ARTIFICIAL INTELLIGENCE INCIDENTS & ETHICS: A NARRATIVE REVIEW

S.Faiza Nasim, Muhammad Rizwan Ali and Umme Kulsoom

Computer Science and Information Technology, Ned University, Pakistan

sfaizaadnan@gmail.com, mrizwanali@gmail.com, kulsoomshah9@gmail.com

ABSTRACT

There has been a lot of media debate about “Artificial Intelligence (AI) Ethics” nowadays and many scientists and researchers have shared their views on this topic. As technology is evolving, security issues are also emerging in new forms. Machines should be ethical, and the “Build and Design” of such machines should be based on ethics. Infact, AI must have Ethics as a part of design within the software code, just like security measures are encoded within. In this review paper, statistics of AI incidents and areas are presented along with the social impact. Using the online AI Incident Database, some areas of AI applications have been identified, which shows unethical use of AI. Applications like Language and Computer vision models, intelligent robots and autonomous driving are in top ranking. Ethical issues also appear in various forms like incorrect use of technology, racism, non-safety and malicious algorithms with biasness. Data collection has helped to identify the AI ethical issues based on Time, Geographic Locations, Application Areas, and Classifications.

Keywords: *AI Ethics, Ethical Principles, Ethical Behavior, Machine Ethics, AI Guidelines*

1. INTRODUCTION

AI offers many benefits and opportunities for society, including the potential to improve healthcare, finance, education, and surveillance. However, AI also poses risks of harm if misused or deployed without careful consideration. Assessing the risks and benefits of AI is challenging, since it is difficult to determine the extent to which impacts are caused by AI systems. Several AI principles have been set up by national and international bodies, worldwide. These consist of generic and contextual [1]. Recent observations show that AI is becoming an essential part in almost every technological advancement, particularly in areas like Banking, Data Analytics, Autonomous Systems (AS), Manufacturing Industries, etc [2]. Hence AI is also facing the same challenges in terms of Safety, Security, Ethics and Privacy [3]. For instance, Banking Chatbots can be manipulated or exploited in such a way that the hackers may try to impersonate a banking officer and get the details of a customer and do fraud. Autonomous weapons are another example, where decisions are being made by these AI based weapons without having human intervention [4].

This review paper is based on the statistics taken from AI incident Database – A catalog of failures in AI. The AI ethical issues have been classified based on the data processed. 13 application areas are identified where AI ethics issues have occurred the most. Results are then discussed with respect to Time [5], Geographical Distributions, Application Areas and Taxonomy of AI issues [6]. Following which discussion on AI guidelines and principles are presented. Finally, a conclusion is made showing the importance of machine and AI ethics in design and implementation.

2. THEORETICAL FRAMEWORK

Mengyi Wei et al. collected AI ethics incidents mainly from the AI Incident Database [1]. 150 AI ethics incidents with detailed information were chosen [7]. After content analysis, below four descriptive attributes were identified namely:

- Time,
- Geographic Locations,
- Application Areas,
- Classification of AI Ethics Issues.

These four attributes cover the critical information of each AI ethics incident.

Time: This attribute refers to the time the AI ethical issue took place.

Geographic locations: Refers to the geographical distribution of AI ethics incidents which took place around the world [8].

Application areas: This attribute provides information about the areas of AI which are most susceptible to ethical issues.

Taxonomy of AI ethics issues: This classifies AI ethical incidents, which comprehensively show the unethical behavior of AI technology and the impacts on society [9].

2.1 Objectives:

1. Develop mechanisms to benefit AI more for human society and to reduce harm.
2. Improve the ability to evaluate the impacts and risks associated with AI.

3. Able to decide in events when there is disagreement and uncertainty.

2.2 Research Questions

There are many questions that are raised for deciding the AI ethics and regulations for the framework. For instance:

1. Copyright ownership of AI Algorithms has been developed.
2. How much AI has the potential to make discriminations, privacy violations and war crimes?
3. To what extent AI can be compromised to fool someone?
4. Design rules for AI regulations
5. Drafting of AI regulations by those judiciary officials who are not technical experts on this domain.
6. How much AI technology is risky in a particular domain?
7. What areas are of low risks where AI can be easily deployed.
8. Ability to produce false and misleading information generated by BigGAN and GPT-3 models.
9. What ethical principles should AI researchers follow?
10. What is the best way to design AI that aligns with human values?
11. Is it possible or desirable to build moral principles into AI systems?
12. When AI systems cause benefits or harm, who is morally responsible?

