Bridging the Gap Between Ethical AI Implementations

Yazan Alahmed[1*], Reema Abadla[2], Nardin Ameen[3], Abdulla Shteiwi[4]

[1,2,3,4]*Faculty of Engineering, Al Ain University; Email: yazan.alahmed@aau.ac.ae, reemaabadla@gmail.com, nardin.ameen@aau.ac.ae, abdullashteiwi@gmail.com*

Corresponding author: Yazan Alahmed (yazan.alahmed@aau.ac.ae)

Abstracts: This paper aims to investigate the existence of artificial intelligence (AI) and its employment has been around for decades, and while it has optimized and advanced various domains of human labour, its developing capabilities have gradually extended into sectors that involve human behaviour and cognition. Naturally, its rapid integration into fields that require unique human behavior, such as creative writing in educational settings, critical thinking, empathetic communication in therapeutic contexts, and many more, has raised ethical issues and concerns from many different aspects throughout the years. In this paper, the ethical implications of AI across various domains and its consequences on societal values, privacy considerations, and human rights are thoroughly examined. Moreover, it highlights the need for robust guidelines, criticizing current frameworks for their lack of enforceability. Examining real-world cases, like IBM Watson's errors and Tesla's autopilot incidents, the paper stresses the urgency for practical and enforceable solutions. A systematic literature review methodology was applied to identify, evaluate, and synthesize existing literature on AI ethics, encompassing aspects like education, healthcare, and social interaction. The findings show the urgent need for robust ethical frameworks that prioritize societal values, privacy, and human rights.

Keywords: Artificial Intelligence, Ethics, Human Rights, Privacy

1. INTRODUCTION

In recent years, Artificial Intelligence (AI) has transformed from a valuable asset, enhancing human productivity across various sectors, into a technology weighted with potential threats. Its rapid evolution, which was once promising, now raises ethical concerns that infiltrate every domain of human interaction – healthcare, social media, finance, and transportation alike. Despite its potential, ethical considerations are very often relegated to an afterthought in technological innovation, rather than being an integral part of the design process. This shift has alarmed tech industry leaders; figures like Elon Musk recently signed an open letter urging the suspension of major AI experiments for six months, claiming that the current rapid advancement of AI systems is happening with no thought or guarantee of the positivity of their impacts or the safety of their applications, and noting their potential risks to humanity [1]. This paper explores the profound ethical issues raised by the quick development of AI. It analyses current approaches and, in a critical manner, draws attention to their shortcomings. The significant gap between current ethical practices and their enforceability is central to the discussion. The ethical ramifications of AI across different domains can be distilled into three primary dimensions as seen in Figure 1: societal values, privacy considerations, and implications for human rights. The figure displays an overview of some of the core components that fall under each dimension. Such categorizations stem from in-depth discussions, real-world observations, and scholarly works reflecting some of the most prevalent ethical concerns. The impact of AI on societal values can be analyzed by considering responsibility, accountability, transparency, addressing biases, and ensuring fairness. As for privacy, AI's impact on it can be viewed through the lens of consent, transparency, and the degree as to which private data is collected and accessed. Lastly, AI's impact on human rights can be considered from the perspective of various human rights including the right to privacy, non-discrimination, equitable access, etc. While ethical values may vary by region, certain moral principles remain consistent across societies and cultures. This paper critically examines existing ethical solutions to the issues arising from AI implementation, the inadequacy of current solutions and emphasizes the lack of enforceability in addressing these concerns. The primary objective of this paper is to shed light on the existing gaps and propose solutions that not only bridge these gaps but also pave the way for enforceable ethical practices in AI technology. By exploring the shortcomings of current approaches and proposing practical solutions, the paper aims to advocate for a more rigorous and accountable framework within companies and businesses, fostering a future where ethics in AI is taken with the seriousness it deserves. The paper is structured as follows. The existing literature on ethical solutions in AI is discussed in section 2. In section 3, a critical analysis of the pros and cons of these existing solutions is discussed. After which the methodology and results are presented in sections 4, 5, and 6.

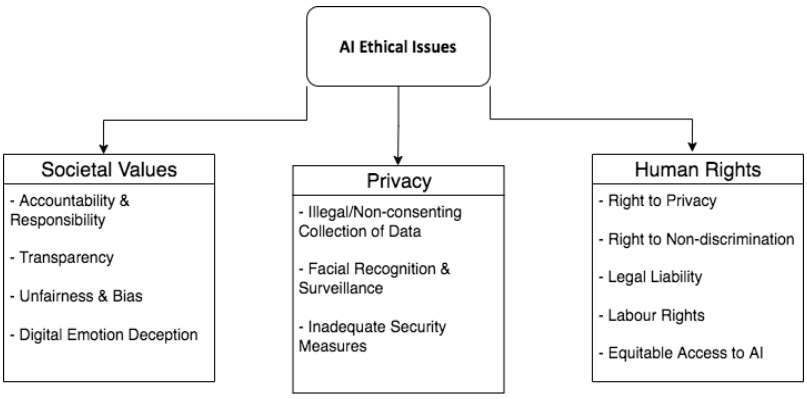

Figure 1. Overview of Ethical Issues in AI.

2. LITERATURE REVIEW

Multiple strategies have been proposed to ensure ethics in AI, both theoretical and technical dimensions. In this section, the literature on such approaches will be delved into and explored to highlight their significance in influencing the ethical climate of artificial intelligence and creating a more accountable and responsible AI ecosystem. These approaches serve as guiding principles for addressing AI's complex challenges.

2.1. Governmental And Organizational Initiatives

A prevalent approach for fostering ethical AI is the creation of guidelines and frameworks. Many governments, including the U.A.E., the EU, some states in the U.S., several countries in Asia (such as South Korea and Singapore), Canada, and more, have taken the initiative to introduce guidelines or proposals for regulations on ethical AI [2, 3, 4, 5, 6]. In addition, prominent tech companies like IBM, Microsoft, and Google have also published their guidelines for ethical AI [7, 8, 9]. Common principles found across all those guidelines and frameworks introduced by governments and organizations are transparency and trust, accountability, privacy, fairness, and the creation of environmentally friendly AI. The present ethical guidelines and frameworks formed by organizations, governments, institutions, and economic powers are numerous and will not be discussed in detail in this paper.

