JOURNAL OF
Emerging Paradigms
in Computing Systems
JEPCS

Journal of Emerging Paradigms in Computing Systems (JEPCS) | Volume 01, Issue No. 01, 2025

ARTICLE TYPE: Review Article

Article ID: JEPCS-2025-0101-0019

Ethical Implications of Artificial Intelligence: A Review, Taxonomy, Governance Framework, and Future Research Agenda

Soban Ahmad[1], Zupash Awais[2,*]

[1] School of Systems and Technology, University of Management and Technology, Lahore, Pakistan

[2] Faculty of Information Technology & Computer Science, University of Central Punjab, Lahore, Pakistan

* Corresponding author: zupash.awais@ucp.edu.pk

Received: September 29, 2024 **Accepted:** December 05, 2024

Abstract- Artifical Intelligence (AI) is used in routine consumer services as well as in systems that support institutional decisions. Health care, education, business, transport, communication, robotics, and public services are some core applications of AI. These applications create ethical questions because they may process personal information or influence decisions that affect people. The discussion therefore includes privacy, fairness, accountability, security, human control, and changes in employment. In this paper, we starts from a keyword-based Systematic Mapping Study (SMS). We searched different databases like IEEE Xplore, ACM Digital Library, Scopus, Web of Science, and ProQuest. In result of searching 1,062 papers returned. After that screening reduced the set to 83 academic papers, and 37 recurring keywords were reported. The keywords were classified into nine categories covering conceptual research, robotics, general philosophical and ethical research, AI-specific philosophical and ethical research, law and regulation, autonomous vehicles, artificial general intelligence and AI risk, human cognition, and technology-oriented research. Selected papers published from 2021 to 2024 are therefore used to update the discussion on trustworthy AI, generative AI, risk management, security, and regulation. Top-down and bottom-up machine ethics, digital policy, data governance, and ethical questions arising during AI development and operation are also reviewed.

Keywords- Artificial Intelligence Ethics; Trustworthy AI; AI Governance; Generative AI; Systematic Mapping Study.

1. Introduction

AI methods perform tasks such as classification, prediction, speech and image recognition, recommendation, planning, and decision support [1]-[3]. Their role depends on the type of application. A model may work in the background, provide information to a user, or contribute to the control of a physical system. These forms of AI are found in online services, health care, education, transport, communication networks, and industry. Accuracy alone does not describe the effect of an AI application on the people who use it or are affected by it. Personal information may be required as input, and the resulting output may influence a decision about a person or group. Questions then arise about the source of the data, unequal treatment, explanation of the result, and responsibility for its use [4]-[6]. Problems like fairness, privacy, transparency, accountability, safety, human control, and responsibility are discussed in the AI-ethics literature. Several of these problems originate in normal development decisions. Training data determine from which cases and groups the model learns. The interface determines what users are told about an automated result. Testing can also reveal differences in error across groups. Where an incorrect result has serious consequences, human review may be required before the output is acted upon. Responsible-AI research considers these questions during design, testing, and deployment [7], [8]. Law introduces requirements that are different from voluntary ethical principles. Robles Carrillo [9] examines the

relationship between AI ethics and legal regulation. Pagallo et al. [10] discuss governance, data protection, and AI. Their work covers issues such as liability, permitted data use, and restrictions placed on particular applications. Renda [11] considers AI within the wider setting of digital policy and governance. International guidance provides another source of requirements. UNESCO's Recommendation on the Ethics of Artificial Intelligence [12] covers human rights, fairness, transparency, human oversight, and governance. NIST [13] approaches AI from the perspective of risk management. The European Union Artificial Intelligence Act [14] provides legal requirements that vary according to the type and use of the AI system. Vakkuri and Abrahamsson [15] studied the earlier AI-ethics literature using a keyword-based SMS. The initial result contained 1,062 papers. After screening, Following screening, 83 academic papers were retained for the keyword analysis. From these papers, 37 recurring keywords were identified. The 37 keywords were placed into nine categories: conceptual research, robotics, general philosophical and ethical research, AI-specific philosophical and ethical research, law and regulation, autonomous vehicles, artificial general intelligence and AI risk, human cognition, and technology-oriented research [15]. The present paper uses these nine categories as the main structure for reviewing the earlier AI-ethics literature. The SMS was published in 2018. AI systems have changed considerably since then, particularly with the wider use of large language models and generative AI. These systems have brought greater attention to training data, generated content, misuse, large-scale deployment, and the difficulty of controlling model behavior. Research after the original mapping has examined harm at different stages of the machine-learning lifecycle [16]. Large language models have been studied in relation to training data, representation, misuse, and other risks [17], [18]. Trustworthy-AI research has also developed further [19]. Studies published after the introduction of ChatGPT examine the use and implications of generative AI in research, education, and other settings [20]-[22]. Recent work has also considered the relationship between trustworthiness and regulation [23] and the security, privacy, and robustness of trustworthy AI systems [24]. The NIST framework [13] and the EU Artificial Intelligence Act [14] form part of the same recent governance discussion. This paper uses the earlier nine-category taxonomy as its starting point and then considers how the discussion has changed in more recent work. The first part reviews the ethical concepts reported in the SMS. Selected literature from 2021 to 2024 is then used to discuss developments related to large language models, generative AI, trustworthy AI, security, risk management, and regulation. The newer studies are treated as an update and are not included in the original SMS counts. The later publications are treated as an update and are not included in the original figures of 1,062 retrieved papers, 83 retained papers, or 37 recurring keywords.

The remainder of the paper is organized as follows. Section 2 describes the systematic mapping method and the literature covered in this paper. Section 3 introduces the AI application areas needed for the later discussion. Section 3 reviews previous work on AI ethics, responsible design, law, and governance. The nine-category taxonomy is discussed in Section 5. Section 6 examines the main ethical issues, while Section 7 discusses top-down and bottom-up approaches to machine ethics. Digital policy and data governance are covered in Section 8. Section 9 discusses selected developments reported between 2021 and 2024. Section 10 considers ethical issues during AI-system development and deployment. Sections 11 and 12 present the discussion and limitations, Section 13 identifies future research directions, and Section 14 concludes the paper.

2. Related Work

Vakkuri and Abrahamsson [15] provide the taxonomy used in this review. Their keyword-based SMS initially retrieved 1,062 papers and retained 83 academic papers after screening. The authors reported 37 recurring keywords and organized them into nine categories. These categories are examined in Section 5. Applied AI raises ethical questions in different ways. Jameel et al. [78] discuss challenges reported in AI-ethics research. Singh et al. [79] focus on multimedia, where dataset construction, privacy, fairness, transparency, and deployment all require attention. Bechmann and Bowker [40] examine assumptions embedded in social-media AI and the processes through which its data and categories are produced. Other studies are concerned with social consequences. Hager et al. [80] consider AI for social good, Francescato [39] examines social networks and political polarization, and Thieltges et al. [81] discuss ethical questions surrounding bot detection. Taken together, these applications place ethical concerns at several points, including data collection, system design, and deployment. Responsible-AI research asks how these concerns can influence actual system development. Peters et al. [7] propose two design frameworks and apply them in digital mental health. Kuleshov et al. [82] discuss codification of AI ethics. Lo Piano [6] uses practical cases to examine ethical principles in machine learning, and Yu et al. [8] review technical approaches for incorporating ethics into AI systems. The common engineering question is how a general principle becomes a design, evaluation, or governance requirement. Machine ethics raises a different question: how should an artificial system behave when its action has an ethical consequence? Etzioni and Etzioni [83] discuss top-down and bottom-up approaches to this problem. Their analysis also covers driverless cars, ethics bots, and trolley-style scenarios. Kim et al. [84] examine value alignment and distinguish mimetic alignment from anchored alignment. They argue that copying observed human behavior does not by itself ensure ethically acceptable behavior. These approaches are examined further in Section 7. Legal research considers the obligations surrounding the development and use of AI. Robles Carrillo [9] examines the movement from ethical principles toward legal requirements. Pagallo et al. [10] study governance in relation to data protection, AI, and the Web of Data. Renda [11] considers broader AI policy and governance. Policy studies on AI and IoT address the governance of connected digital infrastructure [72], [73], while Gourraud and Simon [85] discuss differences between European and United States approaches to AI and digital policy.

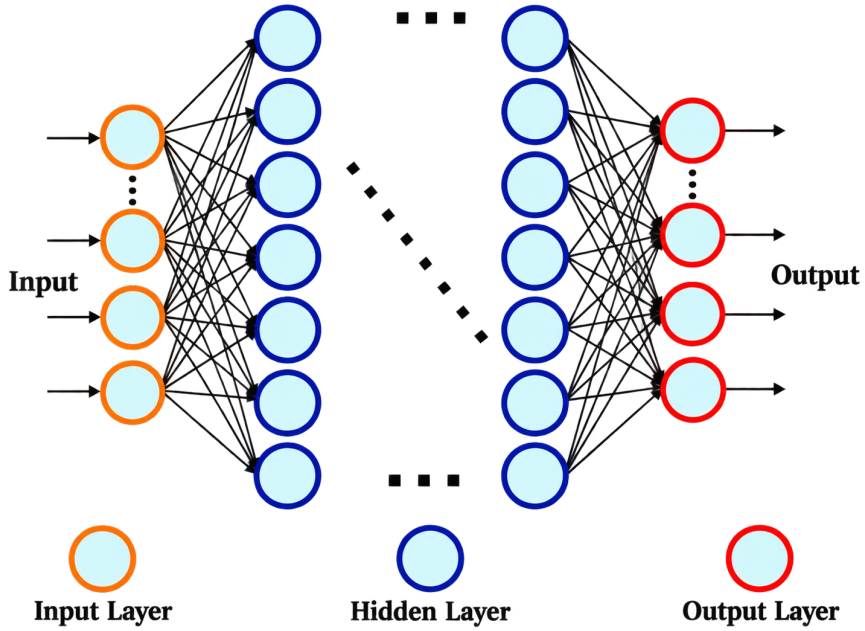

Figure 1: *Simplified neural-network perception model illustrating data-driven AI processing.*

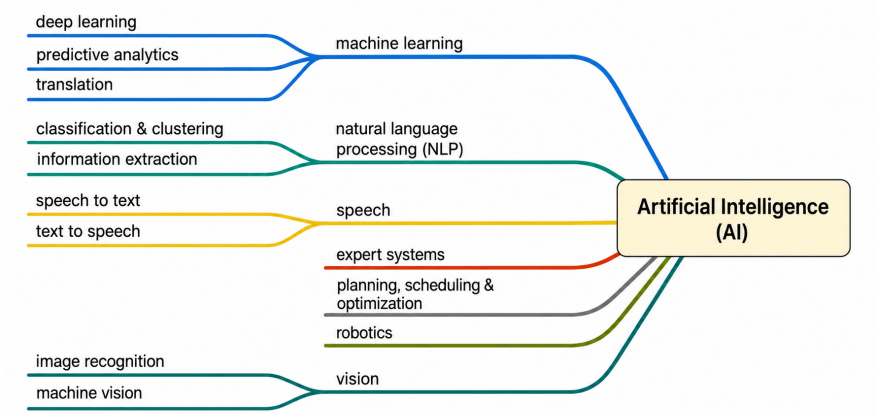

Figure 2: *Major AI areas and representative application families considered in the review.*

Trustworthy AI and the organizational effects of AI form another part of the reviewed literature. Jain et al. [5] consider trustworthiness. Dignum [77] discusses ethical implications within the wider development of intelligent systems. Jarrahi [47] examines how people and AI share decision-making roles at work. Business studies address AI-based transformation [49], [50], while [86]-[88] consider economic, employment, and engineering implications of automation. Technical sources are included where they help explain the systems to which ethical requirements are later applied. They cover DL [29], distributed AI [26], ML for wireless networks [67], health-care applications [51]-[53], and selected energy and engineering applications [57], [58], [62]. Figure 1 separates input, internal processing, and output in a simplified perception model. This distinction is used later when discussing where data- and model-related concerns can arise.

3. AI Application Areas

AI is not based on a single computational method. It includes machine learning (ML), deep learning (DL), expert systems, distributed AI, natural-language processing, planning, optimization, computer vision, and intelligent agents [25]-[27]. ML and DL represent two related but different parts of AI. ML methods learn patterns or decision rules from data while DL methods use multilayer neural networks to learn representations from their inputs [28]-[30]. Kersting [2] discusses the relationship between ML and AI. Russell and Norvig [3] cover a broader set of AI topics, including search, reasoning, planning, and intelligent agents. AI methods are also applied to engineering problems. Examples in the review include power systems [31] and data-based modeling of complex processes [32]. Figure 2 presents the main AI areas considered in this review.

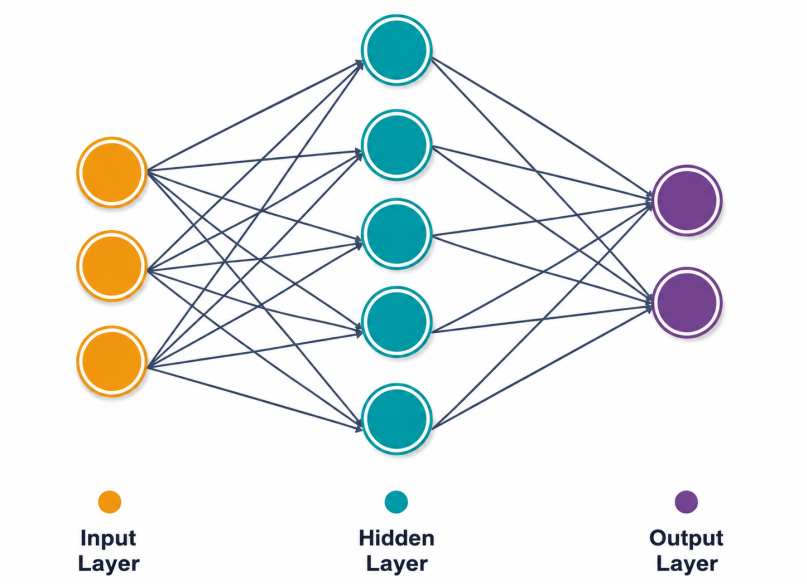

Figure 3: *Illustrative feed-forward neural-network structure with input, hidden, and output layers.*

The methods shown in the figure serve different purposes depending on the problem being solved. Their representative computational functions are summarized in Table 1. It is included as technical background and does not rank the methods.