Can I determine a person’s act for doing crime? Infact most recently there has been a case in which researchers at the Chicago University, have developed an algorithm that predicts crimes a week prior to their occurrence with 90% accuracy [10].

3. LITERATURE REVIEW

Presently there are no defined legitimate processes to develop and deploy AI in the human community. Most of the self- decisions or rules are already established in technology-based companies like Google. Facebook, Microsoft, AWS, etc [11]. But there is no procedure to scrutinize those decisions or rules by a central government entity at national or international level. Also, the public is not aware about the rules. Those rules can have a great impact on society, but they are not accountable [12]. A proper understanding of AI impacts on society needs to be done before assessing risks and benefits of AI. Furthermore, if those risks or benefits would have a high impact on society. Unsafe and insecure AI systems is on the rise which can become more severe in near future [13]. Take for example, “Autonomous Drone Swarms”, an autonomous lethal weapon is an instrument of mass destruction.

To properly assess risks and benefits, we need a thorough understanding of how AI is already impacting society, and how those impacts are likely to evolve in future [14]. Despite these many real and potential benefits, we are already beginning to see harms arise from the use of AI systems, which could become much more severe with more widespread application of increasingly capable systems [15]. For instance, there is a strong case that “armed fully autonomous drone swarms”, one type of lethal autonomous weapon, qualify as a weapon of mass destruction [16].

There is a strong need to develop a regulatory framework for AI before it gets too late [17]. Technology companies and government bodies should form a consortium to build the regulations for this framework. There is a need for governance in AI which should at least cover three objectives [18].

3.1 Content Categorization

Mengyi Wei and his team were not aware of the categories of AI ethical issues, nor was there enough relevant research available. Categories were left to be discovered during the analysis. The reliability of the data was calculated using Krippedorff’s alpha [18], which is regularly used by researchers in the field of content analysis. Krippedorff’s alpha (α) is a reliability coefficient developed to measure the agreement among observers [19], coders, judges, raters, or measuring instruments and helps in how much the resulting data can be trusted to represent something real [20].

A lot of efforts were put in by the research team who continuously tried their best to refine the methodology. They reached a high agreement with Krippedorff’s Alpha larger than 0.8 on most variables and averaged 0.94 on all variables.

The agreement of the identified thirteen application areas and eight AI ethics issues is summarized in Table 1.

Content Category	Krippendorff’s Alpha
AI supervision	0.79
AI recruitment	0.44
Identity Authentication	1
Language/vision model	0.98
Intelligent recommendation	0.96
Autonomous Driving	1
Intelligent Service Robots	1
Smart Healthcare	1
AI Education	1

Predictive policing	1
Smart Home	1
AI Game	1
Smart Finance	1
Privacy	1
Inappropriate Use (Bad Performance)	0.9
Unethical Use (illegal Use)	0.97
Racial Discrimination	1
Gender Discrimination	0.98
Unfair Algorithm (Evaluation)	0.94
Mental Health	0.86
Physical Safety	1
Average	**0.94**

Table 1: Krippedorff's alpha for each variable. The upper part corresponds to application areas, and the lower part corresponds to AI ethics issues.