2.2. Multidimensional Governance in AI

In [10], a multidimensional governance system is proposed. The authors discuss two theories in the selection of AI governance, namely “opposition theory” and “system theory”. The former focuses on issues that arise between AI technology and human rights and emphasizes the need to build regulations,

standards, and penalties to protect the rights and welfare of users by focusing on transparency, privacy, fairness, etc. The latter, on the other hand, is a broader, more holistic theory that involves other social aspects of AI interactions, including education, human interaction, training, infrastructure, etc., and aims to change various components of the sociotechnical system to find systemic solutions. An important distinction is that the opposition theory seeks to protect the users’ interests, rights, and welfare while also studying the potential harms AI systems can have on individuals and society. The system theory, however, takes into consideration the interests of organizations and governments in addition to those of users.

The multidimensional system the authors propose is based upon the system theory and involves practices like education reform, ethics and codes of conduction formulation, technical support, legal regulation, and international cooperation. A significant point brought up by the authors under educational reform is that, in addition to training and educating people about AI, there is a need to take measures to create a system where workers can continue to learn and upgrade their skills throughout their careers to adapt to changes brought about by AI and prevent the loss of their jobs. In addition, a national ethics committee is projected for the creation of AI-specific ethical codes, the assessment of new ethical risks that may arise with new AI ventures, and the endorsement of industry best practices. The paper also encourages companies in different fields to create and publish their own ethical AI

standards in an effort to increase participation in the formation of international standards. In terms of legal regulation, this paper advocates the development of digital human rights within legal frameworks, emphasizes the need for clear responsibility allocation (i.e., identify who is responsible for what in the context of AI), and notes the importance of combining legal regulations with technical measures to ensure responsible AI. It calls for the effective implementation of existing laws related to personal information protection and data security and the creation of new legislation, specifically for autonomous driving. Regulation is also discussed regarding AI algorithms, as the need for measures like ethical AI assessments, algorithm verification and reviews, as well as dataset quality checks.

In other work [11], the authors address the issues of unfairness, bias, and inaccuracy in AI systems, particularly in education systems, where the solutions suggested coming from the programmers of the system, teachers, and users. The paper emphasizes the need for regulation in the design stage of AI systems by programmers, also stating that programmers should put people's well-being and human rights first during robotic designs. In addition, the need for frequent training courses for programmers of AI systems in the ethics of AI to guarantee the absence of bias, prejudice, or illegitimacy in AI algorithms is highlighted in the paper. Such courses are also necessary for the teachers who teach students AI materials. Moreover, the paper addresses the significance of including ethical AI courses in universities and schools for students of both computer-related majors and non-computer-related majors. This is a crucial task in ensuring the cultivation of a moral, righteous environment in the field of AI, regardless of which industry it is used in. However, the authors do note that those with computer-related majors should have stricter, more detailed, and more major courses for ethics in AI than others due to their more probable future direct engagement with such systems.

Two innovative solutions have been put forward to address issues related to fairness and bias in AI applications. A method designed to mitigate gender bias in word embeddings (often employed in natural language processing) is proposed by [12], where it is ensured that AI language models don't perpetuate stereotypes or biases related to gender, promoting more equitable language understanding and generation. For example, it can help AI systems avoid associating certain professions or roles with specific genders. In [13], a solution based on social welfare functions, which are utilized to implement fairness in AI systems' reward mechanisms. This approach aims to tackle the fairness of resource allocation, a crucial aspect of AI, within the framework of deep reinforcement learning (It combines the tasks of approximating functions and optimizing targets by linking states and actions to the rewards they result in [14]). For instance, it can be applied to scenarios where AI systems allocate resources, such as medical treatment, based on various criteria. The goal is to ensure that resource allocation is done fairly and without discrimination, considering the needs and rights of all individuals. Nevertheless, the use of deep reinforcement learning is opposed in [15] where the authors state that it poses the concern that a malicious agent could find multiple ways to evade ethical constraints set by manipulating the reward system, therefore compromising the safety of the algorithmic process. One such method is reward hacking, where the AI system exploits weaknesses in how rewards are determined to gain more rewards than originally intended. For instance, if a robot is trained to pick up red apples and a reward system where the robot receives a point each time it successfully picks up a red apple is set up, the robot can start picking up red apples and then immediately drop them and pick them up again repeatedly to earn more points without performing the task. The authors also propose solutions to ethical problems in AI, including the integration of ethical principles throughout the entire engineering process, as well as the need for interdisciplinary collaboration when it comes to researching AI governance and designing AI systems. The former begins with the requirement analysis phase, where AI tasks are meticulously defined, and ethical guidelines are tailored to specific applications. During data collection and preparation, engineers ensure data quality by eliminating flawed samples and minimizing biases, resulting in a well-balanced dataset. Subsequently, in the model design and training phase, the focus remains on adhering to ethical criteria, after which the initial model undergoes rigorous testing against ethical benchmarks like fairness, robustness, transparency, and task performance. Any failures mean a re-evaluation of the model, data, and training must be done. However, a successful model proceeds to integration with other software components and deployment within the intelligent system. Finally, continuous runtime monitoring takes place to ensure ethical compliance and any ethical violations must be met with further enhancements of the model.

The issue with multidimensional approaches and educational reforms is that resistance or lack of cooperation from industries and organizations may be encountered. Implementing systemic changes across diverse sectors

can be logistically challenging and may face opposition from stakeholders with conflicting interests. Moreover, in terms of educational reforms, curriculum constraints must be made more flexible and continuously updated.

2.3. Technical Approaches & Tools

In a more technical approach, the IMDA (Infocom Media Development Authority) in Singapore created a software toolset and testing framework for the purpose of AI governance called AI Verify [16]. The framework is made up of 11 AI ethical principles (including transparency, explainability, safety, fairness, accountability, human oversight, and more) that are congruent with globally recognized AI frameworks, including those from the EU, OECD, and Singapore's Model AI Governance Framework, and on which jurisdictions from across the world agree. Furthermore, IMDA established a non-profit foundation to act as an open-source platform for communities to share ideas and work on the government and testing of AI systems. Similarly, another toolkit [17] called AI Explainability 360 supports transparency and explainability by offering eight advanced explainability techniques and two evaluation metrics, making it a comprehensive resource for those who want to understand how AI models arrive at their decisions. The taxonomy offered is especially helpful in guiding users through the landscape of explanation techniques, including those from the toolkit as well as those from the larger area of explainability research. To make it user-friendly, the toolkit is built on a flexible software architecture that categorizes these methods based on their relevance within the AI model development process. It's a valuable tool for data scientists and others looking to enhance the transparency and interpretability of AI models. Overall, the challenge with such tools is that may miss distinct real-world ethical challenges and rely on varied user understanding, impacting their effectiveness.

