., Gradus, J. L., Lash, T. L., & Fox, M. P. (2021). Jiang et al. respond to “Quantitative bias analysis.” *American Journal of Epidemiology*, 190(9), 1844–1845. <https://doi.org/10.1093/aje/kwab009>.
- Karachristos, C., Seremeti, L., & Anastasiadou, S. (2026). Exploring Student Attitudes Towards Artificial Intelligence. In S. Anastasiadou, T. Chadjipadelis, N. Koutsoupias, A. Markos, & G. Menexes (Eds.), *Advanced Data Analysis Techniques, Theory, and Applications* (pp. 199-216). IGI Global Scientific Publishing. <https://doi.org/10.4018/979-8-3373-6926-6.ch008>.
- Kanellos, L. I. (2020). *Applications of Artificial Intelligence in Law and Judicial Practice*. Nomiki Bibliothiki Publications.
- Kousouni-Pantazopoulou, A. (2022). *Cloud Computing & Legal Issues*. Nomiki Bibliothiki Publications.
- Laitinen, A & Sahlgren, O. (2021) AI Systems and Respect for Human Autonomy. *Front. Artif. Intell.* 4:705164. doi: 10.3389/frai.2021.705164.
- Li, F., Ruijs, N. & Lu, Y. (2023). Ethics & AI: A Systematic Review on Ethical Concerns and Related Strategies for Designing with AI in Healthcare. *AI* 2023: 4, 28–53. <https://doi.org/10.3390/ai4010003>.
- Mantelero, A., Esposito, M.S. (2021). An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems. *Comput. Law Secur.Rev.* 41, 105561 <https://doi.org/10.1016/j.clsr.2021.105561>.

- Murphy, K., Di Ruggiero, E., Upshur, R., Willison D.J., Malhotra, N., Jia Ce Cai1, Malhotra, N., Lui, V. & Gibson, J. (2021). Artificial intelligence for good health: a scoping review of the ethics literature. *BMC Med Ethics*, 22:14.
- O’Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown.
- Panagopoulou-Koutnatzi, F. (2023). *Artificial Intelligence: The Road Towards Digital Constitutionalism*. Papazisis Publishers.
- Puri, M. & S. Gochhait, S. (2024). Data Security in Healthcare: Enhancing the Safety of Data with CyberSecurity. 8th International Conference on Communication and Electronics Systems (ICCES), Coimbatore, India, 2023, pp. 1779-1783, doi: 10.1109/ICCES57224.2023.10192596.
- Radanliev, P. (2025). AI Ethics: Integrating Transparency, Fairness, and Privacy in AI Development. *Applied Artificial Intelligence*, 39(1). <https://doi.org/10.1080/08839514.2025.2463722>.
- Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. (2028). Ensuring Fairness in Machine Learning to Advance Health Equity. *Ann Intern Med*. 2018 Dec 18;169(12):866-872. doi: 10.7326/M18-1990. Epub 2018 Dec 4. PMID: 30508424; PMCID: PMC6594166.
- Rasoolimanesh, S.M.; Ringle, C.M.; Sarstedt, M.; Olya, H. (2021). The combined use of symmetric and asymmetric approaches: Partial least squares-structural equation modeling and fuzzy-set qualitative comparative analysis. *Int. J. Contemp. Hosp. Manag.* 33, 1571–1592.
- Rubel, A., Clinton, C., and Pham, A. (2021). *Algorithms & Autonomy: The Ethics of Automated Decision Systems*. Cambridge: Cambridge University Press.
- Rudner T. G. & H. Toner, H. (2021). Key concepts in ai safety: an overview,” CSET Issue Briefs, 2021.
- Russell, S. (2021). *Human Compatible: Artificial Intelligence and the Problem of Control*. Travlos Publications.
- Salhab, W., Ameued, D., Jaafar, F., Mcheick, H. (2024). A Systematic Literature Review on AI Safety: Identifying Trends, Challenges and Future Directions. *IEEE Access*, Volume 12, 2024, pp.1-26.