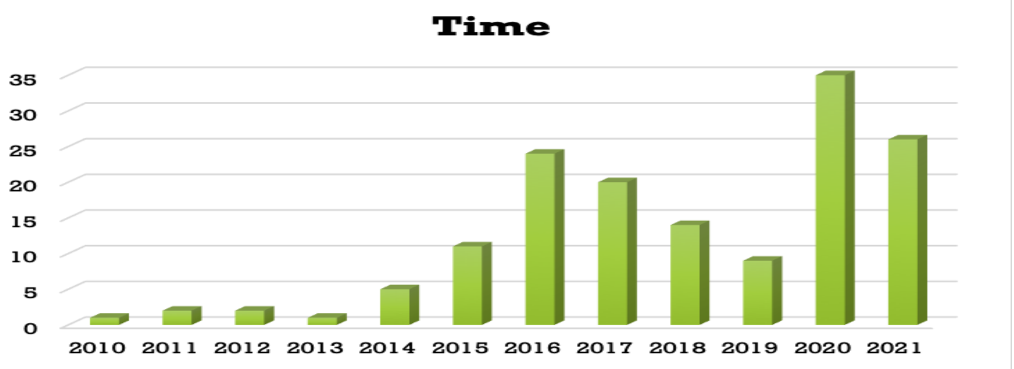

Figure 1: Yearly based AI ethics incidents

3.2 Time-based Evolution of AI Incidents

Figure 1 depicts 150 AI ethics ranging from 2010 to 2021. The incidents annually increased from 2010 to 2016. This was because there were many AI advancements during these years. A decline in the years 2017 to 2019 can be well noticed [21], probably due to a more cautious approach in AI design and implementations. But again, in the year 2020 and 2021, the incidents broke the previous records [22].

3.3 Geographic Distribution of AI Incidents

The AI companies which are in certain developed countries happen to have more AI ethics incidents more in that region than any other part of the world. It was discovered that technology-based organizations are the key players in the development of AI technology and are also the ones who create the technology that causes ethical incidents [23]. Countries like the USA, UK and China where most AI based companies are located have 80 out of 150 incidents.

Another type of location category is Global, depicting those ethical issues which occurred globally, rather than just in a particular specific geographical location or a country [24] as defined here in Figure 2. For example, an incident that happens in a news company is classified into the news and media industry which occurred globally. Another example is the incident of gender bias embedded in NLP, all over the world [25].

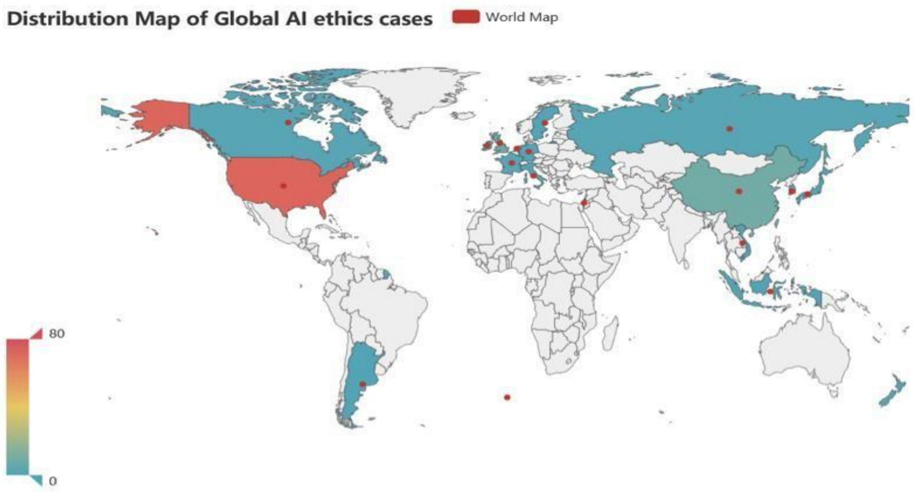

Figure 2: Geographic distribution of AI ethics incidents

4. DATA ANALYSIS

4.1 Application Areas of AI Incidents

Table 2 depicts thirteen areas in which applications of AI were discovered having ethical issues. Following are the areas, according to the number of incidents that took place.

S. No	Fields / Areas of AI	Number of Incidents by Year 2021
1	Intelligent Service Robots	31
2	Language / Vision Models	27
3	Autonomous Cars	17
4	Recommendation Systems	14
5	Identity Authentication	14
6	Supervision / Monitoring under AI	14
7	Smart Health	10
8	AI based recruitment,	10
9	Predictive Policing	8

10	Smart Finance	4
11	AI based Games	2
12	Smart Homes	2
13	AI based Education	2

Table 2: Application Areas of AI ethics incidents

Here in Figure 3 last seven fields, have the occurrences of AI ethics incidents relatively rare, yet these issues have emerged and cannot be ignored in any case.