Additionally, in a recent article by [18], a software capable of assessing AI system’s acquisition of knowledge from a company’s database and determining if there are any potential vulnerabilities in software code that can be exploited by an AI system. Knowing what AI systems know through this software provides reliability and transparency, helps in enumerating knowledge gaps in an AI system, and offers suggestions for development. It also improves privacy and security as the software can detect if an AI system is using or accessing sensitive information. However, this approach can simultaneously be risky in revealing sensitive data due to detailed information access and should therefore be used with robust privacy controls.

In other work [19], a few approaches to solve ethical issues in AI. One of them is the minimization of negative side effects during machine operation. By incorporating advanced predictive algorithms and real-time monitoring, the proposed approach aims to detect potential negative outcomes before they occur. For instance, if a robot aims to move from point A to point B, this stance allows it to adjust its actions or path in real time to avoid negative or unintended consequences like damage to valuable objects, even when pursuing its objectives. Reward hacking, where optimization techniques find unanticipated ways to optimize a fitness function (i.e., evaluation function) that doesn’t match the intended goal or objective, is another issue noted. A fitness function is a computational formula that measures the quality or performance of an action taken in machine learning terms, assisting algorithms to find the best solutions based on certain criteria [20]. The solution proposed is a multipronged approach where the robot is encouraged to mimic expert behavior, collaborate with humans to fulfil their objectives, and train the machine learning (ML) model via reward modelling techniques (providing the model with reward signals based on human preferences [21]) to ensure that it operates in alignment with human preferences. Another strategy proposed in the paper is “safe exploration”, where new solutions are navigated while ensuring that harmful actions are avoided through specifying clear, definite objectives. For instance, in the case of a moving robot, the agent may undergo fitness penalties for breaking something while moving from point A to point B but learns, through training, to avoid such actions. In that way, the AI agent builds an ideal strategy to stop harmful acts after learning the dynamics of its goals. Lastly, the paper highlights the need to enhance AI systems robustness through strategies that can skillfully manage deviations from their training data in terms of encountering unexpected scenarios. It emphasizes the significance of developing new benchmark tasks to assess and enhance the resilience of AI systems in real-world contexts and advises utilizing exploration networks to gather insights. The paper's proposed solutions play a pivotal role in mitigating ethical concerns in the field of AI. By encouraging safe exploration, AI systems are less likely to engage in harmful or unintended actions, aligning their behavior with ethical standards. Additionally, reward modelling techniques ensure that AI models adhere to human preferences, promoting ethical decision-making. Robustness enhancements allow AI systems to navigate unexpected

scenarios responsibly, reducing the risk of unethical actions. These solutions collectively contribute to the development of more ethical AI systems, making their applications more trustworthy and reliable. A potential challenge with such advanced solutions is the need for high costs and skilled workforce, which may hinder smaller organizations and developing countries.

In summary, in synthesizing these diverse efforts, key common threads emerge, highlighting the human-centric focus of ethical AI initiatives. Across many fields and global contexts, an emphasis on transparency, fairness, and accountability prevails. Moreover, literature consistently advocates for interdisciplinary collaboration, recognizing the complexity of ethical challenges. In addition, continuous learning and adaptation stand out as essential themes, emphasizing the need for ongoing education and awareness. The key to tackling the complex ethical field of AI is to use an integrated, holistic approach that incorporates technological, social, and cultural factors. These links work together to shed light on a responsible AI development path that cuts beyond corporate and regional borders.

3. METHODOLOGY

Researches in AI ethics sector start from reflections on how ethical guidelines can be implemented in decision routines of autonomous systems over meta-studies about AI ethics or the empirical analysis on how trolley issues are solved to reflections on certain challenges and comprehensive AI guidelines [15]. This thesis focuses on the latter issue. The ethics guidelines list touched on this paper therefore includes compilations that cover the path of artificial intelligence ethics as comprehensively as possible. This paper doesn't aim at a full analysis of every available soft-law or non-legal norm document on AI, algorithm, robot, or data ethics, but rather a partial systematic overview of issues and normative stances in the field, showing how AI ethics details relate to a larger image.

To respond to alarming questions on "what to do to embrace ethnicity in AI systems", the author of this research applied a systematic literature review methodology to carry out the research. This is a rigorous and comprehensive research method used in academia and various fields to identify, evaluate, and synthesize existing literature on a specific topic or research question [20]. It aims to provide an unbiased, structured, and evidence-based summary of the current state of knowledge in a particular area. The author applied PRISMA rules to implement the technique of planning and realizing this research. Thereafter, a protocol regarding the review was conducted and registered. In this methodology, we followed the right scheme of processes (PRISMA rules) like literature search strategy, eligibility technique, the selection procedure, quality assessment and data synthesis as elaborated below. Three critics were involved in the study process. First and second independent critics performed the selection of the PRISMA and screening tasks in parallel. This reduced the chances of bias in both selection and reporting. The other critic validated the research process.

a) LITERATURE SEARCH STRATEGY

The following five databases' samples were presented for literature research: PubMed, Association for Computing Machinery (ACM), Institute of Electrical and Electronics Engineers (IEEE) Xplore, Nature-SCI and the AI ethics Guidelines Global Inventory. ACM, PubMed, and Nature-SCI are academic databases while on the other hand AI Ethics Guidelines Global Inventory is a searchable inventory of published framework that notes ethical issues of AI. Documents addressing the following aspects were selected: Education, Ethics and guidelines, and artificial intelligence. This literature search was done through an automatic search in each search engine listed, using the key terms and synonyms. After doing the review, it is defined that the period of search as from 2013-2023.