- Seremeti, L. & Anastasiadou, S. (2025). The transformative impact of artificial intelligence (AI) on strategic decision-making. In A. Kavoura, E. Kefallonitis, & P. Theodoridis (Eds.), *Strategic innovative marketing and tourism* (pp. 427–435). Springer. https://doi.org/10.1007/978-3-031-81962-9_47
- Seremeti, L, Anastasiadou, S., Masouras, A., Papalexandris, S. (2025). Addressing the challenges of harmonizing law and artificial intelligence technology in modern society. *IAES International Journal of Artificial Intelligence (IJ-AI)*, Vol. 14, No. 3, June 2025, pp. 2471~2478, ISSN: 2252-8938, DOI: 10.11591/ijai.v14.i5.pp2471-2478-2471, Journal homepage: <http://ijai.iaescore.com>
- Stahl, B. C., Brooks, L., Hatzakis, T., Santiago, N., & Wright, D. (2023). Exploring ethics and human rights in artificial intelligence – A Delphi study. *Technological Forecasting and Social Change*, 191, 122502. <https://doi.org/10.1016/j.techfore.2023.122502>.
- Taraza, E. & Anastasiadou S. (2019). EFQM Excellence Model in Vocational Lyceum: Reliability and Validity of EFQM Instrument. *Proceedings of 13th annual International Technology, Education and Development Conference (INTED2019)*, Valencia, Spain, pp. 2273-2285. doi: 10.21125/inted.2019.0637
- Thapa, K. B., Okalidou, A., & Anastasiadou, S. (2016). Teachers’ screening estimations of speech–language impairments. *International Journal of Language & Communication Disorders*, 51(3), 310–327.
- Tsinorema, S. (2018). Consent in Biomedical and Clinical Research: Epistemological and Ethical Issues. Presentation at the 5th Conference on Medical Liability and Bioethics: Health Data and Genetic Data, Nomiki Bibliothiki, Athens, Greece, January 19, 2018.
- Vellido, A. (2019). Societal Issues Concerning the Application of Artificial Intelligence in Medicine. *Kidney Dis*. 5, 11–17.
- Verghese, A., Shah, NH., Harrington, RA. (2018). What this computer needs is a physician: humanism and artificial intelligence. *JAMA*. 2018;319(1):19. <https://doi.org/10.1001/jama.2017.19198>.
- Yeung, K.; Howes, A.; Pogrebná, G. (2020). AI Governance by Human Rights–Centered Design, Deliberation, and Oversight. In *The Oxford Handbook of Ethics of AI*; Oxford University Press: Oxford, UK, 2020; pp. 77–106.

- Yeng, P.K., Yang, B., Snekenes, E.A. (2019). Observational Measures for Effective Profiling of Healthcare Staffs' Security Practices. In Proceedings of the 2019 IEEE 43rd Annual Computer Software and Applications Conference (COMPSAC), Milwaukee, WI, USA, Volume 2, pp. 397–404.
- Yuste, R., Goering, S., Arcas, BAY, Bi, G., Carmena, JM, Carter, A. et al. (2017). Four ethical priorities for neurotechnologies and AI. *Nature*. 2017;551(7679):159–63.
- Wang, Y., Xiao, J., Wei, Z., Zheng, Y., Tang K.-T. and Chang, C. H. (2024). Security and Functional Safety for AI in Embedded Automotive System—A Tutorial. *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 71, no. 3, pp. 1701-1707, March 2024, doi: 10.1109/TCSII.2023.3334273.
- Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Ue sato, J., Huang, P.-S., Cheng, M., M. Glaese, M., Balle, B., Kasirzadeh, A. et al., (2021). Ethical and social risks of harm from language models, arXiv preprint arXiv:2112.04359.
- Zhuo, Terry Yue, Yujin Huang, Chunyang Chen and Zhenchang Xing. (2023). Exploring AI Ethics of ChatGPT: A Diagnostic Analysis. ArXiv abs/2301.12867. Doi: 10.48550/arXiv.2301.12867Corpus ID: 256390238.