Table 1: *AI techniques and representative computational functions.*

AI technique	Representative computational function
Intelligent agents	Perceive an environment, exchange information when needed, and select actions according to an internal policy or rule set.
Artificial neural networks	Learn nonlinear relationships for classification, prediction, recognition, and other data-driven tasks.
Expert systems	Apply encoded domain knowledge and inference rules to support reasoning or decisions.
Machine learning	Learn predictive or descriptive patterns from data through supervised, unsupervised, or reinforcement-learning methods.
DL	Learn multilayer representations for tasks such as image, speech, text, and high-dimensional pattern analysis.

A feed-forward neural network is shown in Figure 3. Values enter through the input layer, are transformed through hidden layers, and reach the output layer. The figure is a simple illustration of the layered processing used by many neural-network models [25], [29], [30].

An intelligent agent receives information from its environment and produces an action according to its internal function. Figure 4 shows a generic agent-function process. It distinguishes an agent that receives percepts and selects actions from a feed-forward model that maps inputs directly to outputs [3], [27].

2.1. Impact of AI in daily life

AI is now embedded in many services used at home and online. Smart-home systems are one example. They can use AI for sensing, automation, prediction, and assistance to the user [33]. Other smart-home studies combine AI with Internet-of-Things devices, connected appliances, and data-driven services [34], [35]. Recommender systems use information about users and available items to rank or suggest content [36]. In online shopping, related methods support both product search and customer interaction. Yan et al. [37] study a task-oriented dialogue system for shopping, whereas Li et al. [38] examine AI-based customer service. Social-media systems make extensive use of information produced through user activity. Such information can be processed for classification, content organization, and recommendation. The reliance on behavioral data also creates questions about how personal information is collected and used [39], [40]. Figure 5 gives representative examples of AI encountered in everyday services.

2.3. Education and offices

Educational applications of AI include assessment, prediction of academic performance, distance-learning support, and identification of learning styles [41]-[44]. Chattopadhyay et al. [45] consider AI specifically in school assessment. These

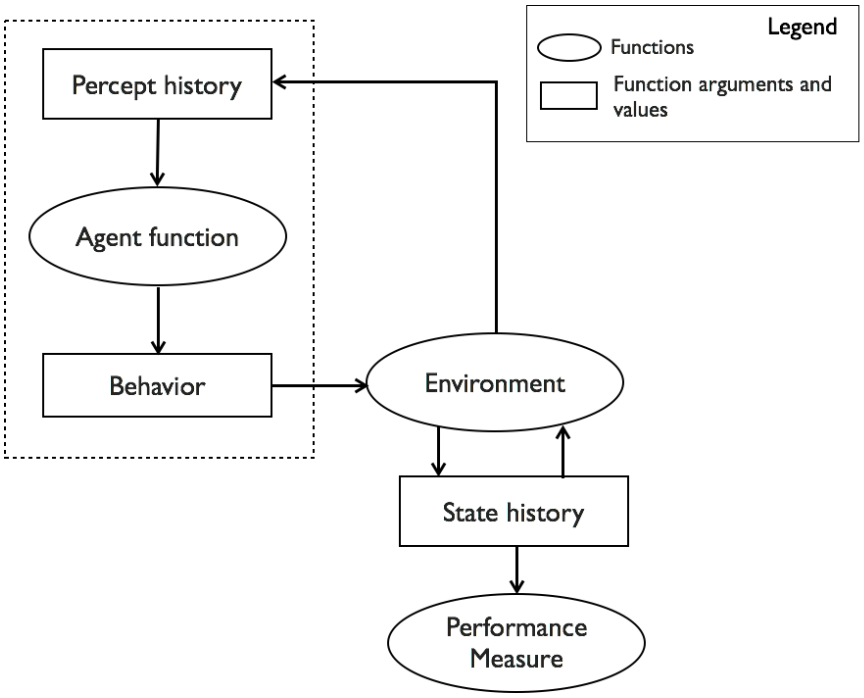

Figure 4: *Generic agent-function process mapping perceived inputs to actions.*

systems can reduce repetitive work for teachers and provide additional information about student performance. Their outputs may also affect how a student is evaluated or supported. Data quality and the interpretation of the output therefore matter when the result contributes to an academic decision. Fairness, privacy, transparency, and human review are relevant in this setting. AI applications in offices range from simple classification tasks to organizational decision support. Email classification assigns incoming messages to predefined categories [46]. Other studies consider workplace information and organizational decision making [47], [48]. At the organizational level, the introduction of AI can also change business processes and employee roles [49], [50].

2.4. Health care

AI has been studied for several clinical applications. The examples reviewed here include prostate-cancer diagnosis, cardiac care, and gastric-cancer detection from medical images [51]-[53]. In these cases, AI provides support for clinical analysis rather than replacing the clinical process itself. The use of AI in medicine also raises questions that are not captured by prediction accuracy alone. Keskinbora [54] discusses these issues from a medical-ethics perspective. Patient data need protection, a model has to be evaluated for the clinical setting in which it will be used, and its limitations need to be understood. Responsibility for diagnosis and care also remains with the health-care professionals and organizations using the system.

2.5. Online shopping and recommender systems

Online shopping is a common setting for AI-based personalization. Recommender systems use available information about users and products to select or rank items that may be relevant to a customer [36]. This can reduce the amount of information that a user has to search manually. AI is also used during direct interaction with customers. Yan et al. [37] developed a task-oriented dialogue approach for online shopping, whereas Li et al. [38] examined AI-based customer service and its relationship with consumer attitude. These applications can make product search and customer support more convenient. Online shopping services often use information from earlier interactions, user preferences, and purchasing activity. This information helps the system produce personalized rankings or recommendations. Its collection and use, however, create practical questions about retention, access, and the information used to generate a recommendation. Users may also need to know when their previous activity affects the products or content shown to them.

2.6. Transport, engineering, energy, and infrastructure

AI methods are used in both transport and engineering applications. Assi et al. [55] compare AI methods for travel-to-school mode choice. Hofmann et al. [56] discuss the use of AI and data science in the automotive industry. Engineering applications cover problems with quite different physical characteristics. AI has been used for improving motor efficiency [57] and for photovoltaic applications [58]. Other studies use AI to estimate hanging-wall stability [59] and

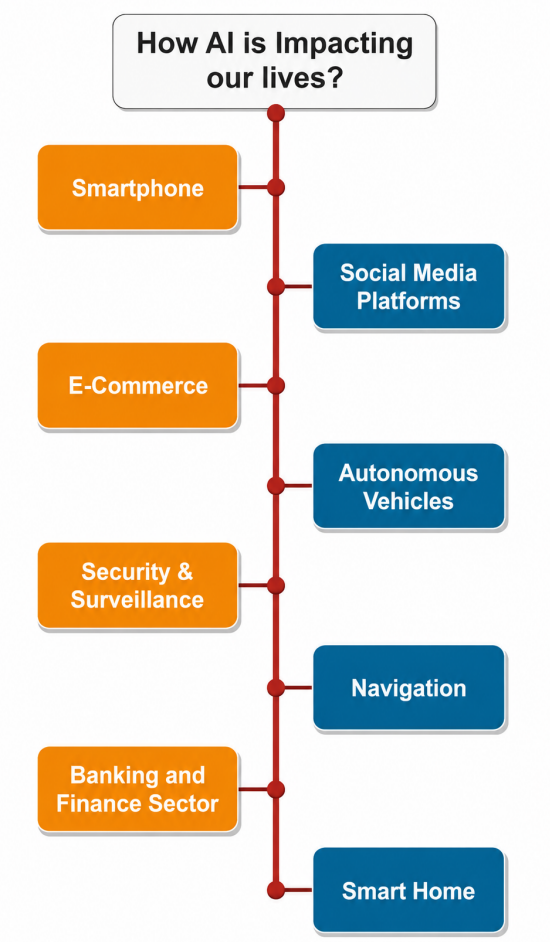

Figure 5: *Representative examples of AI use in everyday life.*

to support assembly design in manufacturing [60]. Prediction is also used in process and energy applications. Examples include coal flotation [61] and hybrid-energy-system optimization [62]. Further applications include the prediction of concrete strength [63], permeability from well-log data [64], micromixing performance [65], and geomechanical properties [66]. Unlike many consumer applications, the output of these models may affect a physical process or an engineering decision. An incorrect prediction can change process operation or lead to an unsuitable engineering choice. In safety-related applications, the consequence may be more serious. Evaluation should therefore include reliability, traceability, and failure behavior in addition to predictive performance.

2.7. Networks, IoT, automotive systems, and smart cities

Communication networks, edge systems, connected vehicles, and smart-city infrastructure provide another group of AI applications. Chen et al. [67] survey neural-network-based machine learning for wireless networks. Wang and Lu [68] discuss AI applications in computer networks, while Debauche et al. [69] consider AI processing at the network edge. Lin et al. [70] use AI-based data analytics in heterogeneous wireless communication. Related applications appear in the automotive sector [56] and in smart-city systems [71]. These systems depend on communication infrastructure as well as computation. Policy studies in [72] and [73] examine the use of AI and IoT from the perspective of technology policy and connected digital services. A networked AI service depends on the timely and correct delivery of its data. Problems in the communication layer can therefore affect the AI service even when the model itself is functioning as intended. Arsalan and Rehman [74] examine interest broadcasting and a timing attack in a Named Software-Defined vehicular network. Burhan et al. [75] review real-time security problems in fog-based IoT applications. Sattar et al. [76] address interest flooding and controlled packet propagation in Named Internet of Vehicles. These cases show that delay, packet flooding, or disruption of communication can affect the information available to a networked application.

2.8. Technical training and ethical awareness

Technical knowledge explains how an AI system is built and how its model produces an output. Ethical assessment asks different questions. A developer may understand the training procedure of a neural network but still need to examine how the training data were obtained. The developer also needs to know whether errors are distributed differently across relevant groups and who remains responsible when the system output is used in a decision. The technical literature explains how machine-learning and deep-learning systems are developed and trained [2], [25], [28]-[30]. Ethical research addresses a different part of the problem. It considers how fairness, privacy, transparency, responsibility, and other requirements can be included in the design and use of AI systems [6]-[8], [77]. Technical and ethical training should therefore complement each other. A development team does not need to treat every engineering decision as a philosophical problem. It does need to recognize when a design choice affects personal data, safety, security, human control, or accountability. Where these issues are outside the team's expertise, the appropriate legal, ethical, security, or domain specialist should be involved.

4. Review Methodology

4.1. Systematic Mapping Study

The nine-category taxonomy used in this review comes from the keyword-based SMS reported by Vakkuri and Abrahamsson [15]. Their study examined the author keywords used in AI-ethics publications. Terms that appeared repeatedly were identified and related terms were grouped into broader categories. The search covered five databases: IEEE Xplore, ACM Digital Library, Scopus, Web of Science, and ProQuest. It retrieved 1,062 papers. After screening, 83 academic papers were retained for keyword analysis. Thirty-seven recurring keywords were then identified and organized into nine categories [15]. The counts reported in [15] are used here as published. They were not recalculated from the original search results. The same nine-category classification is used in Section 5 to organize the taxonomy. Table 2 lists the methodological details reported for the SMS.

Table 2: *Systematic mapping characteristics reported for the SMS [15].*

Element	Reported information
Review type	Systematic Mapping Study (SMS).
Databases	IEEE Xplore, ACM Digital Library, Scopus, Web of Science, and ProQuest.
Initial document pool	1,062 academic documents.
Retained studies	83 documents reported as relevant to the mapping objective.
Recurring terms	37 recurring words/keywords.
Resulting taxonomy	Nine categories covering conceptual, philosophical, legal, risk, cognition, robotics, vehicle, and technology themes.
Primary unit of analysis	Recurring terminology and concepts in the AI-ethics literature.
Unreported details	Complete search strings, database-specific query dates, inter-reviewer agreement, and a full study-selection flow diagram were not reported.

4.2. Screening and classification

The papers retrieved from the five databases were screened before the keyword analysis was carried out [15]. The retained papers were then examined for recurring terms, and related terms were placed into broader categories. This type of keyword analysis has an important limitation. Two authors may use different expressions for nearly the same concept, while an abbreviation and its full form may appear as separate terms. Technical jargon can create the same problem. As a result, the frequency of a term does not necessarily indicate the importance of the ethical issue that it represents. The nine categories are therefore used here to organize the literature, not to rank ethical principles.

4.3. Bibliographic coverage and temporal scope

The SMS provides the taxonomy, but the discussion in this paper also requires literature from the technical and application domains in which AI is used. For this reason, the reference set includes work on AI methods, health care, education, consumer systems, engineering, communication networks, law, policy, and governance. The bibliography contains 94 references. Table 3 identifies the year and primary subject of each source and indicates why it is relevant to the paper. The 94 references should not be interpreted as the study set used in the SMS. The SMS figures of 1,062 retrieved papers and 83 retained papers refer specifically to the study reported in [15].