Table 1. Keywords and synonyms

Keywords	Synonyms
Artificial Intelligence	"Artificial Intelligence" OR "AI" OR "Machine Learning" OR "ML" OR "Deep Learning" OR "Artificial Neural Networks" OR "Computer Vision"
Healthcare	"Healthcare" OR "Medical care" OR "care" OR "Diagnosis processes" OR "treatment protocol development" OR "drug development" OR "personalized medicine" OR "patient monitoring" OR "Medicine" OR "Radiology" OR "Pathology" OR "Decision Support Systems"
Ethics	"Ethics" OR "Fairness" OR "Integrity" OR "Virtues" OR "Value-System" OR "Ethical Values" OR "Rightness" OR "Moral" OR "Morality"
Guidelines	"Guidelines" OR "Recommendations" OR "Instructions" OR "requirements" OR "principles" OR "regulations" OR "suggestions" OR "advice" OR "Rules" OR "standard" OR "criteria"

b*)* **ELIGIBILITY CRITERIA**

The eligibility criteria are specified as inclusion and exclusion criteria [20], as displayed in table 2. Inclusion and exclusion criteria defined the characteristics that prospective subjects must have if they are to be included in the study. Inclusion criteria embraced subjects in which their defined characteristics match the intended course, while exclusion criteria embraced subjects in which their characteristics did not. The author stipulated that the selected documents had to be aligned this way to fit this study.

c*)* **SELECTION PROCESS**

The selection of documents was based on the following steps:

a) Preliminary collection of studies from the database search; here, two independent arbiters applied the search strings to the five selected databases. The arbiters keenly followed the exact period of range that is between 2013 to 2023 and then determine whether the chat was in English form.

b) The next step was screening of titles and abstracts following the eligibility criteria as shown in table 2. The repeating documents were removed. All documents were organized in table with the following columns of reference: Strategies, issues, and healthcare application part. The above issues were synthesized in a list after comparing them. The number of documents referencing every issue was noted.

c) In the third phase, screening of full texts was done regarding inclusion and exclusion criteria. In this phase, a consensus was reached after discussing the reasons for several differences that happened.

Table 2. Inclusion and exclusion criteria (documents published between 1 January 2013 and 18 September 2023).

No.	Inclusion	No.	Exclusion
1.	Written in English	1.	Did not mention ethical issues or guidelines related to AI
2.	Mentions ethical issues related to AI or guidelines	2.	Did not focus on the application area of healthcare
3.	Highlights the application area of healthcare	3.	Did not have more than 10 citations
4.	Published between 1 January 2013 and 18 September 2023		

D) QUALITY ASSESSMENT

In this phase, quality assessment questions were derived according to the quality assessment checklist of Kitchen ham and Charters to measure and execute the quality of the picked basic documents and do away with bias. Every picked primary study was matched with a quality question as described below. Documents addressing quality assessment questions were assigned a full mark while the documents that partially addressed the question

were assigned a half mark.

Table 3: Quality assessment questions.

No.	Quality Questions	Score
1.	Does the adopted research method address the research questions?	1/0.5/0
2.	Does the study have a clear research objective?	1/0.5/0
3.	Does the study have a specific description of each ethical issue?	1/0.5/0
4.	Does the study have a specific description of strategies related to the ethical issue?	1/0.5/0
5.	Do the results of the study add value to the area of research?	1/0.5/0

E) DATA SYNTHESIS

The worldwide landscape of AI ethics guidelines provided by a scholar group (Jobins group) displays 11 aspects based on a scoping review of ethical instructions related to AI solutions applied to general domain. The jobs' group findings are used in this paper. In this case, the author applied the 11 issues, in conjunction with the ethics instances of trustworthy AI provided by the commission from Europe, as a starting point to group the issues identified in this research in the healthcare sector. In this research, the author's goal was to highlight ethical healthcare issues, which may have brought differences due to the differing scope of the recent literature review and its effects. Based on the work by Jobins group, reviewers conducted the thematic code-mapping process. For this scenario, Jobins ethical principles were applied, and the codes highlighted in the existing AI guidelines as the basics and added one code "control" derived from EGTAI. The process of code mapping entailed two iterations of themes. Those themes were ethical issues and related codes. The author used abductive methodology that entailed inductive and deductive techniques, during the code mapping process. The author first applied deductive technique. This technique was aimed at deducting documents selected based on Jobin's idea. The author concluded that ethical issues in the healthcare department have their own focus, which differs from Jobin's work. Due to the difference highlighted, the author also applied an inductive technique in this case with the aim of identifying new ethical instances. The ethical issues realized were different from Jobin's case.

RESULTS

- AI, initially seen as a productivity enhancer, is now raising serious ethical concerns in various domains, including healthcare, social media, finance, and transportation.
- Ethical and moral considerations have often been overlooked in AI design, leading to unintended consequences.
- Prominent figures in the tech industry, such as Elon Musk, have called for the suspension of major AI experiments due to the rapid advancement of AI systems without sufficient consideration of their impacts.
- Ethical ramifications of AI can be categorized into three primary dimensions: societal values, privacy considerations, and implications for human rights.

4. DISCUSSION

In this section, the challenges surrounding ethical guidelines and frameworks in AI systems are discussed. Despite their existence, the lack of enforceability and indifference in certain sectors hinder their effectiveness. The gap between awareness and implementation is explored, emphasizing the need for enforceable regulations to address ethical breaches.

A. Non-technical Approaches

One of the major problems with the approaches present on solving or minimizing ethical issues in AI is the focus on creating guidelines and frameworks. The problem with these guidelines and frameworks lies in the implementation of these principles. These frameworks and guidelines, despite their creation, lack enforceability. While some organizations may choose to comply with them wholly and completely, others may choose partial adherence or even disregard them entirely. In other words, when companies release these guidelines in response to public concerns about AI ethics, it can be likened to adults placating a curious child with a brief diversion in the form of candy. These guidelines, while existing, allow organizations to operate AI without enduring legal scrutiny. This is described by [22] as "toothless principles", where businesses carry out front-page work on ethical

frameworks without worrying that it will fundamentally change the features of their products, the structure of their organizations, or their quarterly profits.

Another core problem which contributes to the futility and ineffectiveness of these guidelines is the indifference of certain members of society towards ethics. This indifference is evident in the academic field of information technology majors like computer science, software engineering, computer engineering, etc., which are notorious for being filled with sexist, misogynist, and ethically lacking males. These students, who eventually enter the industry, bring their non-ethical beliefs and behaviors to the professional field. Endless publicized cases and stories of sexist and racist comments and/or actions made and taken by employees at esteemed, major tech companies prove the ethically toxic culture present in the tech industry [23, 24, 25, 26, 27]. As a result, when AI systems are developed by engineers from within the technology industry, it is unsurprising that these systems may not consistently adhere to established ethical standards. As highlighted in [22], Universities and colleges hyper-fixate on teaching the technical aspects of majors like software engineering, computer science, cybersecurity, etc. and others that can be used to create AI systems but forget to teach and raise awareness of the necessary ethics that go together with technological implementations (i.e., educational reform).