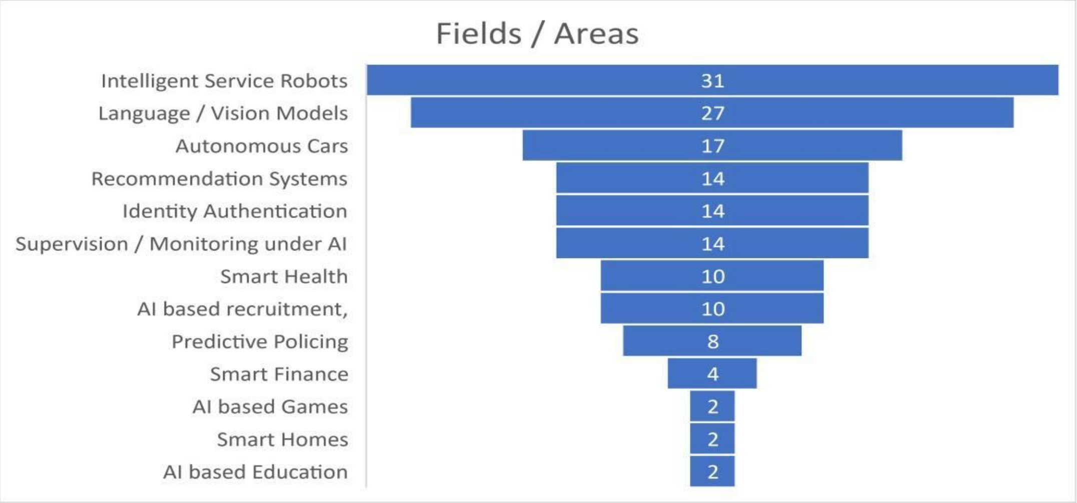

Figure 3: Application fields of AI ethics incidents

4.2 Taxonomy of AI Ethics Issues

Among 150 AI ethical issues, eight categories were discovered during the content analysis. They have been shown here in table 3 in accordance with their respective frequencies.

S. No	Taxonomy of AI ethics issues	Frequency of Events
1	Inappropriate Use (Bad Performance)	48
2	Racial Discrimination	38
3	Physical Safety	32
4	Unfair Algorithm (Evaluation)	22
5	Gender Discrimination	19

6	Privacy	12
7	Unethical (Illegal Use)	11
8	Mental Health	4

Table 3: Taxonomy of AI ethics Issues

It may be the case that an AI application may have more than one ethical issue as classified here in Figure 4 as well. It is worth noticing that one AI application may have multiple AI ethics issues. For example, a chatbot can contain gender discrimination and cause privacy leakage. This exhibits a complex nature of the AI itself.

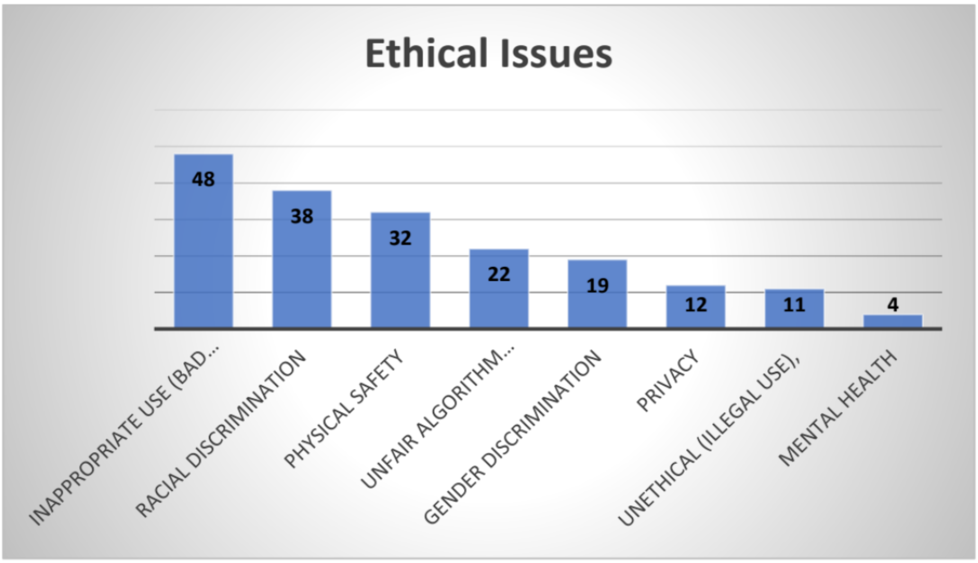

Figure 4: Classification of AI ethics incidents

5. DISCUSSION ON RESULTS

Four attributes, “Time, Geographic Locations, Application Areas and AI Ethic Issues" were discussed. The incidents of AI have increased over time. Developed countries like China, UK and USA have more accidents than other places in the world as the amount of AI work is done is more in these countries than anywhere else. Then Application areas were divided in 13 categories in which Intelligent service robots, language and vision models and autonomous drivers have the most incidents [26]. Lastly, eight categories were identified having AI ethical issues. The most obvious one is “bad performance, which is caused by the consequences due to the limitations in current AI practices [27]. People safety and racism is also the primary concern, the latter being somehow embedded in the designed algorithm, which can have a negative impact for certain groups of people. Incidents have been there where people have been put into risk by robots [28], which were designed for service and manufacturing.