Table 3: *Bibliographic scope of the 94 references used in the review.*

Ref.	Year	Primary focus	Relevance to the review
[1]	2018	General AI and future roles	General AI perspective
[2]	2018	Relationship between ML and AI	Technical foundations
[3]	2003	Artificial intelligence foundations	Textbook / technical foundations
[4]	2020	Ethics in the era of AI	Ethics perspective
[5]	2020	Trustworthiness of AI	Trustworthy AI
[6]	2020	Ethical principles in ML and AI	Applied AI ethics
[7]	2020	Responsible-AI design frameworks	Ethics-by-design
[8]	2018	Embedding ethics in AI	Machine ethics
[9]	2020	AI ethics and law	Legal governance
[10]	2019	Legal governance and data protection	Law and data governance
[11]	2019	AI policy	Policy and governance
[12]	2021	UNESCO Recommendation on AI ethics	International ethics and policy framework
[13]	2023	NIST AI Risk Management Framework 1.0	Operational AI risk management
[14]	2024	European Union Artificial Intelligence Act	Binding risk-based AI regulation
[15]	2018	Core concepts of AI ethics	Ethics taxonomy
[16]	2021	Sources of harm across the ML lifecycle	Responsible-AI lifecycle / fairness
[17]	2021	Risks of very large language models	Language-model ethics and environmental/social risk
[18]	2022	Taxonomy of language-model risks	Generative/language-model risk taxonomy
[19]	2022	Trustworthy artificial intelligence	Trustworthy AI review and validation
[20]	2023	Generative conversational AI and ChatGPT	Research, practice, policy, and ethics
[21]	2023	Large language models in education	Educational opportunities, risks, and responsible use
[22]	2024	Generative AI	Generative-AI foundations, applications, and risks
[23]	2024	Trustworthy AI and the EU AI Act	Regulation, trust, and acceptable risk
[24]	2024	Security, privacy, and robustness for trustworthy AI	Trustworthy AI security review
[25]	2019	DL and information fusion	Technical methods
[26]	2021	Distributed artificial intelligence	Technical foundations / edited book
[27]	1996	AI and expert systems	Technical foundations
[28]	2016	ML and DL	Technical book
[29]	2020	DL	Technical foundations
[30]	2018	Deep-learning history and outlook	Technical perspective
[31]	1997	AI in power systems	Engineering application
[32]	2016	Data-based modeling of complex processes	Technical application
[33]	2019	Smart-home AI applications	Application review
[34]	2020	AI and IoT applications in smart homes	Systematic content review
[35]	2018	Smart homes and new technologies	Consumer application
[36]	2018	Recommender systems	Book / technical methods
[37]	2017	Task-oriented dialogue for online shopping	E-commerce application
[38]	2020	AI customer service in online shopping	E-commerce application
[39]	2018	AI, social networks, and political polarization	Societal implications
[40]	2019	Knowledge production and social-media AI	Bias and data governance
[41]	2008	Student learning assessment	Education application
[42]	2020	Academic achievement prediction	Education application
[43]	2017	AI augmentation for distance learning	Education application
[44]	2018	AI-based learning-style determination	Education application
[45]	2018	AI for assessment in schools	Education application
[46]	2018	Email classification	Communication application

Continued on next page

Table 3 - continued from previous page

Ref.	Year	Primary focus	Relevance to the review
[47]	2018	Human-AI decision making and work	Organizational and labor effects
[48]	2019	Prediction of knowledge-hiding behavior	Organizational application
[49]	2017	Business of artificial intelligence	Business implications
[50]	2020	AI transformation and firm performance	Business implications
[51]	2020	AI diagnosis and grading of prostate cancer	Health-care application
[52]	2018	AI applications in cardiac care	Health-care application
[53]	2018	CNN-based gastric-cancer detection	Health-care application
[54]	2019	Medical ethics and AI	Health-care ethics
[55]	2019	Travel-to-school mode choice	Transport application
[56]	2017	AI and data science in automotive systems	Automotive application
[57]	2005	Motor-efficiency optimization	Engineering application
[58]	2008	Photovoltaic systems	Energy application review
[59]	2018	AI for hangingwall stability	Engineering application
[60]	2017	Data mining for assembly design	Manufacturing application
[61]	2018	AI models for coal flotation	Engineering application
[62]	2016	AI for hybrid-energy optimization	Energy application
[63]	2019	AI for geopolymer-concrete strength	Civil-engineering application
[64]	2018	AI-based permeability formulation	Energy / engineering application
[65]	2019	AI for micromixing evaluation	Chemical-engineering application
[66]	2017	AI and data mining for shale properties	Geomechanical application
[67]	2019	Neural-network-based ML for wireless networks	Network tutorial
[68]	2019	AI in computer networks	Network application
[69]	2020	Edge AI for climatic enclosure	Edge / environmental application
[70]	2019	AI analytics for wireless communications	Network application
[71]	2019	Big data, AI, and smart cities	Smart-city governance
[72]	2018	Industry 5.0, big data, AI, and IoT policy	Technology policy
[73]	2018	AI and IoT policy opportunities	Technology policy
[74]	2022	Interest broadcasting and timing attacks in IoV	Vehicular-network security and resilience
[75]	2023	Fog-computing IoT security survey	IoT/fog security and architecture
[76]	2024	Interest-flooding mitigation in named IoV	Vehicular-network security and resilience
[77]	2018	AI ethics special-issue perspective	Ethics perspective
[78]	2020	AI ethics challenges and solutions	Ethics and governance
[79]	2020	Legal and ethical multimedia research	Ethics and law
[80]	2019	AI for social good	Societal applications
[81]	2016	Ethics of bot detection	Security and research ethics
[82]	2020	Codification of AI ethics	Ethics and governance
[83]	2017	Incorporating ethics into AI	Machine ethics
[84]	2018	Value alignment in AI	Machine ethics / value alignment
[85]	2020	AI and digital-policy comparison	Digital policy
[86]	2018	Economics of AI	Economic implications
[87]	2017	Automation, engineering, and future jobs	Labor and professional ethics
[88]	2019	Economics of AI and future of work	Economic and labor implications
[89]	2017	AI capability and intelligence grading	Conceptual AI assessment
[90]	2016	Ethical aspects and future AI	Ethics perspective
[91]	2017	Explainable artificial intelligence	Transparency / explainability
[92]	2019	Bottom-up constituent parsing	Technical methods
[93]	2019	Bottom-up unsupervised person re-identification	Computer-vision methods
[94]	2018	AI reproducibility	Research integrity

Table 4 shows the publication-year distribution of the bibliography. Most references were published between 2017 and 2020, while the later material extends the discussion through 2024 for the intended 2025 issue.

Table 4: *Publication-year profile of the 94 references used in the review.*

Year	1996	1997	2003	2005	2008	2016	2017	2018	2019	2020	2021	2022	2023	2024
Entries	1	1	1	1	2	5	10	23	18	16	4	3	4	5

The 2021-2024 material is treated as a targeted update to the 2018 SMS [15]. Suresh and Guttag [16] address sources of harm across the machine-learning lifecycle. Bender et al. [17] and Weidinger et al. [18] examine risks associated with large language models. UNESCO provides an international ethics recommendation [12], while Kaur et al. [19] review trustworthy AI and NIST provides a risk-management framework [13]. Generative conversational AI and large language models are discussed in [20]-[22]. The regulatory update includes work on trustworthiness and the EU AI Act [23], a recent trustworthy-AI security review [24], and the 2024 EU Artificial Intelligence Act itself [14]. References [74]-[76] address IoT and vehicular-network security and are used in the sections on communication-layer resilience.

5. Taxonomy and Main Categories

The mapping study identified 37 recurring keywords and grouped them into nine categories [15]. The categories are conceptual research, robotics, general philosophical and ethical research, AI-specific philosophical and ethical research, law and regulation, autonomous vehicles, AGI and AI risk, human cognition, and technology-oriented research. The categories overlap in practice. A paper on an autonomous vehicle, for example, can also involve law, safety, human preferences, and technical validation. Table 5 summarizes the scope assigned to the nine categories in this review.

Table 5: *Nine-category taxonomy reported in the Systematic Mapping Study [15].*

Category	Scope in the mapped literature
Conceptual	Definitions, terminology, conceptual boundaries, taxonomies, and the relationship among ethics-related terms.
Robotics	Moral and safety questions associated with embodied autonomous machines and human-robot interaction.
General philosophical and ethical	Moral philosophy, social values, rights, responsibility, human welfare, and theories of ethical conduct.
AI-specific philosophical and ethical	Machine ethics, value alignment, autonomous decision making, fairness, explainability, and responsibility in AI systems.
Law and regulation	Liability, rights, legal compliance, data protection, jurisdiction, regulation, and governance mechanisms.
Autonomous vehicles	Safety-critical decision making, responsibility, human preferences, uncertainty, and machine behavior in transport.
AGI and AI risk	Highly autonomous or self-improving systems, loss of human control, catastrophic-risk arguments, and long-term governance.
Human cognition	Common sense, preferences, autonomy, judgment, human values, and whether/how machines can represent them.
Technology-oriented	Data, algorithms, ML, security, transparency, software architecture, validation, and implementation controls.

5.1. Conceptual research

The terminology of AI ethics is not always consistent. Fairness, responsibility, transparency, autonomy, and trust can refer to different requirements in different studies. The keyword mapping in [15] helps to identify this variation by collecting the terms that repeatedly appear in AI-ethics research. Some conceptual studies approach AI from the question of capability. Liu et al. [89], for example, proposed intelligence quotient and intelligence-grade measures for artificial intelligence. Measuring the capability of a system is different from judging whether its behavior is ethical. The distinction is important because technical performance and ethical acceptability describe different properties of an AI system.

5.2. Robotics and autonomous systems

Ethical questions become more immediate when an AI system can act in the physical world. A wrong prediction in a software application may produce an incorrect recommendation. A wrong decision by a robot or autonomous vehicle can instead affect physical safety. Yu et al. [8] review technical approaches for introducing ethical considerations into AI systems. Etzioni and Etzioni [83] examine a harder question: whether AI-equipped machines, including driverless cars,

should be expected to make moral decisions independently. Their discussion also shows why safety, responsibility, and human control cannot be separated from the technical design of an autonomous system.

5.3. Philosophical and AI-specific ethical research

Philosophical studies are concerned with the principles used to judge right and wrong behavior. AI ethics takes the next step and asks how such principles should affect the design or use of an artificial system. Studies in this area discuss ethical principles [6], [90], the broader ethical implications of AI [77], and technical methods for including ethical requirements in AI systems [8]. Machine ethics is examined in [83], while Kim et al. [84] discuss the related problem of value alignment. There is no single moral rule that can be inserted into every AI application. A medical system, an autonomous vehicle, and an online recommendation service operate under different risks and responsibilities. The required controls therefore depend on the application, the people affected by it, and the consequences of an incorrect decision.

5.4. Law and regulation

Ethical expectations and legal obligations are related, but they are not the same. Ethics may guide how an AI system should be designed or used. Law determines duties that an organization must follow. These duties can include liability, data protection, and restrictions on particular uses of AI [9], [10]. Renda [11] discusses AI from a broader policy and governance perspective. More recently, the European Union Artificial Intelligence Act has introduced legal requirements based partly on the type and risk level of an AI system [14]. An AI application must therefore be considered within the legal environment in which it is deployed.

6. Ethical Issues of Artificial Intelligence

An AI system can work correctly from a technical point of view and still create an ethical problem. The problem may come from the data used for training, the behavior of the model, the way results are presented to a user, or the way an organization acts on those results. Fairness, privacy, transparency, accountability, safety, security, and human control are therefore discussed alongside model performance [5], [6], [79]. These issues also appear at different stages of development. A biased dataset creates a different problem from an insecure deployed model, but both can affect the people using the system. Responsible-AI and trustworthy-AI research therefore considers ethical requirements during design, testing, deployment, and later operation [7], [19].

6.1. Fairness and bias

Bias does not begin only after a model has been trained. It can enter when the problem is defined, when examples are selected for the dataset, or when labels are assigned. It can also remain hidden when only one overall performance measure is reported. Removing a sensitive field does not always solve the problem. Other variables may carry related information and act as proxies. Singh et al. [79] discuss this type of concern in multimedia applications, including face-related systems. Bechmann and Bowker [40] describe a similar issue in social-media AI, where assumptions can become embedded in data and automated analysis. A fairness assessment should therefore consider who is affected by the system and whether the errors are distributed differently across relevant groups [6], [7]. Suresh and Guttag [16] further show that harms can enter at several stages of the machine-learning lifecycle. Kaur et al. [19] similarly treat fairness as one part of a wider set of requirements for trustworthy AI. Figure 6 gives several examples in which an AI decision can lead to a different ethical concern, including fairness, privacy, accountability, transparency, and user control.

6.2. Privacy and use of data

AI systems often depend on personal, behavioral, image, location, educational, or health data. The existence of a technical way to collect these data does not by itself justify their use. Multimedia research discusses consent, privacy, licensing, and the use of images in datasets [79]. Social-media AI raises related questions about the production and interpretation of behavioral data [40]. At the governance level, data protection, access, reuse, and legal responsibility are addressed in work on AI law and data governance [9], [10]. UNESCO's recommendation includes privacy and data governance within its ethical framework [12], while NIST treats data-related risks as part of AI risk management [13]. Recent trustworthy-AI work connects privacy with security and robustness [24], and the EU AI Act adds binding requirements for specified uses and risk categories [14].