Unfortunately, the rapid progress of AI technology is occurring at a much faster pace than the establishment of thorough and strict ethical guidelines and understanding, resulting in a significant gap between the potential risks and benefits posed by these innovations. Having said that, it is imperative to emphasize that this observation does not diminish the significance of these guidelines and frameworks. They represent a commendable step in acknowledging that organizations and their teams are at least aware of the ethical concerns surrounding AI usage. While this marks a crucial starting point for the discourse, it remains very far from sufficient in addressing the core issue at hand, namely, the absence of enforceable regulations, legislation, and punitive measures related to ethical breaches by AI systems and their creators.

In the next section, AI ethics strategies that provide proactive risk management and adaptability but also present complexities, including intricate technical solutions, potential vulnerabilities, and resource intensiveness, are discussed. It highlights the need for a balance between technical advancements and ethical regulations.

B. Technical Approaches

While technical approaches present promising avenues for addressing ethical challenges in AI, they bring both advantages and complexities to the table. The strategy of minimizing negative side effects by employing advanced predictive algorithms and real-time monitoring offers proactive risk management, enabling AI systems to adapt and avoid unintended consequences. This adaptability is particularly crucial in fast-paced industries where AI operates. However, there is a growing concern about the potential exploitation of these systems, such as the occurrence of reward hacking, where optimization techniques can manipulate reward mechanisms, posing a safety risk that needs careful consideration. Moreover, the complexity of implementing technical measures may present scalability challenges, as not all AI applications or industries may readily accommodate these approaches, and customization may be needed to fit specific use cases. Additionally, solutions like safe exploration help in building ideal strategies but require a proper balance between exploration and responsible behavior to avoid harm, making their implementation intricate. Furthermore, the emphasis on enhancing AI robustness by navigating unexpected scenarios underscores the need for benchmark tasks and exploration networks, yet implementation may demand considerable resources in terms of computing power and data requirements.

To create a thorough AI ethics framework, these technical solutions should ideally complement strong, enforceable guidelines and regulations, as highlighted in the non-technical approaches section. Moreover, collaboration between AI developers, ethicists, and regulatory bodies to create standardized, adaptable technical solutions accessible to organizations of all sizes should be encouraged. In summary, the following points describe some of the main challenges with the technical solution posed:

A. Complexity of Technical Solutions:

Some technical approaches can be highly intricate, making implementation challenging, especially for smaller organizations or teams with limited resources. This complexity may lead to slower adoption rates and potential disparities in AI ethics practices.

B. Potential for Manipulation:

The use of deep reinforcement learning, while promising, can introduce vulnerabilities. The concern of malicious agents exploiting reward systems to evade ethical constraints poses a risk to algorithmic safety and reliability, highlighting the need for robust defenses against such manipulation.

C. Resource Intensiveness:

Implementing technical solutions like real-time monitoring, fairness algorithms, and AI explainability can be resource-intensive in terms of computational power and data requirements. This could limit their accessibility to organizations with limited resources.

5. COMPARATIVE ANALYSIS

The challenges posed by AI ethics are not confined to a single domain but permeate various sectors. In the field of healthcare, for example, IBM Watson's supercomputer, designed to assist medical professionals, once provided incorrect cancer treatment recommendations, raising questions about accountability [28]. Clinicians have a regulatorily enforced professional requirement to be able to account for their actions, whereas technologists do not; instead, ethical codes of practice are employed in this sector. This issue extends beyond healthcare and is strikingly evident in the transportation sector, as exemplified by Tesla's autopilot incidents. In cases where drivers in fatal accidents involving Tesla's autopilot software aren't charged guilty [29], questions surrounding accountability become pronounced, emphasizing the need for a proper examination and new legal frameworks for such situations. Moreover, the overreliance on AI has sparked ethical concerns across various industries, including healthcare and finance. A substantial 90% of organizations' executives have voiced their apprehensions, highlighting the critical ethical issues emerging due to the increasing reliance on AI systems [30].

Another concern across many domains is digital emotion deception, which involves AI simulating human emotions (like in chatbots providing scripted responses). This can pose serious risks, especially in healthcare, undermining values like social interaction and mental health. From companion robots for children and the elderly [31, 32] to Alexa emulating real people's voices [33], the gradual removal of actual human beings from tasks that require human empathy and companionship and the overreliance on such AI systems by vulnerable users can bring about false expectations, disappointment, social isolation, and dehumanization, negatively impacting their well-being. Ethical constraints and human involvement in designing and testing such systems are crucial to ensure safety. For example, in mental health, nurses should evaluate AI systems [34]. Similarly, the same principle of a "person-centered approach" instead of a "person-like solution" [34] should be applied to other fields as well. It is important to note that chatbots themselves are not problematic; it's their application in certain fields, like therapy and education, that raises health and academic integrity concerns [35] respectively, affecting social abilities.

Bias and unfairness represent another significant challenge in AI systems across various domains. These issues often stem from historical data reflecting societal prejudices, forming the basis of AI algorithms [36]. For instance, facial recognition systems predominantly trained on one racial or ethnic group may inaccurately identify individuals from other groups, further amplifying biases [37]. Bias can also originate in algorithm design and development, unintentionally favoring certain features or encoding societal prejudices [36]. In domains such as lending, hiring, and healthcare, AI models may perpetuate discrimination against demographic groups, resulting in inequities. Notable examples include COMPAS disproportionately labeling Black defendants as high risk, exacerbating sentencing disparities [38], and healthcare algorithms exhibiting reduced accuracy for Black patients, potentially causing delayed diagnoses [36]. Gender bias is another concern, affecting various fields. In cardiovascular disease diagnostics, gender differences necessitate AI adjustments [39]. Overreliance on predominantly gender-biased training data can lead to inaccurate risk assessments, particularly for women. Additionally, AI bias has manifested in employment, exemplified by Amazon's automated hiring system, which favored male candidates and penalized terms associated with women due to biased training data [40].

6. IMPACT ASSESSMENT

a. Privacy Considerations

The field of AI-driven solutions has brought forth intricate challenges regarding privacy, including aspects such as illegal data collection, facial recognition, Voice Assistants (VA), and other concerns, all of which necessitate thorough examination of such systems.