Jobin et al. compiled 84 AI guidelines and presented an overview. Considering the highest frequency, three principles are addressed below [29]:

1. Transparency: Many AI incidents occurred due to lack of transparency. Developers and practitioners were not able to explain the real mechanisms working behind a black-box algorithm. The consequence and performance are not predictable and guaranteed, thus leading to various AI incidents [30].

2. Justice and Fairness: It relates to non-discrimination, non-biased and impartiality. Racism, gender biasness is found to be common on many incidents. Most of these occur in language or computer vision models. This should be a very focused area for the AI experts while designing and deployment [31].

3. Non-maleficence: The principle of nonmaleficence holds that there is an obligation not to inflict harm on others. It is closely associated with the ***maxim primum non nocere*** (first do no harm) [32]. This principle relates to Safety, harm, security and protection concerns. This is the third most common issue in the ethical guidelines. Very less-frequent efforts have been made in developing AI algorithms safe for human society [33].

In general, people do know about the rules and theories behind the ethics, but do not know the ways to implement them in AI. Neither they know about the impact it may create if the ethics are not followed. This review paper helps to understand the incidents which are occurring due to the lack of importance in AI ethics and a point to start thinking to design some operable regulatory framework based on ethics [34].

6. CONCLUSION AND RECOMMENDATIONS

As human decisions are not perfect, so do AI systems as well. We need to foresee if the inherent risks are greater or less than the risks of not using AI. Is it tolerable as well? Terms like “Trustworthy AI”, “Beneficial AI” should come into practice in every walk of life [35]. Many have tried to find ways to compute Ethics and add an ethical component in the design of AI. AI has the advantage of improving human life in many ways, but also has the risks of developing dangerous technologies that can be very harmful to humans. We need to design and apply this technology with care and wisdom [36]. Finally, it is the need of the time that research should be made in machine ethics and applied in a lawful and safe manner. Means are required to integrate values concerning morality, society, and legality in technological developments in the field of AI both in design and implementation.

REFERENCES

[1] M. Wei and Z. Zhou, “Ai ethics issues in real world: Evidence from ai incident database,” *arXiv Prepr. arXiv2206.07635,* 2022.

[2] M. Farouk, “The Universal Artificial Intelligence Efforts to Face Coronavirus COVID-19,” *Int. J. Comput. Inf. Manuf.*, vol. 1, no. 1, pp. 77–93, 2021, [Online]. Available: https://doi.org/10.54489/ijcim.v1i1.47.

[3] H. M. Alzoubi, M. Alshurideh, B. A. Kurdi, I. Akour, and R. Aziz, “Does BLE technology contribute towards improving marketing strategies, customers’ satisfaction and loyalty? The role of open innovation,” *Int. J. Data Netw. Sci.*, vol. 6, no. 2, pp. 449–460, 2022.

[4] A. AlHamad *et al.*, “The effect of electronic human resources management on organizational health of telecommuni-cations companies in Jordan,” *Int. J. Data Netw. Sci.*, vol. 6, no. 2, pp. 429–438, 2022.

[5] J. Kasem and A. Al-Gasaymeh, “A COINTEGRATION ANALYSIS FOR THE VALIDITY OF PURCHASING POWER PARITY: EVIDENCE FROM MIDDLE EAST COUNTRIES,” *Int. J. Technol. Innov. Manag.*, vol. 2, no. 1, 2022.

[6] T. M. Ghazal and H. M. Alzoubi, “Fusion-based supply chain collaboration using machine learning techniques,” *Intell. Autom. \& Soft Comput.*, vol. 31, no. 3, pp. 1671–1687, 2022.