6.3. Transparency and explainability

Transparency does not mean the same thing for every person who interacts with an AI system. A user may simply need to know that AI is involved in a decision. A developer needs information about the data, model settings, and software version. An auditor may require records that show how a decision was produced. A person affected by the decision may instead need an explanation that can be understood without specialist knowledge. These needs are discussed in work on AI ethics, responsible AI, and trustworthy AI [4], [5], [7], [19]. Explainability is more specific. It concerns whether the

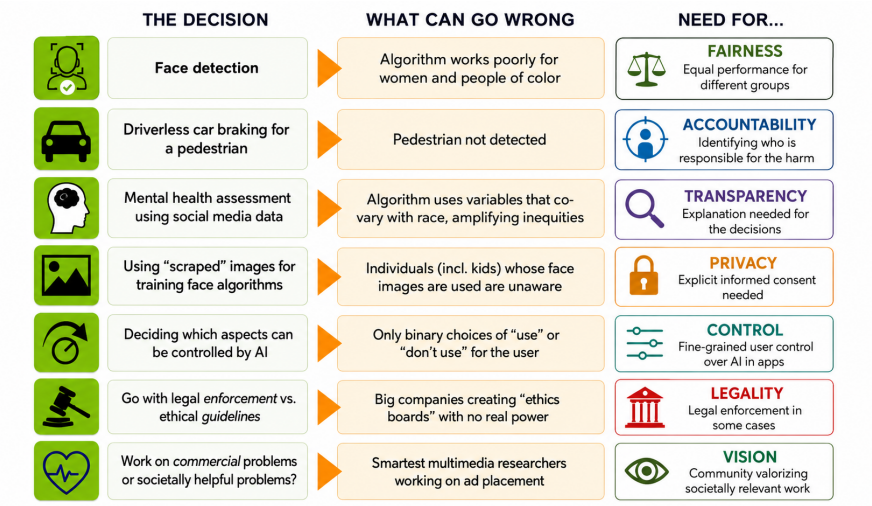

Figure 6: *Examples of AI decisions, possible failure modes, and corresponding ethical requirements.*

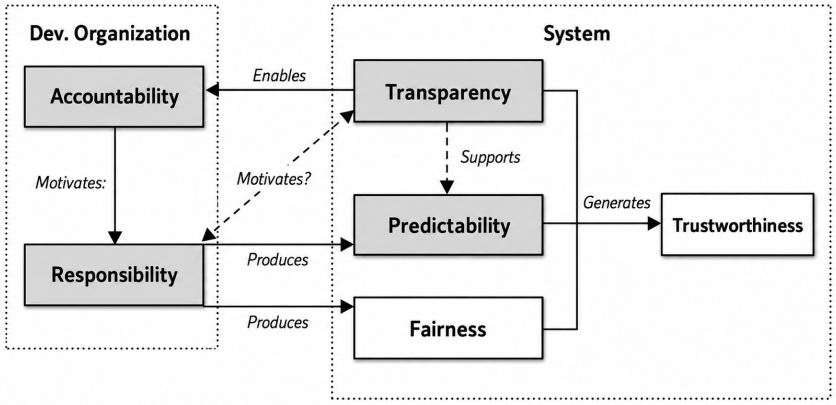

Figure 7: *Relationship among accountability, responsibility, transparency, predictability, fairness, and trustworthiness.*

behavior or output of a model can be interpreted in a useful way. The DARPA Explainable Artificial Intelligence program [91] is one example of research directed toward this technical problem.

6.4. Accountability and responsibility

Using an automated system does not transfer responsibility away from the people or organization that deploy it. Someone still has to approve its use, decide how its output will be used, review failures, and respond when a decision causes harm. Kuleshov et al. [82] discuss responsibility as part of the codification of AI ethics. According to Peters et al. [7], choices made during system design are related to responsibility. The problem is approached from an institutional perspective in legal and governance studies [9], [10]. Although keeping a record of model results is helpful, it does not provide accountability on its own. Additionally, the organization must determine who authorizes the system, who looks into a malfunction, who has the authority to halt its usage, and who is in charge of taking corrective action. International guidelines also reflect this obligation. Accountability and responsibility are two of UNESCO's AI governance tenets [12]. Instead of treating governance as a one-time review prior to deployment, the NIST AI Risk Management Framework [13] views it as a continuous process. The relationship between accountability, responsibility, transparency, predictability, fairness, and trustworthiness is shown in Figure 7.

6.5. Safety, robustness, and security

A model that performs well during testing may behave differently after it is deployed. Real operating conditions are rarely identical to the test environment. Input quality may change, new cases may appear, and an attacker may deliberately manipulate the information supplied to the system. Reliability and robustness are therefore important when the model is expected to operate outside a controlled setting [5], [19]. The importance of safety depends on what happens when the

system fails. An incorrect recommendation may have a limited effect, while an error in a medical, transport, or industrial system can create a much more serious risk. Peters et al. [7] include safety and unintended consequences in responsible-AI design. Saeed and Alsharidah [24] examine security, privacy, and robustness together because a weakness in one part of the system can affect the others. The model is also only one part of a networked AI application. Data may travel through several communication components before reaching the model or the user. Timing attacks, security problems in fog-based IoT systems, and interest-flooding attacks can interfere with this process [74]-[76]. A reliable model cannot compensate for data that have been delayed, altered, or made unavailable by a compromised communication layer.

6.6. Human autonomy and oversight

Adding a human reviewer does not automatically provide meaningful oversight. The reviewer must understand enough about the system to recognize a questionable result. The person must also have enough time and authority to stop, reject, or modify an automated action when needed. Jarrahi [47] examines how people and AI can share decision-making roles inside organizations. In health care, the use of AI does not remove the professional responsibility of clinicians [54]. Work on responsible AI and applied ethics similarly treats human involvement as part of the system design [6], [7]. UNESCO [12] includes human oversight and human determination in its recommendations for AI. Kaur et al. [19] also discuss human agency as a component of trustworthy AI. The required level of oversight will not be the same in every application. A product recommendation, a clinical decision, and an industrial control action involve different levels of risk and therefore require different forms of human involvement.

6.7. Employment and social effects

The effect of AI on work is not limited to replacing complete jobs. Automation may change only one part of an employee's task. It may also change what information is available to workers or move some decision-making authority from people to software. Jarrahi [47] examines this relationship between human and AI decision making in organizations. The economic literature looks at the issue from several other angles. Agrawal et al. [86] discuss the economic value of AI-based prediction. Hoeschl et al. [87] consider automation in relation to engineering jobs. Studies of business use examine how AI can change organizational processes and firm performance [49], [50]. Ernst et al. [88] discuss the possible implications of AI for the future of work. These studies describe different outcomes rather than one fixed effect on employment. The result depends on the type of work, the organization, the industry, and how the technology is introduced. It is therefore more useful to examine which tasks and responsibilities change than to assume that the presence of AI will automatically remove a job.

6.8. Ethical requirements can conflict

Ethical requirements are not always easy to satisfy at the same time. Transparency provides a simple example. More information about how a decision was produced may help a user understand or challenge the result. That information may also expose personal data, confidential system details, or information that could be useful to an attacker. Fairness creates a different type of tension. Evaluating whether a system performs differently across demographic groups may require information about those groups. Collecting that information can itself raise privacy and data-minimization concerns. Automation can make a process consistent while reducing the opportunity for a person to intervene. Applied ethics and responsible-design work discuss these tensions rather than assuming that one principle can always be maximized without cost [6], [7]. Lifecycle research also shows that choices made at one stage can create harm at another stage [16]. Trustworthy-AI reviews and risk-management frameworks therefore treat ethical properties as interacting requirements that need to be balanced in the intended context [13], [19]. Table 6 lists representative design questions.

Table 6: *Examples of ethical requirements that can conflict during system design.*

Requirement	Interacting requirement	Design question
Transparency	Privacy and security	How much explanation can be disclosed without exposing personal data, credentials, proprietary information, or attack surfaces?
Fairness	Data minimization	Can subgroup performance be evaluated adequately when collection of sensitive attributes is restricted?
Human autonomy	Safety	When should the system defer to human choice, and when is intervention justified to prevent a high-consequence failure?
Performance	Explainability	Is a less interpretable model acceptable when it improves task performance, and what additional evidence is required when it is used?
Personalization	Non-discrimination	Does adaptation to individual behavior improve relevance without reproducing or amplifying undesirable patterns in the training data?
Automation	Accountability	Which person or organization retains authority for a decision when the recommendation or action is generated automatically?
Data utility	Consent and purpose limitation	Is reuse of collected data consistent with the context in which the data were originally obtained?

7. Approaches to Machine Ethics

Two broad approaches are commonly used to describe how ethical guidance can enter an AI system: top-down and bottom-up approaches. Etzioni and Etzioni [83] use the same distinction in their discussion of AI-equipped machines and driverless cars. The approaches differ mainly in where the guidance originates.

7.1. Top-down approach

In a top-down approach, ethical principles or rules are specified before the machine acts. The rules may come from law, organizational policy, or a moral theory. The aim is to give the system boundaries that can be inspected in advance. Etzioni and Etzioni [83] note that proposed top-down approaches can draw on religious precepts, utilitarianism, consequentialism, or other moral theories. The difficulty is that moral theories can conflict and may produce unacceptable results in exceptional cases. A designer also cannot write a separate rule for every situation that a learning system may encounter [83]. For this reason, top-down rules are more convincing when they represent clear legal, safety, or organizational constraints than when they are expected to solve every moral question.

7.2. Bottom-up approach

A bottom-up approach begins with examples, observations, or learned behavior rather than a complete rule set. Etzioni and Etzioni [83] discuss driverless-car learning as an illustration. They note, for example, that machine-learning systems can learn driving behavior from human demonstrations. The same idea becomes difficult when the target is ethical behavior because observing what people commonly do does not show that the behavior is morally correct. The term *bottom-up* is also used in technical AI for procedures such as constituent parsing and unsupervised clustering [92], [93]. Those algorithmic uses describe how a method is constructed or learned; they should not be confused with bottom-up machine ethics. This is closely related to the value-alignment problem. Kim et al. [84] distinguish mimetic value alignment, which imitates human behavior or preference, from anchored alignment, which is tied to normative concepts of value. Their analysis also points to a limitation of learning ethical behavior from observation. Human behavior is not always consistent or desirable, so copying observed choices can reproduce the same problems in an artificial system [84]. Figure 8 illustrates the basic difference between top-down and bottom-up machine ethics. The main characteristics of the two approaches are compared in Table 7.

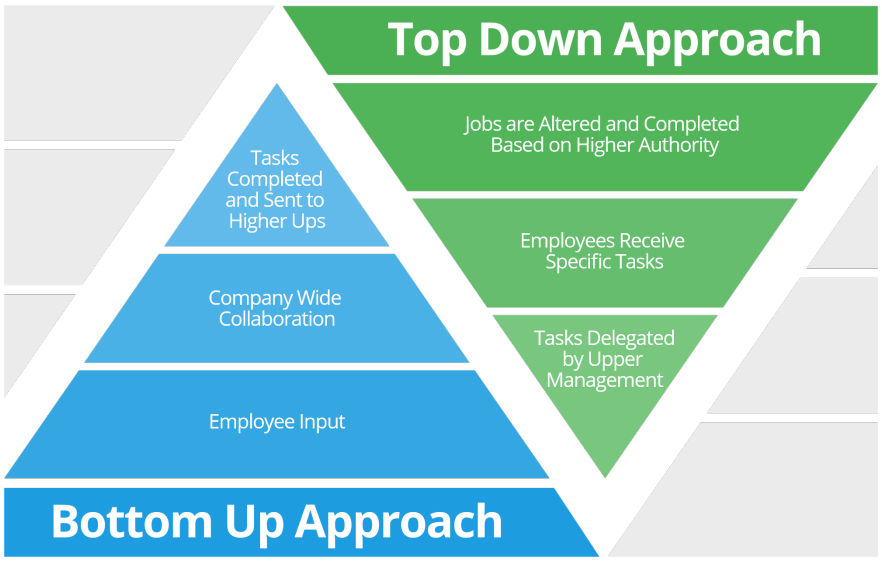

Figure 8: *Conceptual comparison of top-down and bottom-up approaches to machine ethics.*

Table 7: *Comparison of top-down and bottom-up machine-ethics approaches.*

Aspect	Top-down	Bottom-up
Starting point	Rules or principles defined before the system makes a decision.	Examples, observations, preferences, or behavior from which the system learns.
Strength	The rules can be inspected and compared with stated legal, ethical, or organizational requirements.	The system can learn patterns that may be difficult to describe with a fixed set of rules.
Risk	A fixed rule set may not cover every situation and different rules may conflict.	The system may learn bias, unsafe behavior, or undesirable human preferences from the available data.
Typical use	Legal restrictions, safety limits, organizational policies, and other requirements that should not be violated.	Preference learning, adaptation to changing situations, and behavior that is difficult to specify completely in advance.
Governance need	Validation of the rules, management of exceptions, and records showing which rule affected a decision.	Data quality checks, bias assessment, monitoring, and appropriate human oversight.

7.3. Which approach should be adopted?

Top-down and bottom-up approaches solve different parts of the problem. A fixed rule can prevent an AI system from crossing a legal or safety boundary, but it may not describe how the system should behave in every possible situation. A learned policy is more flexible, although that flexibility depends on the examples and feedback from which it learns. Etzioni and Etzioni [83] discuss this difficulty using driverless cars. Their examples also point to a basic limitation of learning from human behavior. A machine can learn what people normally do, but this does not mean that the learned behavior is ethically acceptable. The two approaches do not have to be used separately. Fixed rules can define legal, safety, or organizational limits, while learning can be used for decisions that depend on the situation. Pagallo et al. [10] describe a related middle-out approach to legal governance. In this approach, formal rules, institutions, and practical experience influence one another.

7.4. Ethics bots and preference-aware systems

Etzioni and Etzioni [83] introduce the idea of an "ethics bot." The proposed system observes information about a person's behavior and uses it to estimate that person's moral preferences. These preferences can then be used to guide the behavior of another intelligent system. There is some similarity with a recommender system. A recommender tries to learn which products, films, or other items a user prefers. An ethics bot tries to learn preferences that are related to ethical choices [83]. This does not make the ethics bot an independent moral decision maker. Its guidance is based on preferences inferred from a person, and those preferences may not always be consistent or acceptable. The system would still have to operate within legal and safety constraints. Responsibility would also remain with the people and organization that deploy and use the system.

7.5. Hard cases and trolley-style reasoning

Trolley-style problems are frequently discussed in relation to driverless cars. They describe situations in which a system is forced to choose between two harmful outcomes. Such examples are easy to understand, but they usually assume that the available choices and their consequences are already known. Etzioni and Etzioni [83] question the importance given to these extreme cases in discussions of machine ethics. Autonomous systems frequently have to make choices based on ambiguous or insufficient information. Poor sensor readings, the appearance of an unknown item, a breakdown in communication, or a scenario not included in the training data are all possible. Finding out when the system is not functioning as planned is an engineering difficulty. It might not be suitable to continue operating under such circumstances. Reduced functionality, a fallback mode, or handing over control to a human operator are some possible responses. Damage can occur at several phases of the machine-learning lifecycle, even after deployment, as demonstrated by Suresh and Guttag [16]. Similarly, the NIST framework [13] views risk management as an ongoing process rather than a one-time evaluation.

8. Digital Policy and Data Governance

8.1. Digital policy

AI services depend on more than just the model. Data must be gathered and saved, processed by software components, and transported between users, devices, and computer platforms via communication networks. The service may also include cloud or edge resources. The regulations and governance structures that control the usage of these digital resources are referred to in this assessment as "digital policy." Renda [11] looks at AI from the standpoint of governance and policy. AI is discussed with big data, the Internet of Things, and technology policy by Özdemir and Hekim [72]. Bunz and Janciute [73] concentrate more intently on the policy problems that arise when AI and IoT are used together. The development and operation of an AI service may be impacted by such regulations. They could limit certain automated choices, set security and access criteria, or impose restrictions on the gathering and use of personal data. The jurisdiction in which the system functions determines the appropriate regulations as well. Gourraud and Simon [85] talk on how American and European perspectives on AI and digital policy differ.