Privacy in the context of AI refers to the protection of individuals' personal information, ensuring that their sensitive information is not inappropriately accessed, used, or disclosed without their consent [41]. The traditional approaches to privacy protection may not be able to properly address the complexities posed by data analytics, which often involves processing vast datasets to identify patterns, trends, and correlations among them. Moreover, the growth of data sharing practices, whether voluntary or involuntary, raises concerns about user consent, data ownership, and the possibility of combining data from different sources to create a comprehensive profile of the individual using data from multiple sources.

i. Illegal Data Collection

Increasingly, AI relies on unauthorized data collection practices, encompassing data scraping, unauthorized personal information gathering, and illegal surveillance [42]. This raises significant risks, from privacy breaches to data misuse and identity theft. An example was Cambridge Analytica's scandal, which involved improper data collection via an AI-driven app on Facebook. It harvested data from users and their friends without consent, affecting up to 87 million people. The data was used to create psychographic profiles and allegedly influence political campaigns, raising ethical concerns [43]. The real-world impact of such breaches extends beyond data misuse; they directly affect individuals' lives, their choices, and even the democratic processes they are part of, raising significant ethical and societal concerns.

ii. Biometrics & Smart Home Devices

Today, entities often misuse technologies like Facial Recognition, Surveillance, and Smart Home Devices for unauthorized data collection, bypassing necessary permissions. For instance, during the Hong Kong protests in 2019, authorities used facial recognition technology for surveillance and crowd control, raising concerns about privacy, free speech, and civil liberties [44, 42]. In practical terms, this means individuals participating in peaceful demonstrations could face repercussions based on their identification through AI-powered surveillance systems, highlighting the real-world implications of these technologies on personal freedoms and democratic rights.

In another case [45], Amazon's involved human contractors analyzing Alexa recordings without clear user consent, though no illegal data collection occurred. It prompted calls for transparency, user control, and informed consent in voice assistant technologies, leading to improved privacy settings and data information by tech companies.

iii. Cross-referencing Data

Organizations use AI techniques to cross-reference legally obtained datasets for personalized advertising, as seen with Facebook's ad targeting [46]. Facebook compiles extensive user profiles by collecting data from various sources, including online behavior tracking and third-party data integration. While this practice is legal, it has raised ethical concerns regarding user privacy due to the depth of ad targeting achieved through cross-referencing. In daily life, this translates into a bombardment of targeted advertisements influencing individuals' choices and behaviors without explicit consent. One solution to address this issue is the implementation of stricter regulations and transparency requirements governing the collection, aggregation, and use of personal data in advertising. Additionally, user-centric privacy controls and clear consent mechanisms should empower individuals to have more control over how their data is utilized for targeted advertising.

b. Human Rights Implications

As AI integration becomes increasingly pervasive in various aspects of society, it is paramount to assess the real-world impact of the ethical concerns it poses in terms of human rights, particularly in the areas of privacy, non-discrimination, accountability, labor rights, and equitable access to AI resources.

The right to privacy, a fundamental human right, faces violations through AI systems that engage in non-consensual data collection and facial recognition, suppressing freedom of expression. Non-discrimination, another crucial human right, is frequently compromised by AI systems [36, 38, 39, 40], as they inadvertently perpetuate biases rooted in historical data, often using race, sex, or other characteristics to justify violations of rights and opportunities. These violations have real-world consequences, impacting individuals' experiences and opportunities in various domains. Moreover, AI also poses challenges in terms of legal liability, as traditional notions of accountability hinge on human actions and intentions, raising complex questions when AI systems fail or display biases across various domains. These questions of accountability have practical implications, such as determining responsibility in AI-related incidents.

Additionally, labor rights are potentially undermined by AI's advancement, particularly in terms of job displacement and economic security. While AI offers increased efficiency, it must be balanced with the preservation of employment and livelihoods for workers. As illustrated in [47], by the mid-2030s, one-third of all jobs could be automated, primarily impacting those with lower education levels.

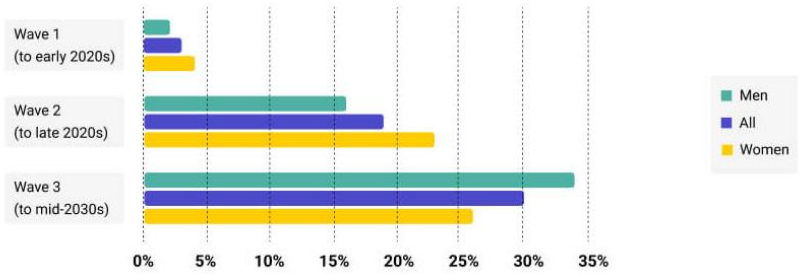

Figure 2. % of Existing Jobs at a Potential Risk of Automation, [38]

The rush to automate certain jobs and showcase AI capabilities can come at a human cost, affecting employment, income inequality, poverty, and social unrest. These impacts are felt in communities and economies, with potential far-reaching consequences. Jobs are not just about earning but also identity and social interaction. Ignoring the human aspects of work, including the sense of identity and social connections it provides, in the pursuit of AI-driven cost-cutting measures highlights the urgent need for ethical awareness in AI deployment. Replacing workers with AI should be approached with careful consideration of these humanitarian concerns, as it influences individuals' well-being and social fabric.

Lastly, equitable access to AI is crucial, ensuring that AI's benefits are accessible to all, regardless of their background, socioeconomic status, or geographic location [48]. Unequal access can perpetuate educational inequalities among students, impacting future job prospects. It is tied to economic opportunity, employment rights, and even privacy rights, as AI-based surveillance might be unevenly deployed in different regions. These disparities in access and opportunity have real-world implications for education, economic prospects, and individual privacy, affecting the fabric of society.

Labor rights and equitable access to AI are two categories of topics that are overlooked when it comes to the discussion of ethics in AI. In [49], this point is supported by the finding that the existing literature on ethical issues in AI rarely focuses on the topics of poverty, labor exploitation, and global inequality, despite being considered significant problems by some. Addressing these issues is essential to safeguard human rights in the age of AI, as they base principles of equality, non-discrimination, education, and economic opportunity, promoting a more just and inclusive society.