[7] R. Yanamandra and H. M. Alzoubi, “Empirical Investigation of Mediating Role of Six Sigma Approach in Rationalizing the COQ in Service Organizations,” *Oper. Supply Chain Manag. An Int. J.*, vol. 15, no. 1, pp. 122–135, 2022.

[8] E. P. Mondol, “THE ROLE OF VR GAMES TO MINIMIZE THE OBESITY OF VIDEO GAMERS,” *Int. J. Comput. Inf. Manuf.*, vol. 2, no. 1, 2022.

[9] B. Kurdi, M. Alshurideh, I. Akour, H. Alzoubi, B. Obeidat, and A. AlHamad, “The role of digital marketing channels on consumer buying decisions through eWOM in the Jordanian markets,” *Int. J. Data Netw. Sci.*, vol. 6, no. 4, pp. 1175–1186, 2022.

[10] H. Alzoubi, M. Alshurideh, B. Kurdi, K. Alhyasat, and T. Ghazal, “The effect of e-payment and online shopping on sales growth: Evidence from banking industry,” *Int. J. Data Netw. Sci.*, vol. 6, no. 4, pp. 1369–1380, 2022.

[11] K. L. Lee, N. A. N. Azmi, J. R. Hanaysha, H. M. Alzoubi, and M. T. Alshurideh, “The effect of digital supply chain on organizational performance: An empirical study in Malaysia manufacturing industry,” *Uncertain Supply Chain Manag.*, vol. 10, no. 2, pp. 495–510, 2022.

[12] H. M. Alzoubi, M. In’airat, and G. Ahmed, “Investigating the impact of total quality management practices and Six Sigma processes to enhance the quality and reduce the cost of quality: the case of Dubai,” *Int. J. Bus. Excell.*, vol. 27, no. 1, pp. 94–109, 2022.

[13] B. Kurdi, M. Alshurideh, I. Akour, E. Tariq, A. AlHamad, and H. Alzoubi, “The effect of social media influencers’ characteristics on consumer intention and attitude toward Keto products purchase intention,” *Int. J. Data Netw. Sci.*, vol. 6, no. 4, pp. 1135–1146, 2022.

[14] T. Eli and L. A. S. Hamou, “INVESTIGATING THE FACTORS THAT INFLUENCE STUDENTSCHOICE OF ENGLISH STUDIES AS A MAJOR: THE CASE OF UNIVERSITY OF NOUAKCHOTT AL AASRIYA, MAURITANIA,” *Int. J. Technol. Innov. Manag.*, vol. 2, no. 1, 2022.

[15] K. L. Lee, P. N. Romzi, J. R. Hanaysha, H. M. Alzoubi, and M. Alshurideh, “Investigating the impact of benefits and challenges of IOT adoption on supply chain performance and organizational performance: An empirical study in Malaysia,” *Uncertain Supply Chain Manag.*, vol. 10, no. 2, pp. 537–550, 2022.

[16] Z. Kallenborn, *Are Drone Swarms Weapons of Mass Destruction?* US Air Force Center for Strategic Deterrence Studies, Air University, 2020.

[17] J. F. Weaver, "Regulation of artificial intelligence in the United States," in *Research Handbook on the Law of Artificial Intelligence*, Edward Elgar Publishing, 2018, pp. 155–212.

[18] J. Whittlestone and S. Clarke, "AI Challenges for Society and Ethics," *Oxford Handb. AI Gov.*, vol. 1, no. 1, pp. 1–20, 2022, doi: 10.1093/oxfordhb/9780197579329.013.3.

[19] M. Shamout, R. Ben-Abdallah, M. Alshurideh, A. Kurdi, and H. B., "S. (2022). A conceptual model for the adoption of autonomous robots in supply chain and logistics industry," *Uncertain Supply Chain Manag.*, vol. 10, no. 2, pp. 577–592.

[20] S. M. Butt, "Management and Treatment of Type 2 Diabetes," *Int. J. Comput. Inf. Manuf.*, vol. 2, no. 1, 2022.

[21] H. M. Alzoubi, G. Ahmed, and M. Alshurideh, "An empirical investigation into the impact of product quality dimensions on improving the order-winners and customer satisfaction," *Int. J. Product. Qual. Manag.*, vol. 36, no. 2, pp. 169–186, 2022.