8.2. Role of data in digital policy

Numerous kinds of data, such as health and educational records, financial transactions, communication logs, geographical data, and internet activity, can be used by AI applications. The job and the context in which the system is utilized determine the information needed. Basic inquiries concerning origin and purpose are the first steps in data governance. An organization should be aware of who gathered the data, why, who has access to it, and how long it should be kept. A new legal or governance evaluation can be necessary if the same data is used for another purpose. Depending on the relevant legislation, cross-border transfers may potentially result in extra restrictions. Data protection is examined by Pagallo et al. [10] as part of the larger regulation of digital data and AI. In its Recommendation on the Ethics of Artificial Intelligence, UNESCO [12] addresses data governance and privacy. Data-related hazards are handled as part of the larger management of AI systems by the NIST AI Risk Management Framework [13]. Prediction, customization, and automation also require appropriate data. Therefore, governance must permit acceptable data usage while imposing suitable restrictions on data collection, access, sharing, and reuse. While too rigorous controls may prohibit legitimate uses of data, weak controls might lead to privacy or security issues. These concerns are covered in policy studies in [11], [72], and [73] in relation to the broader growth of digital and linked services. Table 8 summarizes four perspectives considered in this paper.

Table 8: *Data-governance perspectives considered in the review.*

Perspective	Primary focus	Representative concern
Security	Protection of data and control of authorized access.	Information should remain available to authorized users while being protected from unauthorized access, alteration, or disclosure.
Economy	Use of data in digital services and data-driven business models.	Data can support prediction and personalization, but its use still has to comply with privacy, access, and governance requirements.
Human rights	Privacy, surveillance, consent, jurisdiction, and civil liberties.	Large-scale collection or cross-border transfer can reduce an individual's control over personal information and may create additional legal or governance requirements.
Technology	Standards, software, infrastructure, interoperability, and controlled data exchange.	Incompatible systems, interfaces, or data formats can make secure exchange, auditing, and accountability more difficult.

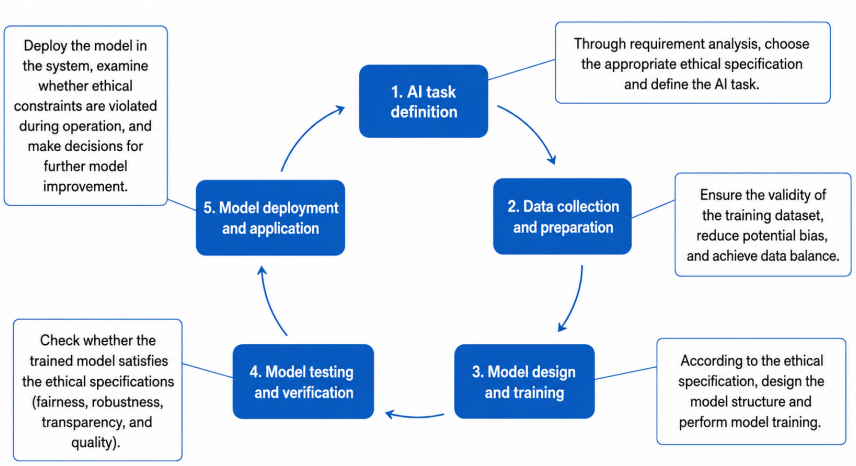

Figure 9: *Lifecycle-oriented integration of ethical principles into AI development and deployment.*

8.3. Open data, interoperability, and trust

Open data can support research and the development of new digital services, but not every dataset is suitable for unrestricted access. Personal, confidential, or otherwise sensitive information may require access controls, consent, or another lawful basis before it can be shared [10], [12]. Interoperability concerns a different problem. Two organizations may be allowed to exchange information but still be unable to do so reliably if their systems use incompatible formats, interfaces, or standards. Research on AI and IoT policy discusses the importance of technical infrastructure and interoperability in connected digital systems [72], [73]. Technical compatibility, however, does not answer the governance question. A system can be capable of exchanging information without having the legal or ethical authority to do so. Decisions about access, reuse, responsibility, and retention still have to be defined separately. Figure 9 places these concerns within an AI-development lifecycle. Ethical and governance checks are associated with task definition, data collection, model development, testing, and deployment.

8.4. Public services and institutional responsibility

The use of AI in a public service differs from many commercial applications because a citizen may have no realistic alternative to the service being provided. An automated decision may affect access to a benefit, an administrative process, or another government function. Responsibility for such a system therefore cannot be assigned only to the model or software vendor. Legal and policy studies discuss accountability, data protection, and the responsibilities of institutions using AI [9]-[11]. UNESCO [12] also includes human rights, transparency, and human oversight in its guidance. For AI systems that fall within its scope, the European Union Artificial Intelligence Act introduces additional requirements according to the system and its risk category [14]. A public organization using AI should be able to explain the purpose of the system and the information it requires. It should also identify who can review a disputed result and who is responsible when the system fails. Where an automated decision affects an individual, there should be a practical way to question or review the outcome. Accessibility is another part of the same problem. A service may work correctly from a technical perspective but still exclude people who cannot access its interface, infrastructure, or required technology. This is an operational issue as well as a governance issue. Figure 10 presents the conceptual relation among legal constraints, organizational factors, technology factors, and ethical considerations.

8.5. Citizen engagement and changing digital services

A digital policy should not be treated as a one-time document. Services, user expectations, security conditions, and legal requirements change. Renda [11] and Gourraud and Simon [85] discuss policy differences and changing governance settings. UNESCO [12] and NIST [13] also treat governance as an ongoing activity rather than a single approval step. This point is relevant to AI because a model may remain unchanged while the data, users, and operating environment change. Monitoring complaints, reviewing failures, and revisiting the purpose of a system are therefore part of responsible operation.

9. Developments from 2021 to 2024

The SMS in [15] was published in 2018, before large language models and generative AI became widely used. The studies published from 2021 to 2024 are therefore treated here as an update. They are not included in the original SMS figures. Suresh and Guttag [16] investigate the many phases of the machine-learning lifecycle when damage may occur.

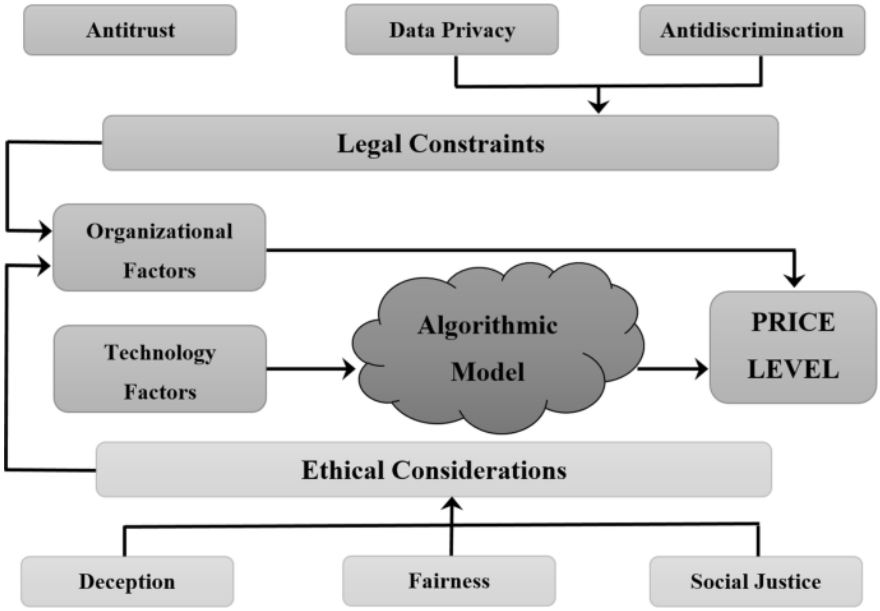

Figure 10: *Conceptual relationship between digital policy and ethically governed AI.*

Issues might arise prior to training, during model creation, or following deployment. For instance, even if the latter approach is theoretically valid, the consequent constraint may still exist if the dataset is poorly chosen. The conversation goes beyond model creation in UNESCO's Recommendation on the Ethics of Artificial Intelligence [12]. Human rights, justice, openness, human supervision, data governance, and social welfare are all included. These concerns relate to the development, introduction, and use of AI in businesses and society. Risks that were not as noticeable in the previous mapping are identified by recent research on large language models. Bender et al. [17] address issues with training data, documentation, and representation in addition to the resource cost of very large language models. A taxonomy of language-model dangers is developed by Weidinger et al. [18] and encompasses damages resulting from human contact with these systems, discrimination, disinformation, and malicious usage. The characteristics of trustworthy AI and the criteria used to assess such systems are reviewed by Kaur et al. [19]. The same broad issue is addressed from a risk-management standpoint in the NIST AI Risk Management Framework [13]. It arranges the procedure according to management, measurement, mapping, and governance. Studies that has been published since ChatGPT's launch looks into generative AI in various contexts. Its ramifications for research, practice, organizations, and policy are covered by Dwivedi et al. [20]. Large language models in education present both potential and problems, according to Kasneci et al. [21]. Feuerriegel et al. [22] outline the technological underpinnings of generative AI as well as its uses and potential hazards. The adoption of the European Union Artificial Intelligence Act in 2024 [14] gave this debate a legally mandated basis. The difference between legally permissible risk and the more general notion of trustworthy AI is examined by Laux et al. [23]. Saeed and Alsharidah [24] examine security, privacy, and robustness in AI systems to address a distinct aspect of trustworthiness. These more recent studies address systems that are different from the majority of the material covered by the older SMS in terms of scale, capabilities, and usage. However, the fundamental governance issues are still intimately tied to the prior ethical debate. Establishing the source and permissible use of data, comprehending system limitations, continuously assessing deployed behavior, and maintaining accountability for decisions based on AI output are all necessary.

10. Designing an Ethical Artificial Intelligence-Based System

The design of an AI system involves ethical as well as technical choices from the beginning of development. Problem definition affects what data are required. Data preparation influences what the model can learn, and model selection affects how the resulting system can be evaluated. Training and testing are followed by implementation, deployment, and monitoring. Problems introduced at an early stage can therefore affect the system later in its lifecycle. Suresh and Guttag [16] identify several points in this lifecycle at which harm can arise. Peters et al. [7] approach the issue from responsible design, while Kaur et al. [19] discuss requirements for trustworthy AI. Taken together, these studies support the need to consider ethical and technical questions during development rather than waiting until the system has already been deployed. Figure 11 shows the main trustworthy-AI requirements considered across the development and operation of an AI system.

Figure 11: *Trustworthy AI requirements considered throughout the AI system lifecycle.*

10.1. Identify the problem

The first step is to decide exactly what the AI system is expected to do. The task, intended users, required output, and environment in which the system will operate should be known before an algorithm is selected. Otherwise, a model can perform well according to a technical metric while still addressing the wrong problem. The same stage should define how much authority will be given to the system. Some tasks can be handled automatically. Others may require a person to review or approve the result before any action is taken. Peters et al. [7] emphasize early consideration of such issues because the objective defined at the beginning affects later choices concerning data, model design, evaluation, and deployment.

10.2. Prepare the data

An AI model can only learn from the information provided to it. The data should therefore be relevant to the task and organized in a form that the selected method can process. Structured data normally follow a defined schema, whereas text, images, audio, and video are common examples of unstructured data. General machine-learning sources discuss data preparation as part of the model-development process [3], [28]. The origin of the data is equally important. A development team should know where the information came from, whether it can legally and appropriately be used, and whether the people represented in the dataset match the population for which the system is intended. Missing or poorly represented groups can affect the behavior of the resulting model. Suresh and Guttag [16] identify data collection and preparation as stages at which harm can already be introduced. Privacy and data-governance requirements also apply before training begins [12], [13]. These checks are easier to perform before the data have been built into a deployed model. Dataset size alone does not determine data quality. Unrelated information can be removed, and the remaining records should be organized in a form that the model can use consistently. If the data are biased, poorly documented, or collected for a different purpose, a more complex algorithm does not automatically correct the problem.

10.3. Choose the learning method

The learning method follows from the task and the available data. Supervised learning uses labeled examples to learn a mapping from input to output. Unsupervised learning looks for structure in data without labeled targets. Reinforcement learning learns actions through interaction and reward. These are different learning paradigms and should not be treated as

the same process [2], [3], [28]. Unsupervised learning and reinforcement learning are distinct paradigms and are therefore treated separately in Table 9. The table also lists ethical questions that commonly arise with each approach.

Table 9: *Learning paradigms and corresponding ethical considerations.*

Paradigm	Technical objective	Ethical considerations
Supervised learning	Learn a mapping from labeled examples to predictions.	Label quality, representativeness, subgroup error, provenance, and the consequences of false positives/negatives.
Unsupervised learning	Discover patterns, clusters, or structure without labeled targets.	Interpretation of discovered groups, proxy discrimination, and misuse of inferred categories.
Reinforcement learning	Learn actions that maximize a reward through interaction.	Reward specification, unsafe exploration, unintended strategies, oversight, and constraints on autonomous behavior.

10.4. Train and test the model

Training begins after the data and learning method have been selected. A good training result is only the first check. The model should also be evaluated on data and conditions that represent its intended use. Where the application affects people directly, evaluation may need to consider relevant groups, error consequences, robustness, security, and human interaction [7], [19]. A failure found during testing may require a change in the data, the model, or the operating rule around it. The data source, model settings, and evaluation procedure should also be recorded. The reproducibility issue in AI research is explained by Hutson [94]. When crucial details regarding the initial setup are absent, it becomes challenging to duplicate an experiment. Therefore, the dataset used, the model configuration, the experimental conditions, and the evaluation process should all be included in the development record.