CONCLUSIONS

In this paper, a rigorous systematic literature review was conducted using the PRISMA methodology, delving into the ethical intricacies of artificial intelligence in various domains, with a focus on healthcare. The research revealed a myriad of ethical challenges, spanning from biased algorithms to privacy breaches and the nuanced implications on human rights. Through a comparative analysis, the universal nature of these issues was highlighted, transcending specific sectors like healthcare and transportation. The findings emphasize the pressing need for extensive ethical frameworks in AI development. As explored in the discussion, the existing ethical guidelines, while well-intentioned, often lack enforceability, rendering them somewhat futile in mitigating the challenges posed by AI technologies. The implications of our research extend beyond technological realms, infiltrating human rights, societal values, and individual privacy. Moreover, the paper’s analysis sheds light on the often-overlooked domains of labor rights and equitable AI access, emphasizing the importance of addressing these aspects in ethical AI discourse. As the era of AI moves forward, it is crucial for technologists, policymakers, and society at large to collaboratively navigate these ethical challenges. By fostering transparency, accountability, and inclusivity in AI development, it can be ensured that the impact of AI aligns with ethical principles, safeguarding the rights and well-being of all individuals.

REFERENCES

[1] K. Fitzgerald, "Elon Musk joins tech leaders in call for pause on 'giant AI experiments,'" The National, Mar. 29, 2023. https://www.thenationalnews.com/business/technology/2023/03/29/elon-musk-joins-tech-leaders-in-call-for-pause-on-giant-ai-experiments/ (accessed Sep. 06, 2023).

[2] United Arab Emirates Minister of state for Artificial Intelligence, Digital Economy, and Remote Work Applications Office, "AI ETHICS PRINCIPLES & GUIDELINES," 2022. Accessed: Sep. 02, 2023. [Online]. Available: file:///Users/a17226953/Downloads/MOCAI-AI-Ethics-EN-1.pdf

[3] AI HLEG, "HIGH-LEVEL EXPERT GROUP ON ARTIFICIAL INTELLIGENCE SET UP BY THE EUROPEAN COMMISSION ETHICS GUIDELINES FOR TRUSTWORTHY AI," Apr. 2019. Available: https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

[4] European Commission, "Proposal for a Regulation laying down harmonised rules on Artificial Intelligence European Commission," Jun. 2021. Accessed: Sep. 03, 2023. [Online]. Available: https://www.managementsolutions.com/sites/default/files/publicaciones/eng/202106-NT-Artificial-Intelligence.pdf

[5] PDPC, "ARTIFICIAL INTELLIGENCE GOVERNANCE FRAMEWORK MODEL SECOND EDITION," Jan. 2020. Available: https://www.pdpc.gov.sg/-/media/files/pdpc/pdf-files/resource-for-organisation/ai/sgmodelaigovframework2.pdf

[6] T. B. of C. Secretariat, "Responsible use of artificial intelligence (AI)," aem, Nov. 22, 2018. https://www.canada.ca/en/government/system/digital-

[7] Google, "Google AI Principles," *Google AI*. https://ai.google/responsibility/principles/

[8] IBM, "AI Ethics | IBM," *www.ibm.com*. https://www.ibm.com/impact/ai-ethics#:~:text=At%20IBM%2C%20we%20believe%20AI (accessed Sep. 03, 2023).

[9] Microsoft, "Responsible AI Principles and Approach | Microsoft AI," *www.microsoft.com*. https://www.microsoft.com/en-us/ai/principles-and-approach

[10] Z. Zhang, J. Zhang, and T. Tan, "Analysis and Strategy of AI Ethical Problems," Bulletin of Chinese Academy of Sciences (Chinese Version), vol. 36, no. 2, 2021, doi: https://doi.org/10.16418/j.issn.1000-3045.20210604002.

[11] L. Sijing and W. Lan, "Artificial Intelligence Education Ethical Problems and Solutions," *IEEE Xplore*, Aug. 01, 2018. https://ieeexplore.ieee.org/document/8468773

[12] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings," *Neural Information Processing Systems*, 2016. https://proceedings.neurips.cc/paper_files/paper/2016/hash/a486cd07e4ac3d270571622f4f316ec5-Abstract.html

[13] P. Weng, "Fairness in Reinforcement Learning," *arXiv.org*, Jul. 24, 2019. https://arxiv.org/abs/1907.10323 (accessed Sep. 09, 2023).

[14] "deep re-inforcement learning - Google Search," *www.google.com*. https://www.google.com/search?q=deep+re-inforcement+learning&oq=deep+re-inforcement+learning&aqs=chrome..69i57j0i13i512l5j46i13i512j0i13i512l3.5392j1j7&sourceid=chrome&ie=UTF-8 (accessed Sep. 09, 2023).

[15] W. Wu, T. Huang, and K. Gong, "Ethical Principles and Governance Technology Development of AI in China," *Engineering*, vol. 6, no. 3, Jan. 2020, doi: https://doi.org/10.1016/j.eng.2019.12.015.

[16] "PDPC | Singapore's Approach to AI Governance," *www.pdpc.gov.sg*. https://www.pdpc.gov.sg/help-and-resources/2020/01/model-ai-governance-framework#:~:text= (accessed Sep. 06, 2023).

[17] V. Arya *et al.*, "One Explanation Does Not Fit All: A Toolkit and Taxonomy of AI Explainability Techniques," *arXiv:1909.03012 [cs, stat]*, Sep. 2019, Available: https://arxiv.org/abs/1909.03012

[18] M. Hassan, "New cyber software 'can test the limits of AI's knowledge,'" *The National*, Apr. 04, 2023. https://www.thenationalnews.com/business/technology/2023/04/04/new-cyber-software-can-test-the-limits-of-ais-knowledge/ (accessed Sep. 06, 2023).

[19] T. Jameel, R. Ali, and I. Toheed, "Ethics of Artificial Intelligence: Research Challenges and Potential Solutions," *2020 3rd International Conference on Computing, Mathematics and Engineering Technologies (iCoMET)*, Jan. 2020, doi: https://doi.org/10.1109/icomet48670.2020.9073911.

[20] V. Mallawaarachchi, "How to define a Fitness Function in a Genetic Algorithm?," *Medium*, Nov. 10, 2017. https://towardsdatascience.com/how-to-define-a-fitness-function-in-a-genetic-algorithm-be572b9ea3b4

[21] D. S. Research, "Scalable agent alignment via reward modeling," Medium, Nov. 20, 2018. https://deepmindsafetyresearch.medium.com/scalable-agent-alignment-via-reward-modeling-bf4ab06dfd84#:~:text=Schematic%20illustration%20of%20reward%20modeling (accessed Sep. 04, 2023).