[22] H. M. Alzoubi, H. Elrehail, J. R. Hanaysha, A. Al-Gasaymeh, and R. Al-Adaileh, "The Role of Supply Chain Integration and Agile Practices in Improving Lead Time During the COVID-19 Crisis," *Int. J. Serv. Sci. Manag. Eng. Technol.*, vol. 13, no. 1, pp. 1–11, 2022.

[23] M. T. Alshurideh *et al.*, "Fuzzy assisted human resource management for supply chain management issues," *Ann. Oper. Res.*, pp. 1–19, 2022.

[24] G. M. Qasaimeh and H. E. Jaradeh, "THE IMPACT OF ARTIFICIAL INTELLIGENCE ON THE EFFECTIVE APPLYING OF CYBER GOVERNANCE IN JORDANIAN COMMERCIAL BANKS," *Int. J. Technol. Innov. Manag.*, vol. 2, no. 1, 2022.

[25] B. Kurdi, H. Alzoubi, I. Akour, and M. Alshurideh, "The effect of blockchain and smart inventory system on supply chain performance: Empirical evidence from retail industry," *Uncertain Supply Chain Manag.*, vol. 10, no. 4, pp. 1111–1116, 2022.

[26] M. Alshurideh, B. Kurdi, H. Alzoubi, B. Obeidat, S. Hamadneh, and A. Ahmad, "The influence of supply chain partners' integrations on organizational performance: The moderating role of trust," *Uncertain Supply Chain Manag.*, vol. 10, no. 4, pp. 1191–1202, 2022.

[27] A. Alzoubi, "Renewable Green hydrogen energy impact on sustainability performance," *Int. J. Comput. Inf. Manuf.*, vol. 1, no. 1, pp. 94–105, 2021, [Online]. Available: https://doi.org/10.54489/ijcim.v1i1.46.

[28] G. Ahmed and N. Al Amiri, "THE TRANSFORMATIONAL LEADERSHIP OF THE FOUNDING LEADERS OF THE UNITED ARAB EMIRATES: SHEIKH ZAYED BIN SULTAN AL NAHYAN AND SHEIKH RASHID BIN SAEED AL MAKTOUM," *Int. J. Technol. Innov. Manag.*, vol. 2, no. 1, 2022.

[29] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," *Nat. Mach. Intell.*, vol. 1, no. 9, pp. 389–399, 2019.

[30] A. J. Obaid, "Assessment of Smart Home Assistants as an IoT," *Int. J. Comput. Inf. Manuf.*, vol. 1, no. 1, pp. 18–38, 2021, [Online]. Available: https://doi.org/10.54489/ijcim.v1i1.34.

[31] N. Alsharari, "THE IMPLEMENTATION OF ENTERPRISE RESOURCE PLANNING (ERP) IN THE UNITED ARAB EMIRATES: A CASE OF MUSANADA CORPORATION," *Int. J. Technol. Innov. Manag.*, vol. 2, no. 1, 2022.

[32] M. Farouk, "STUDYING HUMAN ROBOT INTERACTION AND ITS CHARACTERISTICS," *Int. J. Comput. Inf. Manuf.*, vol. 2, no. 1, 2022.

[33] V. Victoria, "IMPACT OF PROCESS VISIBILITY AND WORK STRESS TO

IMPROVE SERVICE QUALITY: EMPIRICAL EVIDENCE FROM DUBAI RETAIL INDUSTRYIMPACT OF PROCESS VISIBILITY AND WORK STRESS TO IMPROVE SERVICE QUALITY: EMPIRICAL EVIDENCE FROM DUBAI RETAIL INDUSTRY," *Int. J. Technol. Innov. Manag.*, vol. 2, no. 1, 2022.

[34] N. Ratkovic, "IMPROVING HOME SECURITY USING BLOCKCHAIN," *Int. J. Comput. Inf. Manuf.*, vol. 2, no. 1, 2022.

[35] N. Radwan, "THE INTERNET'S ROLE IN UNDERMINING THE CREDIBILITY OF THE HEALTHCARE INDUSTRY," *Int. J. Comput. Inf. Manuf.*, vol. 2, no. 1, 2022.

[36] A. Alzoubi, "MACHINE LEARNING FOR INTELLIGENT ENERGY CONSUMPTION IN SMART HOMES," *Int. J. Comput. Inf. Manuf.*, vol. 2, no. 1, 2022.