10.5. Programming and implementation

To implement the model, it must be integrated with the remainder of the application. AI libraries, data interfaces, storage, authentication, and software components that supply inputs to the model or utilize its output are typically involved in this. This phase also establishes whether certain governance standards are effective in real-world situations. If the software doesn't use access control to enforce a restriction limiting access to personal data, it won't have much of an impact. Because they enable later verification of the system's configuration and usage, logs, version records, retention restrictions, and test records are also crucial. Design practice and responsible-AI ideas are discussed by Peters et al. [7]. Requirements decided upon during design must be implemented as working controls in the software, not just in design or policy documents.

10.6. Deployment and monitoring

Typically, specific data and predetermined evaluation criteria are used for training and testing. After the system is placed into use, it interacts with software services, actual users, and organizational procedures while also receiving fresh inputs. As a result, the conditions used during development may not be the same as those encountered during use. They might also evolve over time. Software components can be upgraded, and user behavior can change. An organization may also alter how it applies the results of the model. Deployment is one phase of the machine-learning lifecycle when harm might occur, according to Suresh and Guttag [16]. After release, monitoring must continue. It is necessary to document and look into reported occurrences and changes in performance. Issues that were not discovered during testing may also be revealed by complaints. The operating process should specify how a disputed result can be evaluated and when the system can be suspended if it directly affects an individual. Measurement and management are considered as ongoing risk-management tasks in the NIST AI Risk Management Framework [13]. Therefore, rather than being a task saved for a significant failure, post-deployment monitoring is a regular element of operations. A summary of representative technical activities and governance controls across the AI lifecycle can be found in Table 10.

Table 10: *Technical activities and ethical controls across the AI lifecycle.*

Stage	Technical activity	Ethical and governance control
Task definition	Define the objective, intended users, expected output, and operating conditions.	Identify who may be affected, consider possible harms and applicable legal constraints, and determine which decisions require human review or authority.
Data preparation	Collect, clean, organize, label, and transform the required data.	Record the origin of the data and permission for its use. Check privacy, representation of relevant groups, possible bias, access rights, and retention conditions.
Model design	Choose the learning method, architecture, and optimization objective.	Check whether the design can meet the fairness, transparency, robustness, and safety requirements that are relevant to the intended application.
Training and testing	Train the model and evaluate its behavior.	Compare performance across relevant groups where applicable, and examine known failure cases, security, robustness, and model limitations.
Deployment	Integrate the model with users, software, and organizational processes.	Assign responsibility, specify when human intervention is required, keep traceable records, and define how incidents will be handled.
Monitoring	Observe the system during continued use.	Review changes in performance, complaints, incidents, and model drift. Specify when further review, suspension, or withdrawal is required.

10.7. Practical characteristics of an AI service

Even if a model does well in testing, it may still be challenging to deploy as a service. After implementation, access, affordability, accountability, interoperability with the surrounding system, and user-accessible data are all important. Availability, affordability, accountability, compatibility, and comprehensibility are the five service-level attributes into which this assessment divides these pragmatic concerns. The grouping is not meant to serve as a general framework for ethical AI, rather, it is used here to address deployed AI services. In practical terms, availability refers to access. The service must be accessible to intended users as needed. Even when the basic model functions properly, access may be restricted by frequent outages, inadequate infrastructure, or technological constraints that certain users are unable to fulfill. When the cost of accessing the service puts a barrier in the way of its intended consumers, affordability becomes important. The application and the availability of workable alternatives determine how important it is. Identifiable responsibility is necessary for accountability. The organization must be aware of who authorized deployment, who assesses issues, who has the authority to modify or halt the system, and who is still in charge of making decisions based on its results. The primary problem with compatibility is operational. The software, hardware, data formats, communication interfaces, and organizational processes that the application requires must all be compatible with the AI component. The function of the user of the system determines its comprehensibility. Developers could require in-depth technical data. Operating restrictions and intervention protocols must be understood by operators. Regular customers require sufficient information to comprehend the service's goal and the significance of its results. The five service-level characteristics utilized in this review are summarized in Table 11.

Table 11: *Service-level characteristics considered in this review.*

Characteristic	Interpretation for AI services
Availability	The intended users can reach and use the service with the infrastructure available to them.
Affordability	The cost of access is considered in relation to the intended users and the practical alternatives available to them.
Accountability	Responsibility for approval, deployment, review, intervention, and decisions based on system output is assigned to identifiable people or roles.
Compatibility	The system can operate with the required infrastructure, software components, data formats, organizational procedures, and human workflow.
Comprehensibility	Users and operators receive enough information about the system's purpose, limitations, and outputs for their respective roles.

11. Discussion

Practical decisions made during development or operation are the starting point for many of the ethical issues discussed in this work. What a model can learn is altered by inadequate data. Weak testing could miss significant failure scenarios, and

an inappropriate goal could steer the system in the wrong direction. Even when the model functions as intended, insecure communication might cause issues in a networked application. When an organization does not specify how an automated outcome should be perceived or contested, a different issue emerges. Recommendation systems, autonomous systems, and medical applications can all have these flaws, but the outcomes are vastly different. One application's mistake could be inconvenient. In another, physical safety or health may be impacted by the same kind of failure. AI ethics cannot be reduced down to a single category of issues, as the SMS in [15] explains. Conceptual, philosophical, legal, human-centered, and technical activities are all included in its classification. In the broader development of intelligent systems, Dignum [77] takes ethical issues into account. Etzioni and Etzioni [83] concentrate more specifically on how AI-powered computers behave morally and make decisions. Pagallo et al. [10] and Robles Carrillo [9] discuss governance, data security, and legal accountability. These viewpoints address several aspects of the same implemented system. The components that can be directly developed, tested, or observed are the focus of engineering study. Responsible AI design is examined by Peters et al. [7]. While the DARPA XAI program [91] focuses on explainability, Saeed and Alsharidah [24] concentrate on security, privacy, and resilience. Developers can investigate some of the more general ethical and governance needs covered in the literature with the help of such work. Organizational accountability is not eliminated by technical controls. The organization determines the purpose of the system's deployment as well as the appropriate level of authority for its output. It must also determine who evaluates a contested outcome, when human involvement is necessary, and who takes corrective action following a failure. It is not possible to assign these duties to the model. With each application, the balance of issues shifts. Smart-home systems frequently depend on linked gadgets and ongoing sensing. Privacy, security, and user control are therefore directly relevant in this setting [33]-[35]. In education, automated assessment or prediction may affect how a student is evaluated or supported [41], [42], [44], [45]. Fairness, transparency, protection of student data, and human review therefore become important concerns [41], [42], [44], [45]. Health-care systems have a different risk profile. Clinical validation, patient privacy, safety, explanation, and professional responsibility are particularly important because an incorrect result may affect medical care [51]-[54]. Network and IoT applications depend on secure and reliable communication [67], [68], [75]. Organizational systems raise questions about decision authority, work practices, and employment [47], [50], [86], [88]. Table 12 compares these domains and the ethical requirements that are most relevant to each one.

Table 12: *AI application domains and associated ethical requirements.*

Domain	Primary AI role	Priority ethical requirements
Consumer/smart home	Personalization, assistance, sensing, and recommendation.	Privacy, consent, user control, data retention, security, and transparency.
Education	Assessment, prediction, and adaptive learning.	Fairness, review of disputed results, transparency, human involvement, and protection of student information.
Health care	Diagnostic support, image analysis, and patient monitoring.	Safety, clinical validation, privacy, explanation, accountability, and human oversight.
Transport	Prediction, navigation, and autonomous control.	Safety, uncertainty handling, human intervention, responsibility, and legal compliance.
Industry/energy	Optimization, prediction, and process control.	Robustness, traceability, uncertainty assessment, safety margins, and secure operation.
Networks/IoT	Analytics, communication, and control.	Cybersecurity, privacy, resilience, data governance, and interoperability.
Public sector	Service delivery and administrative decision support.	Rights, review procedures, accessibility, transparency, accountability, and non-discrimination.

11.1. Comparison across application domains

An ethical requirement cannot always be evaluated in the same way across different applications. Privacy is handled differently across application domains. In a smart home, the concern is often continuous sensing inside a private environment and control over the information collected by connected devices [33], [34]. Privacy has different implications in health care. Patient records, medical images, and other clinical information require appropriate protection and control over access [54]. Automated decisions in education create a separate concern. A prediction may affect how a student is assessed or what support is offered. Performance across relevant student groups and the possibility of teacher review should therefore be considered when AI contributes to an academic decision [42], [45]. The effect of an error depends strongly on the application. A poor product recommendation may inconvenience a user. An incorrect result in a clinical or autonomous system can have a much more serious effect because health or physical safety may be involved. The amount of testing, human review, and operational control should therefore match the risks associated with the intended use. Networked AI systems depend on the communication path as well as the model. Even when the model behaves correctly, the service may fail if data are delayed, lost, or altered during transmission. This reliance on communication infrastructure is demonstrated by wireless and smart-city applications [70], [71]. Similar issues can arise in IoT and

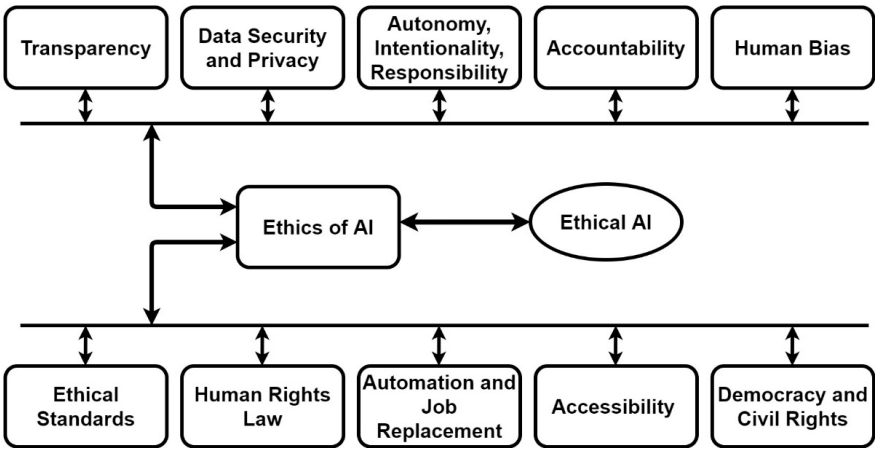

Figure 12: *Relationships among key ethical requirements in AI systems.*

automotive systems, where flooding, communication interruption, and timing assaults can impact the data accessible to the service [74]–[76]. While a checklist can assist in identifying problems that need to be addressed, it is unable to determine which control is appropriate for each application. The individuals impacted, how the system is used, and the repercussions of failure all influence that choice.

11.2. From principles to implementation

Only when ethical ideas can be translated into requirements that can be verified in real-world situations can they become valuable in system development. Comparing model performance across pertinent groups and looking for significant discrepancies is a helpful measure for fairness. Establishing guidelines for data collection, access, retention, and reuse can help protect privacy. The audience also affects transparency. Records outlining the dataset, model settings, and system behavior may be required by developers. In general, users want an explanation of the system's function, the significance of its output, and any restrictions pertaining to their use of that information. Accountability requires named responsibilities for approval, review, and corrective action. Safety also has to be expressed through controls that can be tested and applied in practice. These may include testing under abnormal conditions, defined operating limits, failure detection, and a procedure for suspending or reviewing the system when necessary. Responsible-AI and trustworthy-AI research links these general ethical principles with technical and organizational controls [6], [7], [19]. Figure 12 brings together several ethical requirements discussed in this section.

Table 13 links the nine taxonomy categories with stages of AI development. The check marks represent a conceptual synthesis proposed in this review; they are not frequency counts from the 83 papers reported in the SMS.

Table 13: *Conceptual relationship between the nine-category taxonomy and stages of the AI lifecycle.*

Taxonomy category	Task definition	Data preparation	Model design	Validation	Deployment	Monitoring
Conceptual	✓		✓		✓	
Robotics	✓		✓	✓	✓	✓
General philosophical and ethical	✓		✓	✓	✓	✓
AI-specific philosophical and ethical	✓	✓	✓	✓	✓	✓
Law and regulation	✓	✓		✓	✓	✓
Autonomous vehicles	✓		✓	✓	✓	✓
AGI and AI risk	✓		✓	✓	✓	✓
Human cognition	✓	✓	✓	✓	✓	✓
Technology-oriented		✓	✓	✓	✓	✓

11.3. Practical control of an AI-based service

A practical question remains after the principles are listed: what happens when the system behaves outside its intended conditions? A physical autonomous system may need an immediate safe-stop function. A software decision-support system may need manual review, suspension, or rollback. The people operating the system should know when such controls are to be used. Control also depends on understanding the limits of the model. An operator who cannot interpret the purpose and limitations of a system may trust it in the wrong situation. A user who cannot identify the organization responsible for an automated decision has little practical recourse. The characteristics discussed in Section 10.7 matter

when the system is placed in actual use. Conditions around the service, not only the characteristics of the underlying model, are described by availability, cost, accountability, compatibility, and comprehensibility. If the service cannot be accessed, comprehended, integrated, or held responsible, a model may function well technically but still be inappropriate.

12. Limitations

The SMS used in this paper cannot be reproduced completely from the information that is available. Vakkuri and Abrahamsson [15] report the databases used in the search, an initial set of 1,062 papers, 83 retained academic papers, and 37 recurring keywords. Complete database-specific search strings and exact search dates are not available. The available information also does not provide a duplicate-removal record or inter-reviewer agreement. The numerical results reported in [15] are therefore retained as published, but the full search and screening process cannot be repeated independently from the information provided. The use of keywords introduces another limitation. Different authors may use different words for the same or closely related concepts. An abbreviation and its complete form may also be recorded separately. Domain-specific terminology can create the same problem. The nine categories are therefore more useful for organizing the topics found in the literature than for comparing their importance by frequency.