[22] L. Munn, "The uselessness of AI ethics," *AI and Ethics*, Aug. 2022, doi: https://doi.org/10.1007/s43681-022-00209-w.

[23] N. Burleigh, "What Silicon Valley Thinks of Women," vol. 28, no. 5, 2015, Available: http://iuf317.live.s3.amazonaws.com/What%20Silicon%20Valley%20Thinks%20of%20Women33a70b95-cb51-4043-8e9e-aa31fc7ca366.pdf

[24] "Girlfriend 'Complains A Lot ... Interrupts,' Developer Tells Conference," *NPR*, Jun. 04, 2014. https://www.npr.org/sections/alltechconsidered/2014/06/04/318882549/women-complain-a-lot-interrupt-developer-says-at-conference

[25] K. Conger, "Exclusive: Here's The Full 10-Page Anti-Diversity Screed Circulating Internally at Google [Updated]," Gizmodo.com, 2019. https://gizmodo.com/exclusive-heres-the-full-10-page-anti-diversity-screed-1797564320

[26] "The Elephant in the Valley," *The Elephant in the Valley*. https://www.elephantinthevalley.com/

[27] "Sexism, racism and bullying are driving people out of tech, US study finds," the Guardian, Apr. 27, 2017. https://www.theguardian.com/technology/2017/apr/27/tech-industry-sexism-racism-silicon-valley-study

[28] J. Brown, "IBM Watson Reportedly Recommended Cancer Treatments That Were 'Unsafe and Incorrect,'" Gizmodo, Jul. 25, 2018. https://gizmodo.com/ibm-watson-reportedly-recommended-cancer-treatments-tha-1827868882

[29] N. W. Communications, "When a Tesla on Autopilot Kills Someone, Who Is Responsible?," *www.nyu.edu*. https://www.nyu.edu/about/news-

publications/news/2022/march/when-a-tesla-on-autopilot-kills-someone--who-is-responsible--.html#:~:text=The%20driver%20is%20still%20responsible

[30] "How consumers view the transparency of their AI-enabled interactions," Help Net Security, Jul. 11, 2019. https://www.helpnetsecurity.com/2019/07/11/ai-enabled-interactions/

[31] Arnold, L. (2016). "Emobie™: A Robot Companion for Children with Anxiety." In: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI). Christchurch, New Zealand: IEEE

[32] A. Abou Allaban, M. Wang, and T. Padır, "A Systematic Review of Robotics Research in Support of In-Home Care for Older Adults," *Information*, vol. 11, no. 2, p. 75, Jan. 2020, doi: https://doi.org/10.3390/info11020075.

[33] J. Dastin, "Amazon has a plan to make Alexa mimic anyone's voice," Reuters, Jun. 23, 2022. Available: https://www.reuters.com/technology/amazon-has-plan-make-alexa-mimic-anyones-voice-2022-06-22/

[34] R. L. Wilson et al., "Artificial intelligence: An eye cast towards the mental health nursing horizon," International Journal of Mental Health Nursing, Jan. 2023, doi: https://doi.org/10.1111/inm.13121.

[35] Y. A. Ahmed and A. A. Sharo, "On the Education Effect of CHATGPT: Is AI CHATGPT to Dominate Education Career Profession?," *ICCNS*, Jun. 2023, doi: 10.1109/iccns58795.2023.10192993.

[36] OECD, *Artificial Intelligence in Society*. PARIS: OECD, 2019.

[37] A. Najibi, "Racial discrimination in face recognition technology," Science in the News, https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/ (accessed Sep. 10, 2023).

[38] J. Larson, J. Angwin, L. Kirchner, and S. Mattu, "How we analyzed the compas recidivism algorithm," ProPublica, https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm (accessed Sep. 10, 2023).

[39] E. Tat, D. L. Bhatt, and M. G. Rabbat, Addressing bias: Artificial Intelligence in cardiovascular medicine, https://www.thelancet.com/journals/landig/article/PIIS2589-7500(20)30249-1/fulltext (accessed Sep. 10, 2023).

[40] J. D. Rey, "A leaked Amazon memo may help explain why the tech giant is pushing out so many recruiters," Vox, https://www.vox.com/recode/2022/11/23/23475697/amazon-layoffs-buyouts-recruiters-ai-hiring-software (accessed Sep. 10, 2023).

[41] C. Bartneck, C. Lütge, A. Wagner, and S. Welsh, *Introduction to Ethics in Robotics and Ai*. Springer Nature, 2021.

[42] S. M. Liao, *Ethics of Artificial Intelligence*. New York, NY, United States of America: Oxford University Press, 2020.

[43] N. Confessore, "Cambridge Analytica and Facebook: The scandal and the fallout so far," The New York Times, https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html (accessed Sep. 9, 2023).

[44] P. Mozur, "In Hong Kong protests, faces become weapons," The New York Times, https://www.nytimes.com/2019/07/26/technology/hong-kong-protests-facial-recognition-surveillance.html (accessed Sep. 9, 2023).

[45] G. Turner , N. Drozdiak , and M. Day, "Thousands of Amazon workers listen to Alexa conversations," Time, https://time.com/5568815/amazon-workers-listen-to-alexa/#:~:text=Amazon.com%20Inc.,Echo%20owners%27%20homes%20and%20offices. (accessed Sep. 9, 2023).

[46] G. Venkatadri *et al.*, "Privacy risks with Facebook's PII-based targeting: Auditing a data broker's advertising interface," *2018 IEEE Symposium on Security and Privacy (SP)*, 2018. doi:10.1109/sp.2018.00014

[47] B. H. Attarbashi, "How Artificial Intelligence Impacts the Future of Work," *https*, Dec. 09, 2022. https://www.ai-bees.io/post/how-artificial-intelligence-impacts-the-future-of-work

[48] admin, "Equitable AI in the Workplace," Peatworks. https://www.peatworks.org/ai-disability-inclusion-toolkit/equitable-ai-in-the-workplace/#:~:text=%E2%80%9CEquitable%20AI%E2%80%9D%20refers%20to%20AI

[49] O. Bakiner, "What do academics say about artificial intelligence ethics? An overview of the scholarship," *AI and Ethics*, Jun. 2022, doi: https://doi.org/10.1007/s43681-022-00182-4

DOI: https://doi.org/10.15379/ijmst.v10i3.2953