13. Future Research Directions

The mapping study should be repeated using a procedure that allows the search and screening process to be reproduced. Reproducing the mapping requires a more complete record of the search and screening process. The full search strings should be reported together with the date of each database search, the screening criteria, the duplicate-removal procedure, and agreement between reviewers. A new mapping could then examine whether the 37 recurring keywords and nine categories reported in [15] remain evident in AI-ethics research after the growth of large language models and generative AI. Fairness, transparency, privacy, safety, and human oversight also need application-specific measures. These ideas are too general to assess without clarifying their implications for the system under study. The initial step should be to identify the impacted users or groups. The research should specify precisely what is being assessed as well as the evidence that will be utilized to determine the outcome. Although frameworks for trustworthy and responsible AI can direct this effort [7], [19], they do not eliminate the necessity to provide criteria for the specific application. It takes more than just assigning a reviewer for human monitoring to be successful. The individual evaluating an AI output must be aware of the system's limitations and have adequate time to challenge the outcome. Additionally, the reviewer must be able to halt or modify an automatic judgment as needed. Future research should investigate if these circumstances are indeed present when using the technology. In fields like health care, education, public services, and autonomous systems, where an automated outcome might directly impact individuals, this topic is crucial. Data gathered from several sources can be used by contemporary AI services. The same data may travel across several software components and communication networks before arriving at the model. This necessitates tracking the data's origin, processing method, and system components. Development records should therefore identify where the data originated, whether their use was permitted, how they were transformed, which components processed them, and under what conditions they were shared or reused. Security testing should also consider attacks against the data, the model, and the communication infrastructure [16], [24], [75], [76]. Evaluation should remain specific to the application domain. The evidence required for a smart-home service is not the same as the evidence required for a hospital, a public-sector system, or an autonomous vehicle. Future benchmarks should therefore report technical performance together with the privacy, fairness, safety, human-control, and accountability requirements that are relevant to the intended deployment setting.

14. Conclusion

AI is now used in systems that support everyday services as well as decisions in health care, education, industry, communication, and public administration. The technology can reduce manual work and process information at a scale that is difficult for people, but model performance answers only part of the question. The way data are collected, the way a result is used, and the responsibility attached to that result remain important. The SMS reported in [15] retrieved 1,062 papers, retained 83 academic papers, and identified 37 recurring keywords organized into nine categories. The nine categories cover conceptual, philosophical, legal, human-centered, application-specific, and technical aspects of AI ethics. They offer a standard method for classifying issues that arise in various AI applications. Throughout the system lifespan, ethical concerns must also be taken into consideration. Even if a more powerful method is subsequently used, an issue that was established during the task definition or dataset preparation may still exist. Monitoring, deployment, testing, and implementation can all cause new issues. When setting organizational, safety, or legal boundaries, top-down regulations are helpful. When the system must adjust to circumstances that are not fully predetermined, bottom-up learning can be applied. The individuals and institutions who create, authorize, and utilize the deployed system are still in charge of it [7], [83]. The prior subject is expanded upon by the material released between 2021 and 2024. Large language models, trustworthy-AI frameworks, generative AI, lifecycle hazards, risk management, and current regulations are all included [12]–[14], [16]–[24]. The review is updated using these subsequent research. Neither the initial SMS numbers nor the

counts stated in [15] are altered by them. Reproducible reviews, quantifiable ethical standards, data governance, human supervision, security, and domain-specific evaluation are all strengthened by them.

Author Declarations

Author Contributions (CRediT): Soban Ahmad: Conceptualization, Methodology, Investigation, Literature Review, Data Curation, Formal Analysis, Visualization, Writing - Original Draft. Zupash Awais: Conceptualization, Methodology, Validation, Supervision, Writing - Review & Editing. All authors reviewed and approved the final manuscript.

Funding: No external funding was received for this research.

Data Availability Statement: No new datasets were generated or analyzed in this review.

Acknowledgments: The authors have no acknowledgments to report.

Conflicts of Interest: The authors declare no conflicts of interest.

Ethics Statement: Not applicable.

Declaration of Generative AI: Quillbot tool was used during manuscript preparation for language editing and formatting support. The authors reviewed and verified the revised content and remain fully responsible for the accuracy, integrity, interpretation, citations, and final content of the manuscript.

References

[1] J. Shabbir and T. Anwer, "Artificial intelligence and its role in near future," *arXiv preprint arXiv:1804.01396*, 2018.

[2] K. Kersting, "Machine learning and artificial intelligence: Two fellow travelers on the quest for intelligent behavior in machines," *Frontiers in Big Data*, vol. 1, Art. no. 6, 2018. doi: 10.3389/fdata.2018.00006.

[3] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 2nd ed. Upper Saddle River, NJ, USA: Prentice Hall, 2003.

[4] A. T. Johnson, "Ethics in the era of artificial intelligence," *IEEE Pulse*, vol. 11, no. 3, pp. 44-47, 2020. doi: 10.1109/MPULS.2020.2993667.

[5] S. Jain, M. Luthra, S. Sharma, and M. Fatima, "Trustworthiness of artificial intelligence," in *Proc. 6th Int. Conf. Advanced Computing and Communication Systems (ICACCS)*, pp. 907-912, 2020. doi: 10.1109/ICACCS48705.2020.9074237.

[6] S. Lo Piano, "Ethical principles in machine learning and artificial intelligence: Cases from the field and possible ways forward," *Humanities and Social Sciences Communications*, vol. 7, Art. no. 9, 2020. doi: 10.1057/s41599-020-0501-9.

[7] D. Peters, K. Vold, D. Robinson, and R. A. Calvo, "Responsible AI-Two frameworks for ethical design practice," *IEEE Transactions on Technology and Society*, vol. 1, no. 1, pp. 34-47, 2020. doi: 10.1109/TTS.2020.2974991.

[8] H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser, and Q. Yang, "Building ethics into artificial intelligence," in *Proc. 27th Int. Joint Conf. Artificial Intelligence (IJCAI)*, pp. 5527-5533, 2018. doi: 10.24963/ijcai.2018/779.

[9] M. Robles Carrillo, "Artificial intelligence: From ethics to law," *Telecommunications Policy*, vol. 44, no. 6, Art. no. 101937, 2020. doi: 10.1016/j.telpol.2020.101937.

[10] U. Pagallo, P. Casanovas, and R. Madelin, "The middle-out approach: Assessing models of legal governance in data protection, artificial intelligence, and the Web of Data," *The Theory and Practice of Legislation*, vol. 7, no. 1, pp. 1-25, 2019. doi: 10.1080/20508840.2019.1664543.

[11] A. Renda, *Artificial Intelligence: Ethics, Governance and Policy Challenges-Report of a CEPS Task Force*. Brussels, Belgium: Centre for European Policy Studies, 2019.

[12] UNESCO, *Recommendation on the Ethics of Artificial Intelligence*. Paris, France: United Nations Educational, Scientific and Cultural Organization, adopted Nov. 23, 2021.

[13] E. Tabassi, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1. Gaithersburg, MD, USA: National Institute of Standards and Technology, 2023. doi: 10.6028/NIST.AI.100-1.

[14] European Parliament and Council of the European Union, "Regulation (EU) 2024/1689 of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," *Official Journal of the European Union*, L 2024/1689, Jul. 12, 2024.

[15] V. Vakkuri and P. Abrahamsson, "The key concepts of ethics of artificial intelligence: A keyword based systematic mapping study," in *Proc. IEEE Int. Conf. Engineering, Technology and Innovation (ICE/ITMC)*, pp. 1-6, 2018. doi: 10.1109/ICE.2018.8436265.

[16] H. Suresh and J. Guttag, "A framework for understanding sources of harm throughout the machine learning life cycle," in *Proc. Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO)*, pp. 1-9, 2021. doi: 10.1145/3465416.3483305.

[17] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proc. 2021 ACM Conf. Fairness, Accountability, and Transparency (FAccT)*, pp. 610-623, 2021. doi: 10.1145/3442188.3445922.

[18] L. Weidinger, J. Uesato, M. Rauh, C. Griffin, P.-S. Huang, J. Mellor, et al., "Taxonomy of risks posed by language models," in *Proc. 2022 ACM Conf. Fairness, Accountability, and Transparency (FAccT)*, pp. 214-229, 2022. doi: 10.1145/3531146.3533088.

[19] D. Kaur, S. Uslu, K. J. Rittichier, and A. Durresi, "Trustworthy artificial intelligence: A review," *ACM Computing Surveys*, vol. 55, no. 2, Art. no. 39, pp. 1-38, 2022. doi: 10.1145/3491209.

[20] Y. K. Dwivedi, N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar, et al., "So what if ChatGPT wrote it? Multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy," *International Journal of Information Management*, vol. 71, Art. no. 102642, 2023. doi: 10.1016/j.ijinfomgt.2023.102642.

[21] E. Kasneci, K. Sessler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, et al., "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, Art. no. 102274, 2023. doi: 10.1016/j.lindif.2023.102274.

[22] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech, "Generative AI," *Business & Information Systems Engineering*, vol. 66, pp. 111-126, 2024. doi: 10.1007/s12599-023-00834-7.

[23] J. Laux, S. Wachter, and B. Mittelstadt, "Trustworthy artificial intelligence and the European Union AI Act: On the conflation of trustworthiness and acceptability of risk," *Regulation & Governance*, vol. 18, no. 1, pp. 3-32, 2024. doi: 10.1111/rego.12512.

[24] M. M. Saeed and M. Alsharidah, "Security, privacy, and robustness for trustworthy AI systems: A review," *Computers & Electrical Engineering*, vol. 119, Part B, Art. no. 109643, 2024. doi: 10.1016/j.compeleceng.2024.109643.

[25] E. Blasch, I. Kadar, L. L. Grewe, G. Stevenson, U. K. Majumder, and C.-Y. Chong, "Deep learning in AI and information fusion panel discussion," in *Proc. SPIE 11018, Signal Processing, Sensor/Information Fusion, and Target Recognition XXVIII*, Art. no. 110180Q, 2019. doi: 10.1117/12.2519230.

[26] S. P. Yadav, D. P. Mahato, and N. T. D. Linh, Eds., *Distributed Artificial Intelligence: A Modern Approach*, 1st ed. Boca Raton, FL, USA: CRC Press, 2021.

[27] C. S. Krishnamoorthy and S. Rajeev, *Artificial Intelligence and Expert Systems for Engineers*. Boca Raton, FL, USA: CRC Press, 1996.

[28] D. Cook, *Practical Machine Learning with H2O: Powerful, Scalable Techniques for Deep Learning and AI*. Sebastopol, CA, USA: O'Reilly Media, 2016.

[29] Y. Bengio, "Deep learning for AI," invited talk, *AAAI Conference on Artificial Intelligence (AAAI-20)*, New York, NY, USA, 2020.

[30] L. Deng, "Artificial intelligence in the rising wave of deep learning: The historical path and future outlook [Perspectives]," *IEEE Signal Processing Magazine*, vol. 35, no. 1, pp. 177-180, 2018. doi: 10.1109/MSP.2017.2762725.

[31] M. A. Laughton, "Artificial intelligence techniques in power systems," in *IEE Colloquium on Artificial Intelligence Techniques in Power Systems*, Digest no. 1997/354, pp. 1/1-1/18, 1997. doi: 10.1049/ic:19971179.

[32] K. Vassiljeva, E. Petlenkov, V. Vansovits, and A. Tepljakov, "Artificial intelligence methods for data based modeling and analysis of complex processes: Real life examples," in *Proc. IEEE First Int. Conf. Data Stream Mining & Processing (DSMP)*, pp. 363-368, 2016. doi: 10.1109/DSMP.2016.7583579.

[33] X. Guo, Z. Shen, Y. Zhang, and T. Wu, "Review on the application of artificial intelligence in smart homes," *Smart Cities*, vol. 2, no. 3, pp. 402-420, 2019. doi: 10.3390/smartcities2030025.

[34] S. M. E. Sepasgozar, R. Karimi, L. Farahzadi, F. Moezzi, S. Shirowzhan, S. M. Ebrahimzadeh, F. Hui, and L. Aye, "A systematic content review of artificial intelligence and the Internet of Things applications in smart home," *Applied Sciences*, vol. 10, no. 9, Art. no. 3074, 2020. doi: 10.3390/app10093074.

[35] V. Kopytko, L. Shevchuk, L. Yankovska, Z. Semchuk, and R. Strilchuk, "Smart home and artificial intelligence as environment for the implementation of new technologies," *Path of Science*, vol. 4, no. 9, pp. 2007-2012, 2018. doi: 10.22178/pos.38-2.

[36] F. Kane, *Building Recommender Systems with Machine Learning and AI: Help People Discover New Products and Content with Deep Learning, Neural Networks, and Machine Learning Recommendations*. Birmingham, U.K.: Packt Publishing, 2018.

[37] Z. Yan, N. Duan, P. Chen, M. Zhou, J. Zhou, and Z. Li, "Building task-oriented dialogue systems for online shopping," in *Proc. 31st AAAI Conf. Artificial Intelligence*, pp. 4618-4625, 2017.

[38] C. Li, R. Pan, H. Xin, and Z. Deng, "Research on artificial intelligence customer service on consumer attitude and its impact during online shopping," *Journal of Physics: Conference Series*, vol. 1575, Art. no. 012192, 2020. doi: 10.1088/1742-6596/1575/1/012192.

[39] D. Francescato, "Globalization, artificial intelligence, social networks and political polarization: New challenges for community psychologists," *Community Psychology in Global Perspective*, vol. 4, no. 1, pp. 20-41, 2018. doi: 10.1285/i24212113v4i1p20.

[40] A. Bechmann and G. C. Bowker, "Unsupervised by any other name: Hidden layers of knowledge production in artificial intelligence on social media," *Big Data & Society*, vol. 6, no. 1, Art. no. 2053951718819569, 2019. doi: 10.1177/2053951718819569.

[41] P. S. Almeida, P. Novais, E. Costa, M. Rodrigues, and J. Neves, "Artificial intelligence tools for student learning assessment in professional schools," in *Proc. 7th European Conf. on e-Learning*, Cyprus, pp. 17-28, 2008.

[42] F. Cruz-Jesus, M. Castelli, T. Oliveira, R. Mendes, C. Nunes, M. Sá-Velho, et al., "Using artificial intelligence methods to assess academic achievement in public high schools of a European Union country," *Heliyon*, vol. 6, no. 6, Art. no. e04081, 2020. doi: 10.1016/j.heliyon.2020.e04081.

[43] D. Crowe, M. LaPierre, and M. Kebritchi, "Knowledge based artificial augmentation intelligence technology: Next step in academic instructional tools for distance learning," *TechTrends*, vol. 61, no. 5, pp. 494-506, 2017. doi: 10.1007/s11528-017-0210-4.

[44] R. Bajaj and V. Sharma, "Smart education with artificial intelligence based determination of learning styles," *Procedia Computer Science*, vol. 132, pp. 834-842, 2018. doi: 10.1016/j.procs.2018.05.095.

[45] S. Chattopadhyay, S. Shankar, R. B. Gangadhar, and K. Kasinathan, "Applications of artificial intelligence in assessment for learning in schools," in *Handbook of Research on Digital Content, Mobile Learning, and Technology Integration Models in Teacher Education*, IGI Global, pp. 185-206, 2018. doi: 10.4018/978-1-5225-2953-8.ch010.

[46] A. Alghoul, S. Al Ajrami, G. Al Jarousha, G. Harb, and S. S. Abu-Naser, "Email classification using artificial neural network," *International Journal of Academic Engineering Research*, vol. 2, no. 11, pp. 8-14, 2018.

[47] M. H. Jarrahi, "Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making," *Business Horizons*, vol. 61, no. 4, pp. 577-586, 2018. doi: 10.1016/j.bushor.2018.03.007.

[48] A. M. Abubakar, E. Behravesh, H. Rezapouraghdam, and S. B. Yildiz, "Applying artificial intelligence technique to predict knowledge hiding behavior," *International Journal of Information Management*, vol. 49, pp. 45-57, 2019. doi: 10.1016/j.ijinfomgt.2019.02.006.

[49] E. Brynjolfsson and A. McAfee, "The business of artificial intelligence," *Harvard Business Review*, Jul. 18, 2017.

[50] S.-L. Wamba-Taguimdje, S. F. Wamba, J. R. K. Kamdjoug, and C. E. T. Wanko, "Influence of artificial intelligence (AI) on firm performance: The business value of AI-based transformation projects," *Business Process Management Journal*, vol. 26, no. 7, pp. 1893-1924, 2020. doi: 10.1108/BPMJ-10-2019-0411.

[51] P. Ström, K. Kartasalo, H. Olsson, L. Solorzano, B. Delahunt, D. M. Berney, et al., "Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: A population-based, diagnostic study," *The Lancet Oncology*, vol. 21, no. 2, pp. 222-232, 2020. doi: 10.1016/S1470-2045(19)30738-7.

[52] D. Tsay and C. Patterson, "From machine learning to artificial intelligence applications in cardiac care: Real-world examples in improving imaging and patient access," *Circulation*, vol. 138, no. 23, pp. 2569-2575, 2018.

[53] T. Hirasawa, K. Aoyama, T. Tanimoto, S. Ishihara, S. Shichijo, T. Ozawa, et al., "Application of artificial intelligence using a convolutional neural network for detecting gastric cancer in endoscopic images," *Gastric Cancer*, vol. 21, no. 4, pp. 653-660, 2018. doi: 10.1007/s10120-018-0793-2.

[54] K. H. Keskinbora, "Medical ethics considerations on artificial intelligence," *Journal of Clinical Neuroscience*, vol. 64, pp. 277-282, 2019. doi: 10.1016/j.jocn.2019.03.001.

[55] K. J. Assi, M. Shafiullah, K. M. Nahiduzzaman, and U. Mansoor, "Travel-to-school mode choice modelling employing artificial intelligence techniques: A comparative study," *Sustainability*, vol. 11, no. 16, Art. no. 4484, 2019. doi: 10.3390/su11164484.

[56] M. Hofmann, F. Neukart, and T. Bäck, "Artificial intelligence and data science in the automotive industry," *arXiv preprint arXiv:1709.01989*, 2017.

[57] G. Papa and B. Koroušić-Seljak, "An artificial intelligence approach to the efficiency improvement of a universal motor," *Engineering Applications of Artificial Intelligence*, vol. 18, no. 1, pp. 47-55, 2005. doi: 10.1016/j.engappai.2004.09.003.

[58] A. Mellit and S. A. Kalogirou, "Artificial intelligence techniques for photovoltaic applications: A review," *Progress in Energy and Combustion Science*, vol. 34, no. 5, pp. 574-632, 2008. doi: 10.1016/j.pecs.2008.01.001.

[59] C. Qi, A. Fourie, G. Ma, X. Tang, and X. Du, "Comparative study of hybrid artificial intelligence approaches for predicting hangingwall stability," *Journal of Computing in Civil Engineering*, vol. 32, no. 2, Art. no. 04017086, 2018. doi: 10.1061/(ASCE)CP.1943-5487.0000737.

[60] R. Kretschmer, A. Pfouga, S. Rulhoff, and J. Stjepandić, "Knowledge-based design for assembly in agile manufacturing by using data mining methods," *Advanced Engineering Informatics*, vol. 33, pp. 285-299, 2017. doi: 10.1016/j.aei.2016.12.006.

[61] D. Ali, M. B. Hayat, L. Alagha, and O. K. Molatlhegi, "An evaluation of machine learning and artificial intelligence models for predicting the flotation behavior of fine high-ash coal," *Advanced Powder Technology*, vol. 29, no. 12, pp. 3493-3506, 2018. doi: 10.1016/j.apt.2018.09.032.

[62] S. Zahraee, M. K. Assadi, and R. Saidur, "Application of artificial intelligence methods for hybrid energy system optimization," *Renewable and Sustainable Energy Reviews*, vol. 66, pp. 617-630, 2016. doi: 10.1016/j.rser.2016.08.028.

[63] D. V. Dao, H.-B. Ly, S. H. Trinh, T.-T. Le, and B. T. Pham, "Artificial intelligence approaches for prediction of compressive strength of geopolymer concrete," *Materials*, vol. 12, no. 6, Art. no. 983, 2019. doi: 10.3390/ma12060983.

[64] T. Moussa, S. Elkatatny, M. Mahmoud, and A. Abdulraheem, "Development of new permeability formulation from well log data using artificial intelligence approaches," *Journal of Energy Resources Technology*, vol. 140, no. 7, Art. no. 072903, 2018. doi: 10.1115/1.4039270.

[65] M. Izadi, M. Rahimi, and R. Beigzadeh, "Evaluation of micromixing in helically coiled microreactors using artificial intelligence approaches," *Chemical Engineering Journal*, vol. 356, pp. 570-579, 2019. doi: 10.1016/j.cej.2018.09.052.

[66] G. K. Parapuram, M. Mokhtari, and J. B. Hmida, "Prediction and analysis of geomechanical properties of the Bakken Shale using artificial intelligence and data mining," in *Unconventional Resources Technology Conference (URTeC)*, Austin, TX, USA, pp. 2815-2833, 2017. doi: 10.15530/urtec-2017-2692746.

[67] M. Chen, U. Challita, W. Saad, C. Yin, and M. Debbah, "Artificial neural networks-based machine learning for wireless networks: A tutorial," *IEEE Communications Surveys & Tutorials*, vol. 21, no. 4, pp. 3039-3071, 2019. doi: 10.1109/COMST.2019.2926625.

[68] Q. Wang and P. Lu, "Research on application of artificial intelligence in computer network technology," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 33, no. 5, Art. no. 1959015, 2019. doi: 10.1142/S0218001419590158.

[69] O. Debauche, S. Mahmoudi, S. A. Mahmoudi, P. Manneback, and F. Lebeau, "Edge computing and artificial intelligence semantically driven: Application to a climatic enclosure," *Procedia Computer Science*, vol. 175, pp. 542-547, 2020. doi: 10.1016/j.procs.2020.07.077.

[70] K. Lin, C. Li, D. Tian, A. Ghoneim, M. S. Hossain, and S. U. Amin, "Artificial-intelligence-based data analytics for cognitive communication in heterogeneous wireless networks," *IEEE Wireless Communications*, vol. 26, no. 3, pp. 83-89, 2019. doi: 10.1109/MWC.2019.1800351.

[71] Z. Allam and Z. A. Dhunny, "On big data, artificial intelligence and smart cities," *Cities*, vol. 89, pp. 80-91, 2019. doi: 10.1016/j.cities.2019.01.032.

[72] V. Özdemir and N. Hekim, "Birth of Industry 5.0: Making sense of big data with artificial intelligence, the Internet of Things and next-generation technology policy," *OMICS: A Journal of Integrative Biology*, vol. 22, no. 1, pp. 65-76, 2018. doi: 10.1089/omi.2017.0194.

[73] M. Bunz and L. Janciute, *Artificial Intelligence and the Internet of Things: UK Policy Opportunities and Challenges*. London, U.K.: University of Westminster Press, CAMRI Policy Brief, 2018.

[74] A. Arsalan and R. A. Rehman, "Interest broadcasting and timing attack in IoV (IBTA-IoV): A novel architecture using Named Software Defined Network," *Computer Networks*, vol. 213, Art. no. 109121, 2022. doi: 10.1016/j.comnet.2022.109121.

[75] M. Burhan, H. Alam, A. Arsalan, R. A. Rehman, M. Anwar, M. Faheem, and M. W. Ashraf, "A comprehensive survey on the cooperation of fog computing paradigm-based IoT applications: Layered architecture, real-time security issues, and solutions," *IEEE Access*, vol. 11, pp. 73303-73329, 2023. doi: 10.1109/ACCESS.2023.3294479.

[76] M. U. Sattar, R. A. Rehman, A. Arsalan, and B.-S. Kim, "Mitigation of interest flooding attack and controlled packet propagation in Named Internet of Vehicles," *IEEE Access*, vol. 12, pp. 125398-125415, 2024. doi: 10.1109/ACCESS.2024.3452281.

[77] V. Dignum, "Ethics in artificial intelligence: Introduction to the special issue," *Ethics and Information Technology*, vol. 20, no. 1, pp. 1-3, 2018. doi: 10.1007/s10676-018-9450-z.

[78] T. Jameel, R. Ali, and I. Toheed, "Ethics of artificial intelligence: Research challenges and potential solutions," in *Proc. 3rd Int. Conf. Computing, Mathematics and Engineering Technologies (iCoMET)*, pp. 1-6, 2020. doi: 10.1109/iCoMET48670.2020.9073911.

[79] V. K. Singh, E. André, S. Boll, M. Hildebrandt, and D. A. Shamma, "Legal and ethical challenges in multimedia research," *IEEE MultiMedia*, vol. 27, no. 2, pp. 46-54, 2020. doi: 10.1109/MMUL.2020.2994823.

[80] G. D. Hager, A. Drobnis, F. Fang, R. Ghani, A. Greenwald, T. Lyons, et al., "Artificial intelligence for social good," *arXiv preprint arXiv:1901.05406*, 2019.

[81] A. Thieltges, F. Schmidt, and S. Hegelich, "The devil's triangle: Ethical considerations on developing bot detection methods," in *Proc. AAAI Spring Symposium Series*, pp. 253-257, 2016.

[82] A. Kuleshov, A. Ignatiev, A. Abramova, and G. Marshalko, "Addressing AI ethics through codification," in *Proc. Int. Conf. Engineering Technologies and Computer Science (EnT)*, pp. 24-30, 2020. doi: 10.1109/EnT48576.2020.00011.

[83] A. Etzioni and O. Etzioni, "Incorporating ethics into artificial intelligence," *The Journal of Ethics*, vol. 21, no. 4, pp. 403-418, 2017. doi: 10.1007/s10892-017-9252-2.

[84] T. W. Kim, T. Donaldson, and J. Hooker, "Mimetic vs. anchored value alignment in artificial intelligence," *arXiv preprint arXiv:1810.11116*, 2018.

[85] P.-A. Gourraud and F. Simon, "Differences between Europe and the United States on AI/digital policy: Comment response to roundtable discussion on AI," *Gender and the Genome*, vol. 4, Art. no. 2470289720907103, 2020. doi: 10.1177/2470289720907103.

[86] A. Agrawal, J. Gans, and A. Goldfarb, *Prediction Machines: The Simple Economics of Artificial Intelligence*. Boston, MA, USA: Harvard Business Review Press, 2018.

[87] M. B. Hoeschl, T. C. Bueno, and H. C. Hoeschl, "Fourth industrial revolution and the future of engineering: Could robots replace human jobs? How ethical recommendations can help engineers rule on artificial intelligence," in *Proc. 7th World Engineering Education Forum (WEEF)*, pp. 21-26, 2017. doi: 10.1109/WEEF.2017.8466973.

[88] E. Ernst, R. Merola, and D. Samaan, "Economics of artificial intelligence: Implications for the future of work," *IZA Journal of Labor Policy*, vol. 9, 2019. doi: 10.2478/izajolp-2019-0004.

[89] F. Liu, Y. Shi, and Y. Liu, "Intelligence quotient and intelligence grade of artificial intelligence," *Annals of Data Science*, vol. 4, no. 2, pp. 179-191, 2017. doi: 10.1007/s40745-017-0109-0.

[90] N. Kumar, N. Kharkwal, R. Kohli, and S. Choudhary, "Ethical aspects and future of artificial intelligence," in *Proc. Int. Conf. Innovation and Challenges in Cyber Security (ICICCS-INBUSH)*, pp. 111-114, 2016. doi: 10.1109/ICICCS.2016.7542339.

[91] D. Gunning, "Explainable Artificial Intelligence (XAI)," Defense Advanced Research Projects Agency (DARPA), Arlington, VA, USA, Program Information, 2017.

[92] D. Fernández-González and C. Gómez-Rodríguez, "Faster shift-reduce constituent parsing with a non-binary, bottom-up strategy," *Artificial Intelligence*, vol. 275, pp. 559-574, 2019. doi: 10.1016/j.artint.2019.07.006.

[93] Y. Lin, X. Dong, L. Zheng, Y. Yan, and Y. Yang, "A bottom-up clustering approach to unsupervised person re-identification," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 33, no. 1, pp. 8738-8745, 2019. doi: 10.1609/aaai.v33i01.33018738.

[94] M. Hutson, "Artificial intelligence faces reproducibility crisis," *Science*, vol. 359, no. 6377, pp. 725-726, 2018. doi: 10.1126/science.359.6377.725.